

IBM SPSS Modeler Text Analytics 18.4
User's Guide



Nota

Antes de utilizar esta información y el producto al que se refiere, lea la información en [“Avisos” en la página 231](#).

Información del producto

Esta edición se aplica a la versión 18.4.0 de IBM® SPSS Modeler Text Analytics y a todos los releases y modificaciones posteriores hasta que se indique lo contrario en nuevas ediciones.

© Copyright International Business Machines Corporation .

Contenido

Prefacio.....	ix
Acerca de IBM Business Analytics.....	ix
Soporte técnico.....	ix
 Capítulo 1. Acerca deIBM SPSS Modeler Text Analytics.....	1
actualizaciónIBM SPSS Modeler Text Analytics.....	1
Acerca de la minería de textos.....	2
Cómo funciona la extracción.....	5
Cómo funciona la categorización.....	6
IBM SPSS Modeler Text Analyticsnodos.....	7
Aplicaciones.....	8
 Capítulo 2. Lectura de texto de origen.....	11
Nodo Lista de archivos.....	11
Nodo Lista de archivos: pestaña Configuración.....	11
Nodo Lista de archivos: Otras pestañas.....	12
Utilización del nodo Lista de archivos en minería de textos.....	12
Nodo canal de información web.....	13
Nodo canal de información web: pestaña Entrada.....	13
Nodo Canal de información web: pestaña Registros.....	14
Nodo canal de información web: pestaña Filtro de contenido.....	16
Utilización del nodo Canal de información web en Minería de textos.....	16
Nodo de idioma.....	17
Nodo de idioma: Pestaña Valores.....	17
 Capítulo 3. Minería para conceptos y categorías.....	19
Nodo de modelado de minería de textos.....	20
Nodo Minería de textos: pestaña Campos.....	21
Nodo Minería de textos: pestaña Modelo.....	23
Nodo Minería de textos: pestaña Experto.....	27
Muestreo en sentido ascendente para ahorrar tiempo.....	29
Utilización del nodo Minería de texto en una ruta.....	29
Nugget de minería de textos: Modelo de concepto.....	30
Modelo de concepto: Pestaña Modelo.....	31
Modelo de concepto: pestaña Configuración.....	33
Modelo de concepto: pestaña Campos.....	34
Modelo de concepto: pestaña Resumen.....	35
Utilización de nuggets de modelo de concepto en una ruta.....	35
Nugget de minería de textos: Modelo de categoría.....	38
Nugget de modelo de categoría: Pestaña Modelo.....	39
Nugget de modelo de categoría: pestaña Configuración.....	40
Nugget de modelo de categoría: Otras pestañas.....	41
Utilización de nuggets de modelo de categoría en una ruta.....	42
 Capítulo 4. Minería para enlaces de texto.....	45
Nodo Análisis para los enlaces de texto.....	45
Nodo Análisis para los enlaces de texto: pestaña Campos.....	45
Nodo Análisis de enlaces de texto: pestaña Experto.....	47
Salida de nodo TLA.....	48
Almacenamiento en caché de resultados TLA.....	49

Utilización del nodo Análisis para los enlaces de texto en una ruta.....	49
Capítulo 5. Examinar texto de origen externo.....	53
Nodo Visor de archivos.....	53
Configuración del nodo Visor de archivos.....	53
Utilización del nodo Visor de archivo.....	53
Capítulo 6. Propiedades de nodo para los scripts.....	57
Nodo Lista de archivos: filelistnode.....	57
Nodo Canal de información web: webfeednode.....	57
Nodo de idioma: languageidentifier.....	58
Nodo minería de textos: TextMiningWorkbench.....	59
Nugget de modelo de minería de textos: TMWBModelApplier.....	62
Nodo Análisis de enlaces de texto: textlinkanalysis.....	64
Capítulo 7. Modalidad de área de trabajo interactiva.....	67
La vista de Categorías y Conceptos.....	67
La vista de clústeres.....	70
La vista Análisis de enlaces de texto.....	71
La vista Editor de recursos.....	73
Opciones de configuración.....	74
Opciones: separador Sesión.....	75
Opciones: Visualizar pestaña.....	75
Opciones: Pestaña Sonidos.....	76
Configuración de Microsoft Internet Explorer para obtener ayuda.....	76
Generación de los nuggets de modelo y los nodos de modelado.....	76
Guardar y actualizar nodos de modelado.....	77
Cierre y finalización de sesiones.....	77
Accesibilidad desde el teclado.....	77
Métodos abreviados de los cuadros de diálogo.....	78
Capítulo 8. Extracción de conceptos y tipos.....	81
Resultados de extracción: Conceptos y tipos.....	81
Extraer datos.....	82
Filtración de resultados de extracción.....	85
Exploración de mapas de conceptos.....	86
Creación de índices de mapas de conceptos.....	88
Refinar resultados de extracción.....	89
Añadir sinónimos.....	90
Añadir conceptos a tipos.....	91
Excluir conceptos de una extracción.....	92
Forzado de palabras en la extracción.....	93
Capítulo 9. Categorización de datos de texto.....	95
El panel de categorías.....	96
Métodos y estrategias para crear categorías.....	98
Métodos para crear categorías.....	98
Estrategias para crear categorías.....	98
Sugerencias para crear categorías.....	99
Selección de los mejores descriptores.....	100
Acerca de las categorías.....	102
Propiedades de categoría.....	103
El panel Datos.....	103
Relevancia de categoría.....	105
Marcar respuestas.....	105
Crear categorías.....	106
Configuración lingüística avanzada.....	108

Acerca de técnicas lingüísticas.....	110
Configuración avanzada de frecuencia.....	114
Ampliación de categorías.....	115
Creación manual de categorías.....	118
Creación de categorías nuevas o cambio de nombre de categorías.....	118
Creación de categorías mediante el método de arrastrar y soltar.....	119
Uso de reglas de categoría.....	119
Sintaxis de regla de categoría.....	120
Uso de patrones TLA en las reglas de categoría.....	122
Uso de comodines en reglas de categoría.....	124
Ejemplos de reglas de categoría.....	127
Creación de reglas de categoría.....	129
Edición y eliminación de reglas.....	130
Importación y exportación de categorías predefinidas.....	130
Importación de categorías predefinidas.....	130
Exportación de categorías.....	134
Uso de los paquetes de análisis de texto.....	135
Creación de paquetes de análisis de texto.....	136
Carga de los paquetes de análisis de texto.....	136
Actualización de los paquetes de análisis de texto.....	137
Edición y refinamiento de categorías.....	138
Añadir descriptores a las categorías.....	138
Edición de descriptores de categoría.....	139
Cómo mover categorías.....	139
Aplanamiento de categorías.....	140
Fusión o combinación de categorías.....	140
Forzar los documentos de las categorías.....	140
Eliminación de categorías.....	141
Capítulo 10. Análisis de clústeres.....	143
Creando clústeres.....	144
Calcular valores de similitud de enlaces.....	145
Exploración de los clústeres.....	146
Definiciones de clúster.....	147
Capítulo 11. Exploración del Análisis de enlaces de texto.....	149
Extracción de resultados del patrón TLA.....	150
Patrones de tipo y concepto.....	151
Filtración de resultados TLA.....	151
Panel Datos.....	153
Marcar respuestas.....	154
Reglas de reasignación de tipo.....	155
Capítulo 12. Visualización de gráficos.....	159
Gráficos de categoría.....	159
Gráfico de barras de categorías.....	160
Gráfico de malla de categorías.....	160
Tabla de malla de categorías.....	160
Gráficos de clúster.....	161
Gráfico concepto web.....	161
Gráfico clúster web.....	162
Gráficos de Análisis de enlace de texto.....	162
Gráfico concepto web.....	163
Tipo de gráfico web.....	163
Uso de barras de herramientas y paletas de gráficos.....	163
Capítulo 13. Editor de recursos de sesión.....	165

Edición de recursos en el editor de recursos.....	165
Creación y actualización de plantillas.....	166
Conmutar entre plantillas de recursos.....	166
Capítulo 14. Plantillas y recursos.....	169
El editor de plantillas frente al editor de recursos.....	170
La interfaz Editor.....	170
Apertura de plantillas.....	173
Guardado de plantillas.....	173
Actualización de los recursos en el nodo después de cargar.....	174
Administración de plantillas.....	175
Importación y exportación de plantillas.....	175
Cómo salir del Editor de plantillas.....	176
Copia de seguridad de los recursos.....	176
Importar archivos de recursos.....	176
Capítulo 15. Trabajo con bibliotecas.....	179
Bibliotecas suministradas.....	179
Creación de bibliotecas.....	180
Añadir bibliotecas públicas.....	180
Búsqueda de términos y tipos.....	181
Visión de bibliotecas.....	182
Administración de las bibliotecas locales.....	182
Cambio de nombre de las bibliotecas locales.....	182
Desactivación de bibliotecas locales.....	182
Eliminación de bibliotecas locales.....	183
Administración de bibliotecas públicas.....	183
Compartimiento de bibliotecas.....	184
Publicación de bibliotecas.....	185
Actualización de bibliotecas.....	185
Resolución de conflictos.....	186
Capítulo 16. Acerca de los diccionarios de biblioteca.....	187
Diccionarios de tipo.....	187
Tipos incorporados.....	188
Crear tipos.....	189
Añadir términos.....	190
Forzar términos.....	192
Renombrar tipos.....	193
Cómo mover tipos.....	193
Inhabilitar y suprimir tipos.....	193
Diccionarios de sustitución/sinónimos.....	194
Definir sinónimos.....	195
Definir elementos opcionales.....	196
Desactivación y eliminación de sustituciones.....	196
Diccionarios de exclusión.....	197
Capítulo 17. Acerca de los recursos avanzados.....	199
Buscar.....	200
Reemplazo.....	200
Idioma de destino para Recursos.....	201
Agrupación difusa.....	201
Entidades no lingüísticas.....	202
Definiciones de expresiones regulares.....	203
Normalización.....	205
Configuración.....	205
Gestión de idiomas.....	206

Patrones de extracción.....	207
Definiciones forzadas.....	209
Abreviaturas.....	210
Capítulo 18. Sobre las reglas de enlaces de texto.....	211
Dónde trabajar en las reglas de enlaces de texto.....	211
Dónde comenzar.....	212
Cuándo editar o crear reglas.....	212
Simulación de resultados del análisis de enlace de texto.....	213
Definición de datos para la simulación.....	213
Descripción de los resultados de simulación.....	214
Navegación por reglas y macros en el árbol.....	215
Cómo trabajar con macros.....	216
Creación y edición de macros.....	216
Inhabilitación y supresión de macros.....	217
Comprobación de errores, guardado y cancelación.....	217
Macros especiales: mTopic, mNonLingEntities, SEP.....	218
Cómo trabajar con reglas de enlace de texto.....	219
Creación y edición de reglas.....	221
Inhabilitación y supresión de reglas.....	222
Comprobación de errores, guardado y cancelación.....	222
Orden de proceso para reglas.....	223
Cómo trabajar con conjuntos de reglas (varios pasos).....	224
Elementos soportados para reglas y macros.....	224
Visualizar y trabajar en la modalidad de origen.....	226
Avisos.....	231
Marcas comerciales.....	232
Índice.....	233

Prefacio

IBM SPSS Modeler Text Analytics ofrece potentes prestaciones de análisis de texto, que utiliza tecnologías de lingüística avanzada y Procesamiento del lenguaje natural (NLP) para que procese rápidamente una gran variedad de datos de texto sin estructurar, extraer y organizar los elementos clave. Además, IBM SPSS Modeler Text Analytics puede agrupar estos conceptos en categorías.

Alrededor del 80% de los datos retenidos dentro de una organización están en formato de documentos de texto (por ejemplo, informes, páginas web, correos electrónicos y notas de centro de atención telefónica). El texto es un factor clave para habilitar a una organización a obtener una mayor comprensión del comportamiento de sus clientes. Un sistema que incorpora tecnología NLP puede extraer conceptos de forma inteligente (incluidas frases compuestas). Además, el conocimiento del lenguaje subyacente permite la clasificación de términos en grupos relacionados (como por ejemplo, productos, organizaciones o personas) utilizando el significado y el contexto. Consecuentemente, puede determinar de forma rápida la relevancia de la información según sus necesidades. Estos conceptos y categorías extraídos se pueden combinar con los datos estructurados existentes, como pueden ser datos demográficos y se pueden aplicar al modelado en el conjunto completo de herramientas de minería de datos de IBM SPSS Modeler para tomar decisiones mejores y más certeras.

Los sistemas lingüísticos son sensibles al conocimiento: cuanto mayor sea la cantidad de información contenida en sus diccionarios, mayor será la calidad de los resultados. IBM SPSS Modeler Text Analytics se entrega con un conjunto de recursos lingüísticos, como diccionarios de términos y sinónimos, bibliotecas y plantillas. Además, este producto permite desarrollar y refinar dichos recursos lingüísticos para su contexto. El ajuste preciso de los recursos lingüísticos suele ser un proceso iterativo que resulta necesario para la precisión de la recuperación y la categorización de los conceptos. También se incluyen plantillas, bibliotecas y diccionarios personalizados para dominios específicos, como puede ser la terminología CRM y genómica.

Acerca de IBM Business Analytics

IBM Business Analytics proporciona información completa, coherente y precisa en la que confían para mejorar el rendimiento de su negocio quienes toman las decisiones. Un conjunto integral de inteligencia empresarial, análisis predictivo, rendimiento financiero y gestión de estrategias y aplicaciones de análisis que ofrece una perspectiva clara, inmediata e interactiva del rendimiento actual y la capacidad de predecir resultados futuros. En combinación con extensas soluciones sectoriales, prácticas probadas y servicios profesionales, las organizaciones de cualquier tamaño pueden conseguir el máximo de productividad, automatizar las decisiones de forma fiable y alcanzar mejores resultados.

Como parte de estos documentos, IBM SPSS Predictive Analytics ayuda a las organizaciones a predecir situaciones futuras y a actuar de forma proactiva con esa información para mejorar sus resultados. Clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología IBM SPSS como mejora competitiva para atraer, conservar y aumentar la clientela reduciendo el fraude y los riesgos. Al incorporar el software de IBM SPSS en sus operaciones diarias, las organizaciones se convierten en empresas predictivas, capaces de dirigir y automatizar decisiones para alcanzar los objetivos comerciales y lograr una ventaja considerable sobre la competencia. Para obtener más información o contactar con un representante, visite <http://www.ibm.com/spss>.

Soporte técnico

Hay asistencia técnica disponible para los clientes de mantenimiento. Los clientes podrán ponerse en contacto con el servicio de asistencia técnica si desean recibir ayuda sobre la utilización de los productos de IBM Corp. o sobre la instalación en los entornos de hardware admitidos. Para ponerse en contacto con el servicio de asistencia técnica, consulte el sitio web de IBM Corp. en <http://www.ibm.com/support>. Tenga a mano su acuerdo de asistencia y esté preparado para identificarse a sí mismo y a su organización al solicitar ayuda.

Capítulo 1. Acerca de IBM SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics ofrece potentes prestaciones de análisis de texto, que utiliza tecnologías de lingüística avanzada y Procesamiento del lenguaje natural (NLP) para que procese rápidamente una gran variedad de datos de texto sin estructurar, extraer y organizar los elementos clave. Además, IBM SPSS Modeler Text Analytics puede agrupar estos conceptos en categorías.

Alrededor del 80% de los datos retenidos dentro de una organización están en formato de documentos de texto (por ejemplo, informes, páginas web, correos electrónicos y notas de centro de atención telefónica). El texto es un factor clave para habilitar a una organización a obtener una mayor comprensión del comportamiento de sus clientes. Un sistema que incorpora tecnología NLP puede extraer conceptos de forma inteligente (incluidas frases compuestas). Además, el conocimiento del lenguaje subyacente permite la clasificación de términos en grupos relacionados (como por ejemplo, productos, organizaciones o personas) utilizando el significado y el contexto. Consecuentemente, puede determinar de forma rápida la relevancia de la información según sus necesidades. Estos conceptos y categorías extraídos se pueden combinar con los datos estructurados existentes, como pueden ser datos demográficos y se pueden aplicar al modelado en el conjunto completo de herramientas de minería de datos de IBM SPSS Modeler para tomar decisiones mejores y más certeras.

Los sistemas lingüísticos son sensibles al conocimiento: cuanto mayor sea la cantidad de información contenida en sus diccionarios, mayor será la calidad de los resultados. IBM SPSS Modeler Text Analytics se entrega con un conjunto de recursos lingüísticos, como diccionarios de términos y sinónimos, bibliotecas y plantillas. Además, este producto permite desarrollar y refinar dichos recursos lingüísticos para su contexto. El ajuste preciso de los recursos lingüísticos suele ser un proceso iterativo que resulta necesario para la precisión de la recuperación y la categorización de los conceptos. También se incluyen plantillas, bibliotecas y diccionarios personalizados para dominios específicos, como puede ser la terminología CRM y genómica.

Despliegue. Puede desplegar rutas de minería de textos al utilizar IBM SPSS Modeler Solution Publisher para la puntuación en tiempo real de datos no estructurados. La posibilidad de desplegar estas rutas le garantiza implementaciones de minería de texto de circuito cerrado satisfactorias. Por ejemplo, su organización ahora puede analizar anotaciones de llamadas entrantes o salientes al aplicar los modelos predictivos para incrementar la precisión del mensaje de marketing en tiempo real.

Para ejecutar IBM SPSS Modeler Text Analytics con IBM SPSS Modeler Solution Publisher, añada el directorio `<install_directory>/ext/bin/spss.TMWBServer` a la variable de entorno `$LD_LIBRARY_PATH`.

Nota: El adaptador japonés para IBM SPSS Modeler Text Analytics ha pasado a estar en desuso a partir de la versión 18.1.

actualización IBM SPSS Modeler Text Analytics

Antes de instalar IBM SPSS Modeler Text Analytics, deberá guardar y exportar todos los TAP, plantillas y bibliotecas de la versión actual que desee utilizar en la nueva versión. Le recomendamos que guarde estos archivos en un directorio que no sea borrado o sobrescrito por la instalación de la versión actualizada.

Tras instalar la última versión de IBM SPSS Modeler Text Analytics, puede cargar el archivo TAP guardado y cualquier biblioteca guardada, o importar y cargar cualquier plantilla guardada para utilizarlos en la última versión.

Importante: Si desinstala su versión actual sin guardar y exportar primero los archivos que necesita, cualquier trabajo de TAP, plantilla y biblioteca pública realizado en la versión anterior se perderá y no se podrá utilizar en la última versión de IBM SPSS Modeler Text Analytics.

Acerca de la minería de textos

Hoy en día, cada vez más información se está almacenando en formatos no estructurados y semiestructurados, como correos electrónicos de clientes, notas del centro de atención telefónica, respuestas de encuestas abiertas, canales de noticias, formularios web, etc. Esta abundancia de información plantea un problema a muchas organizaciones que se preguntan: "¿Cómo podemos recoger, explorar y aprovechar esta información?"

La *minería de textos* es el proceso de analizar colecciones de materiales de texto con el objeto de capturar los temas y conceptos clave y descubrir las relaciones ocultas y las tendencias existentes sin necesidad de conocer las palabras o los términos exactos que los autores han utilizado para expresar dichos conceptos. La minería de textos y la acción de recuperar información son conceptos que a veces se confunden, aunque son bastante diferentes. Una recuperación precisa de la información y su almacenamiento supone un reto importante, pero la extracción y administración de contenido de calidad, de terminología y de las relaciones contenidas en la información son procesos cruciales y determinantes.

Minería de textos y minería de datos

Para cada artículo de texto, la minería de textos basada en la lingüística devuelve un índice de conceptos e información sobre los mismos. Esta información estructurada y desglosada puede combinarse con otros orígenes de datos para abarcar preguntas como:

- ¿Qué conceptos aparecen juntos?
- ¿A qué otras cosas están vinculados?
- ¿Qué categorías de nivel superior pueden crearse a partir de la información extraída?
- ¿Qué es lo que predicen los conceptos o las categorías?
- ¿Cómo predicen el comportamiento los conceptos o las categorías?

La combinación de minería de textos y minería de datos ofrece un punto de vista más amplio que el de los datos estructurados o el de los datos no estructurados por separado. Este proceso suele incluir los pasos siguientes:

1. **Identificar el texto en el que se va a realizar la minería.** Preparar el texto para el proceso de minería. Si el texto existe en varios archivos, guarde los archivos en una misma ubicación. Para las bases de datos, determine el campo que contiene el texto.
2. **Minar el texto y extraer datos estructurados.** Aplicar los algoritmos de minería de textos al texto de origen.
3. **Crear modelos de categoría y concepto.** Identificar los conceptos clave y/o crear categorías. El número de conceptos que se devuelven de los datos no estructurados suelen ser muy alto. Identificar los mejores conceptos y categorías para puntuar.
4. **Analizar los datos estructurados.** Emplear técnicas de minería de datos convencionales, como el clúster, la clasificación y el modelado predictivo, con el objeto de descubrir las relaciones entre los conceptos. Fusionar los conceptos extraídos con otros datos estructurados para predecir comportamientos futuros basados en los conceptos.

Análisis de textos y categorización

El análisis de texto, un tipo de análisis cualitativo, es la extracción de información útil del texto de manera que las ideas o los conceptos clave que contiene el texto pueden agruparse en una serie de categorías apropiadas. El análisis del texto puede realizarse en textos de cualquier tipo y longitud, aunque el enfoque del análisis puede ser diferente.

Los registros o documentos más breves se categorizan con mayor facilidad, porque no son tan complejos y suelen tener menos palabras y respuestas ambiguas. Por ejemplo, una encuesta con preguntas abiertas breves; si pedimos a la gente que nombre sus tres actividades vacacionales favoritas, podemos esperar muchas respuestas cortas del tipo *ir a la playa*, *visitar parques nacionales* o *no hacer nada*. Las respuestas preguntas abiertas, por otro lado, pueden ser bastante complejas y muy largas, sobre todo si los encuestados son cultos, están motivados y tienen tiempo suficiente para completar un cuestionario. Si

pedimos a la gente que nos hable sobre sus creencias políticas en una encuesta, o tenemos una fuente blog sobre política, encontraremos comentarios más largos sobre todo tipo de temas y posturas.

La capacidad de extraer conceptos clave y crear categorías intuitivas para estos orígenes de texto más largos en un breve período de tiempo es la principal ventaja de utilizar IBM SPSS Modeler Text Analytics. Esta ventaja se obtiene mediante la combinación de técnicas automáticas lingüísticas y estadísticas, gracias a las que se obtienen los resultados más fiables en cada fase del proceso de análisis de texto.

Procesamiento lingüístico y NLP

El problema principal en la administración de todos estos datos de texto no estructurados radica en la ausencia de reglas estándares para escribir texto y que el ordenador pueda entenderlo. El idioma, y por consiguiente el significado, varía en cada documento y en cada elemento del texto. La única forma de recuperar y organizar con precisión estos datos no estructurados es analizar el idioma y descubrir su significado. Existen diversos métodos automáticos diferentes para extraer conceptos a partir de información no estructurada. Estos métodos pueden desglosarse en dos tipos, lingüísticos y no lingüísticos.

Algunas organizaciones han intentado emplear soluciones automáticas no lingüísticas basadas en estadísticas y redes neuronales. Mediante la tecnología informática, estas soluciones pueden explorar y categorizar los conceptos clave con más rapidez que los lectores humanos. Por desgracia, la precisión de estas soluciones es muy baja. La mayoría de los sistemas basados en estadística solamente hacen un recuento del número de veces que se repiten las palabras, y calculan una proximidad estadística a los conceptos relacionados. Generan muchos resultados irrelevantes, datos innecesarios, y pasan por alto resultados que deberían haberse encontrado, a los que se les llama "silenciosos".

Para compensar esta limitación de la precisión, algunas soluciones incorporan reglas no lingüísticas complejas que ayudan a distinguir entre resultados relevantes e irrelevantes. A esto se le conoce como *minería de textos basada en reglas*.

Minería de textos basada en lingüística, por otro lado, aplica los principios de procesamiento de lenguaje natural (NLP), análisis asistido por sistema de lenguajes humanos, al análisis de palabras, frases y sintaxis, o estructura, del texto. Un sistema que incorpora tecnología NLP puede extraer conceptos de forma inteligente (incluidas frases compuestas). Además, el conocimiento del lenguaje subyacente permite la clasificación de conceptos en grupos relacionados (como por ejemplo, productos, organizaciones o personas) utilizando el significado y el contexto.

La minería de textos basada en lingüística encuentra significado en el texto del modo en que lo hacen la personas, reconociendo una variedad de formas de palabra como similares en su significado y analizando la estructura de la oración para proporcionar una infraestructura para entender el texto. Este método ofrece la velocidad y la rentabilidad propia de los sistemas basados en estadísticas, pero proporciona un grado de precisión mucho más alto y menos intervención humana.

Para ilustrar la diferencia entre enfoques basados en estadísticas y basados en lingüística durante el proceso de extracción, considere cómo respondería cada uno a una consulta sobre *reproduction of documents*. Tanto las soluciones basadas en estadísticas como las basadas en lingüística tendrían que ampliar la palabra *reproduction* para incluir sinónimos, como *copy* y *duplication*. De lo contrario, se pasaría por alto información relevante. Pero si una solución basada en estadísticas intenta realizar este tipo de búsqueda de sinonimia para otros términos con el mismo significado, es probable que también incluya el término *birth*, generando un número de resultados irrelevantes. La comprensión del idioma reduce la ambigüedad del texto, lo que convierte a la minería de textos basada en lingüística, por definición, en el método más fiable.

Comprender el funcionamiento del proceso de extracción puede ayudarle a tomar decisiones clave a la hora de ajustar los recursos lingüísticos (bibliotecas, tipos, sinónimos, etc.). Estos son los pasos del proceso de extracción:

- Conversión de datos de origen en un formato estándar
- Identificar los términos candidatos
- Identificar las clases de equivalencias y la integración de sinónimos

- Asignar un tipo
- Crear índices y, cuando se le pida, extraer patrones con un verificador de datos secundario

Paso 1. Conversión de datos de origen en un formato estándar

En este primer paso, los datos que se importan se convierten en un formato uniforme que puede utilizarse para realizar otros análisis. Esta conversión se lleva a cabo internamente y no cambia los datos originales.

Paso 2. Identificar los términos candidatos

Es importante comprender el rol de los recursos lingüísticos en la identificación de los términos candidatos durante la extracción lingüística. Los recursos lingüísticos se utilizan siempre que se ejecuta una extracción. Existen en forma de plantillas, bibliotecas y recursos compilados. Las bibliotecas incluyen listas de palabras, relaciones y otra información que se utiliza para especificar o ajustar la extracción. Estos recursos compilados no se pueden ver ni editar. Sin embargo, los recursos restantes sí pueden editarse en el Editor de plantillas o, si se encuentra en una sesión de entorno de trabajo interactivo, en el Editor de recursos.

Los recursos compilados son componentes internos principales del motor de extracción dentro de IBM SPSS Modeler Text Analytics. Estos recursos incluyen un diccionario general que contiene una lista de formatos base con un código de categoría léxica (sustantivo, verbo, adjetivo, etc.).

Además de los recursos compilados, se entregan varias bibliotecas con el producto y pueden utilizarse para complementar los tipos y las definiciones de conceptos en los recursos compilados, así como para ofrecer sinónimos. Estas bibliotecas (y las personalizadas que cree) se componen de varios diccionarios. Estos incluyen diccionarios de tipos, diccionarios de sinónimos y diccionarios de exclusiones.

Una vez importados y convertidos los datos, el motor de extracción empezará a identificar los términos candidatos para la extracción. Los términos candidatos son palabras o grupos de palabras que se utilizan para identificar conceptos en el texto. Durante el proceso del texto, las palabras simples (*unitérminos*) y palabras compuestas (*multitérminos*) se identifican mediante extractores de patrones de categorías léxicas. A continuación, las palabras clave de sentimiento candidatas se identifican utilizando el análisis para los enlaces de textos de sentimientos.

Nota: Los términos del diccionario general compilado antes mencionado representan una lista de todas las palabras que puede que no sean interesantes o lingüísticamente ambiguas como unitérminos. Estas palabras se excluyen de la extracción cuando se están identificando los unitérminos. Sin embargo, volverán a evaluarse cuando determine las categorías léxicas no cuando busque palabras compuestas candidatas más largas (multitérminos).

Paso 3. Identificar las clases de equivalencias y la integración de sinónimos

Después de identificar los unitérminos y multitérminos candidatos, el software utiliza un diccionario de normalización para identificar las clases de equivalencias. Una clase de equivalencia es una forma base de una frase o una única forma de dos variantes de la misma frase. Para determinar qué concepto utilizar para la clase de equivalencia el motor de extracción aplica las reglas siguientes en el orden especificado:

- La forma especificada por el usuario en una biblioteca.
- La forma más frecuente, según lo definido por recursos compilados con anterioridad.

Paso 4. Asignación de tipo

A continuación, se asignan tipos a los conceptos extraídos. Un tipo es una agrupación semántica de conceptos. En este paso se utilizan tanto los recursos compilados como las bibliotecas. Los tipos incluyen elementos como conceptos de nivel más alto, palabras positivas y negativas, nombres propios, lugares, organizaciones, etc. Para obtener más información, consulte el tema [“Diccionarios de tipo” en la página 187](#).

Los sistemas lingüísticos son sensibles al conocimiento: cuanto mayor sea la cantidad de información contenida en sus diccionarios, mayor será la calidad de los resultados. La modificación del contenido del diccionario, como definiciones de sinónimos, puede simplificar la información resultante. A veces resulta un proceso iterativo, pero es necesario para una recuperación conceptual precisa. NLP es un elemento principal de IBM SPSS Modeler Text Analytics.

Cómo funciona la extracción

Durante la extracción de conceptos clave e ideas de sus respuestas, IBM SPSS Modeler Text Analytics se basa en el análisis de textos basado en lingüística. Este método ofrece la velocidad y la rentabilidad propia de los sistemas basados en estadística. Pero ofrece un grado de precisión mucho mayor, y con menos intervención humana. El análisis de texto basado en lingüística se basa en el ámbito de estudio conocido como proceso de idioma natural, conocido también como lingüística computacional.

Comprender el funcionamiento del proceso de extracción puede ayudarle a tomar decisiones clave a la hora de ajustar los recursos lingüísticos (bibliotecas, tipos, sinónimos, etc.). Estos son los pasos del proceso de extracción:

- Conversión de datos de origen en un formato estándar
- Identificar los términos candidatos
- Identificar las clases de equivalencias y la integración de sinónimos
- Asignar un tipo
- Indexación
- Extraer patrones y eventos de relaciones

Paso 1. Conversión de datos de origen en un formato estándar

En este primer paso, los datos que se importan se convierten en un formato uniforme que puede utilizarse para realizar otros análisis. Esta conversión se lleva a cabo internamente y no cambia los datos originales.

Paso 2. Identificar los términos candidatos

Es importante comprender el rol de los recursos lingüísticos en la identificación de los términos candidatos durante la extracción lingüística. Los recursos lingüísticos se utilizan siempre que se ejecuta una extracción. Existen en forma de plantillas, bibliotecas y recursos compilados. Las bibliotecas incluyen listas de palabras, relaciones y otra información que se utiliza para especificar o ajustar la extracción. Estos recursos compilados no se pueden ver ni editar. Sin embargo, los recursos restantes (plantillas) sí pueden editarse en el Editor de plantillas o, si se encuentra en una sesión de entorno de trabajo interactivo, en el Editor de recursos.

Los recursos compilados son los componentes internos principales del motor de extracción en IBM SPSS Modeler Text Analytics. Estos recursos incluyen un diccionario general que contiene una lista de formatos base con un código de categoría léxica (sustantivo, verbo, adjetivo, adverbio, participio, conjunción, determinante o preposición). Los recursos también incluyen tipos incorporados reservados que se utilizan para asignar muchos términos extraídos a los tipos siguientes: <Location>, <Organization>, o <Person>. Para obtener más información, consulte el tema [“Tipos incorporados”](#) en la página 188.

Además de los recursos compilados, se entregan varias bibliotecas con el producto y pueden utilizarse para complementar los tipos y las definiciones de conceptos en los recursos compilados, así como para ofrecer otros tipos y sinónimos. Estas bibliotecas (y las personalizadas que cree) se componen de varios diccionarios. Entre ellos están los diccionarios de tipo, los diccionarios de sustitución (sinónimos y elementos opcionales) y los diccionarios de exclusión. Para obtener más información, consulte el tema [Capítulo 15, “Trabajo con bibliotecas”](#), en la página 179.

Una vez importados y convertidos los datos, el motor de extracción empezará a identificar los términos candidatos para la extracción. Los términos candidatos son palabras o grupos de palabras que se utilizan para identificar conceptos en el texto. Durante el proceso del texto, las palabras simples (*unitérminos*) que no están en los recursos compilados se consideran como extracciones de términos candidatos. Las palabras compuestas candidatas (*multitérminos*) se identifican mediante extractores de patrones de categorías léxicas. Por ejemplo, el multitérmino `sports car`, que sigue el patrón de componentes del discurso "nombre adjetivo", tiene dos componentes. El multitérmino `fast sports car`, que sigue el patrón de componentes del discurso "nombre adjetivo adjetivo", tiene tres componentes.

Nota: Los términos del diccionario general compilado antes mencionado representan una lista de todas las palabras que puede que no sean interesantes o lingüísticamente ambiguas como unitérminos. Estas

palabras se excluyen de la extracción cuando se están identificando los unitérminos. Sin embargo, volverán a evaluarse cuando determine las categorías léxicas no cuando busque palabras compuestas candidatas más largas (multitérminos).

Por último, se utiliza un algoritmo especial para gestionar las cadenas de letras en mayúsculas, como cargos laborales, de manera que estos patrones especiales pueden extraerse.

Paso 3. Identificar las clases de equivalencias y la integración de sinónimos

Una vez identificados los unitérminos y los multitérminos candidatos, el software emplea un conjunto de algoritmos para compararlos y para identificar clases de equivalencias. Una clase de equivalencia es la forma básica de una frase o una forma simple de dos variantes de la misma frase. El propósito de asignar frases a clases de equivalencia es asegurarse de que, por ejemplo, `president of the company` y `company president` no se tratan como conceptos separados. Para determinar qué concepto utilizar para la clase de equivalencia, es decir, si se utiliza `president of the company` o `company president` como término principal, el motor de extracción aplica las siguientes reglas en el orden listado:

- La forma especificada por el usuario en una biblioteca.
- La forma más frecuente en el cuerpo completo del texto.
- La forma más corta en el cuerpo completo del texto (que generalmente se corresponde a la forma básica).

Paso 4. Asignación de un tipo

A continuación, se asignan tipos a los conceptos extraídos. Un tipo es una agrupación semántica de conceptos. En este paso se utilizan tanto los recursos compilados como las bibliotecas. Los tipos incluyen elementos como conceptos de nivel más alto, palabras positivas y negativas, nombres propios, lugares, organizaciones, etc. El usuario puede definir tipos adicionales. Para obtener más información, consulte el tema [“Diccionarios de tipo” en la página 187](#).

Paso 5. Indexación

Se crea el índice del conjunto completo de registros o documentos estableciendo una marca entre una posición de texto y el término representativo de cada clase de equivalencia. De esta manera se presupone que todos los casos de la forma declinada de un concepto candidato se indexa como forma básica candidata. Para cada forma básica se calcula la frecuencia global.

Paso 6. Extraer patrones y eventos de relaciones

IBM SPSS Modeler Text Analytics puede detectar no solamente tipos y conceptos, sino también las relaciones entre ellos. Hay varios algoritmos y bibliotecas disponibles en el producto que proporcionan la capacidad de extraer patrones de relaciones entre tipos y conceptos. Son particularmente útiles cuando se intentan detectar opiniones específicas (por ejemplo, reacciones ante productos) o los enlaces relacionales entre personas y objetos (por ejemplo, enlaces entre grupos políticos o genomas).

Cómo funciona la categorización

Al crear modelos de categoría en IBM SPSS Modeler Text Analytics, existen varias técnicas distintas que puede elegir para crear categorías. Puesto que cada conjunto de datos es exclusivo, el número de técnicas y el orden en el que las aplica puede cambiar con el tiempo. Puesto que su interpretación de los resultados puede ser diferente de la de otra persona, es posible que tenga que experimentar con las distintas técnicas para ver cuáles generan los mejores resultados para los datos de texto. En IBM SPSS Modeler Text Analytics, puede crear modelos de categoría en una sesión de área de trabajo en la que puede explorar y afinar las categorías más aún.

En esta guía, la **generación de categorías** hace referencia a la generación de definiciones de categoría y clasificación mediante el uso de una o más técnicas incorporadas, y **categorización** hace referencia al

proceso de puntuación o etiquetaje por el que se asignan identificadores exclusivos (nombre/ID/valor) a las definiciones de categorías para cada registro o documento.

Durante la generación de categorías, los conceptos y los tipos que se extrajeron se utilizan como los cimientos para las categorías. Cuando crea categorías, los registros o documentos se asignan automáticamente a categorías si contienen texto que coincida con un elemento de una definición de categoría.

IBM SPSS Modeler Text Analytics le ofrece varias técnicas de creación de categorías automatizadas para ayudarle a categorizar los documentos o registros rápidamente.

Agrupación de técnicas

Cada una de las técnicas disponibles resulta idónea para determinados tipos de datos y situaciones, pero a menudo conviene combinar técnicas en el mismo análisis para capturar el rango completo de documentos o registros. Puede ver un concepto en diversas categorías o detectar categorías redundantes.

Derivación de raíz de conceptos. Esta técnica crea categorías tomando un concepto y buscando otros conceptos que estén relacionados con el primero analizando si alguno de los componentes de los conceptos están morfológicamente relacionados o comparten raíces. Esta técnica es muy útil para identificar conceptos de palabras compuestas sinónimas, puesto que los conceptos de cada categoría generada son sinónimos o tienen un significado muy similar. Funciona con datos de extensión diversa y genera un número más reducido de categorías compactas. Por ejemplo, el concepto *opportunities to advance* se agruparía con los conceptos *opportunity for advancement* y *advancement opportunity*. Consulte el tema [“Derivación de raíz de conceptos”](#) en la página 110 para obtener más información.

Red semántica. Esta técnica comienza identificando los posibles sentidos de cada concepto a partir de un amplio índice de relaciones de palabras, y luego crea categorías agrupando los conceptos relacionados. Esta técnica resulta idónea cuando los conceptos son conocidos en la red semántica y no son muy ambiguos. Es menos idónea si el texto contiene terminología específica o jerga desconocida en la red. En un ejemplo, el concepto *granny smith apple* se podría agrupar con *gala apple* y *winesap apple* ya que son hermanos de Granny Smith. En otro ejemplo, el concepto *animal* podría agruparse con *cat* y *kangaroo* ya que son hipónimos de *animal*. En esta versión esta técnica está disponible solo para texto en inglés. Para obtener más información, consulte el tema [“Redes semánticas”](#) en la página 112.

Inclusión de conceptos. Esta técnica genera categorías agrupando los conceptos multitérmino (palabras compuestas) basándose en si contienen palabras que son subconjuntos o superconjuntos de una palabra en la otra. Por ejemplo, el concepto *seat* se agruparía con *safety seat*, *seat belt* y *seat belt buckle*. Para obtener más información, consulte el tema [“Inclusión de conceptos”](#) en la página 112.

Co-ocurrencia. Esta técnica crea categorías a partir de las coocurrencias que se encuentran en el texto. La idea radica en que cuando en los documentos y registros a menudo se encuentran conceptos o patrones de conceptos que aparecen juntos, esa co-ocurrencia refleja una relación subyacente que probablemente sea valiosa para las definiciones de categorías. Cuando la coocurrencia de algunas palabras es significativa, se crea una regla de coocurrencia que puede utilizarse como un descriptor de categoría para una nueva subcategoría. Por ejemplo, si muchos registros contienen las palabras *price* y *availability* (pero hay pocos registros que contengan solo una de las dos), estos conceptos se podrían agrupar en una regla de coocurrencia, (*price & available*) y asignarse a una subcategoría de la categoría *price* por ejemplo. Para obtener más información, consulte el tema [“Reglas de coocurrencia”](#) en la página 113.

Número mínimo de documentos. Para ayudar a determinar la relevancia de las coocurrencias, defina el número mínimo de registros documentos o registros que deben contener una coocurrencia determinada para que se utilice como descriptor en una categoría.

IBM SPSS Modeler Text Analytics nodos

Junto con los diversos nodos estándar proporcionados con IBM SPSS Modeler, también puede trabajar con nodos de minería de textos para incorporar el poder del análisis de texto en las rutas. IBM SPSS

Modeler Text Analytics le ofrece varios nodos de minería de textos para hacer precisamente eso. Estos nodos se almacenan en la pestaña IBM SPSS Modeler Text Analytics de la paleta de nodo.

Se incluyen los siguientes nodos:

- El **nodo fuente Lista de archivos** genera una lista de nombres de documentos como entrada al proceso de minería de textos. Esto es útil cuando el texto reside en un documento externo en lugar de en una base de datos o cualquier otro archivo estructurado. El nodo da como resultado un único campo con un registro para cada documento o carpeta indicado, que se puede seleccionar como entrada en un nodo de Minería de textos subsiguiente. Para obtener más información, consulte el tema [“Nodo Lista de archivos”](#) en la página 11.
- El **nodo fuente Canal de información web** hace posible leer texto de canales de información web, como pueden ser los blogs o los canales de información de noticias en formato RSS o HTML y utiliza estos datos en el proceso de minería de textos. El nodo da como resultado uno o más campos para cada registro encontrado en las fuentes, que se puede seleccionar como entrada en un nodo de Minería de textos subsiguiente. Para obtener más información, consulte el tema [“Nodo canal de información web”](#) en la página 13.
- El **nodo identificador de idioma de** es un nodo de proceso que explora el texto de origen para determinar en qué lenguaje humano está escrito y, a continuación, lo marca en un nuevo campo. Principalmente diseñado para ser utilizado con grandes cantidades de datos, este nodo resulta especialmente útil cuando tiene más de un idioma en los orígenes de datos. Para obtener más información, consulte el tema [“Nodo de idioma”](#) en la página 17.
- El **nodo Minería de textos** utiliza métodos lingüísticos para extraer conceptos clave de los textos, permite crear categorías con estos conceptos y otros datos y ofrece la posibilidad de identificar relaciones y asociaciones entre conceptos basados en patrones conocidos (llamado análisis de enlace de texto). Este nodo se puede utilizar para explorar el contenido de los datos de texto o para producir un modelo de concepto o de categoría. Tanto el concepto como la categoría se pueden combinar con datos estructurados existentes, como datos demográficos, y aplicarlos al modelado. Para obtener más información, consulte el tema [“Nodo de modelado de minería de textos”](#) en la página 20.
- El **nodo Análisis de enlaces de texto** extrae conceptos y también identifica relaciones entre conceptos basados en patrones conocidos dentro del texto. La extracción de patrones puede utilizarse para descubrir relaciones entre los conceptos, así como cualquier opinión o calificador adjunto a estos conceptos. El nodo Análisis de enlaces de texto ofrece una forma más directa de identificar y extraer patrones de los textos y después añadir los resultados de patrones en el conjunto de datos en la ruta. Pero también puede realizar AET al utilizar una sesión de área de trabajo interactiva en el nodo de modelado de minería de textos. Para obtener más información, consulte el tema [“Nodo Análisis para los enlaces de texto”](#) en la página 45.
- Cuando se realiza la minería de texto en documentos externos, el **nodo Salida de minería de textos** puede utilizarse para generar una página HTML que contenga enlaces a los documentos desde donde se extrajeron los conceptos. Para obtener más información, consulte el tema [“Nodo Visor de archivos”](#) en la página 53.

Aplicaciones

Por lo general, cualquiera que necesite de manera regular revisar grandes volúmenes de documentos para identificar elementos clave para su exploración más profunda puede beneficiarse de IBM SPSS Modeler Text Analytics.

Algunas aplicaciones específicas incluyen:

- **Investigación científica y médica.** Explorar materiales de investigación secundarios, como informes de patentes, artículos en periódicos y publicaciones de protocolo. Identificar asociaciones previamente desconocidas (como un doctor asociado a un producto en particular), presentar nuevas avenidas para la exploración adicional. Minimizar el tiempo empleado en el proceso de descubrimiento de una droga. Utilizarlo como una ayuda en la investigación genómica.
- **Investigación de inversión.** Revisar informes de análisis diarios, artículos de noticias y comunicados de prensa de empresas para identificar puntos de estrategia claves o cambios en el mercado. El análisis

de tendencias de este tipo de información revela problemas emergentes u oportunidades para una empresa o industria durante un período de tiempo.

- **Detección de fraudes.** Utilizado en el fraude a la banca o la asistencia médica para detectar anomalías y descubrir asuntos preocupantes en grandes cantidades de texto.
- **Investigación de mercado.** Utilizada en esfuerzos de investigación de mercado para identificar temas clave en respuestas a encuestas abiertas.
- **Análisis de canales de información web y blogs.** Explore y construya modelos utilizando ideas clave encontradas en canales de información, blogs, etc.
- **CRM.** Construya modelos utilizando datos de todos los puntos de contacto de los clientes, como ser los correos electrónicos, transacciones y encuestas.

Capítulo 2. Lectura de texto de origen

Los datos para la minería de textos pueden residir en cualquiera de los formatos estándar que utiliza IBM SPSS Modeler, incluyendo bases de datos u otros formatos "rectangulares" que representen datos en filas y columnas, o en formatos de documentos, como Microsoft Word, Adobe PDF, o HTML, que no se ajustan a esta estructura.

- Para leer en texto en documentos que no se ajustan a la estructura de datos estándar, incluyendo Microsoft Word, Microsoft Excel y Microsoft PowerPoint, además de Adobe PDF, XML, HTML y otros, el nodo Lista de archivos puede utilizarse para generar una lista de documentos o carpetas como entrada para el proceso de minería de textos. Para obtener más información, consulte [“Nodo Lista de archivos” en la página 11](#).
- Para leer en texto en canales de información web, como ser blogs en formato RSS o HTML, el nodo de canal de información web se puede utilizar para formatear los datos del canal de información web para entrada en el proceso de minería de textos. Para obtener más información, consulte [“Nodo canal de información web” en la página 13](#).
- Para leer en texto desde cualquiera de los formatos de datos estándar utilizados por SPSS Modeler como, por ejemplo, una base de datos con uno o más campos para comentarios de clientes, se puede utilizar cualquiera de los nodos de origen de SPSS Modeler. Si desea obtener más información, consulte la documentación del nodo SPSS Modeler.
- Cuando procese grandes cantidades de datos, que puede incluir texto en varios idiomas diferentes, utilice el nodo de idioma para identificar el idioma utilizado en un campo específico. Para obtener más información, consulte [“Nodo de idioma” en la página 17](#).

Nodo Lista de archivos

Para leer texto de documentos sin estructura guardados en formatos como Microsoft Word, Microsoft Excel y Microsoft PowerPoint, así como Adobe PDF, XML, HTML, y otros, el nodo Lista de archivos puede utilizarse para generar una lista de documentos o carpetas como entrada al proceso de minería de textos. Esto es necesario porque los documentos de texto sin estructura no pueden ser representados por campos ni registros (filas y columnas) de la misma forma en que lo son otros datos utilizados por IBM SPSS Modeler.

El nodo Lista de archivos funciona como un nodo de origen.

Puede encontrar este nodo en la pestaña IBM SPSS Modeler Text Analytics de la paleta de nodos, en la parte inferior de la ventana de IBM SPSS Modeler. Para obtener más información, consulte el tema [“IBM SPSS Modeler Text Analytics nodos” en la página 7](#).

Importante: No se admite ningún nombre de directorio ni de archivo que contenga caracteres no incluidos en la codificación local de la máquina. Al intentar ejecutar una ruta que contiene un nodo de lista de archivos, los nombres de archivo o directorio que contienen estos caracteres harán que la ejecución de la ruta falle. Esto se puede producir con nombres de directorio o nombres de archivo en un idioma extranjero como, por ejemplo, un nombre de archivo en alemán en un entorno local francés.

Soporte de datos locales. Si está conectado a un IBM SPSS Modeler Text Analytics Server remoto y tiene una ruta con un nodo de lista de archivos, los datos deberían residir en la misma máquina que el IBM SPSS Modeler Text Analytics Server – o debería asegurarse de que la máquina del servidor tiene acceso a la carpeta donde los datos de origen están almacenados en el nodo de lista de archivos.

Nota: No se puede utilizar el nodo Lista de archivos para la puntuación dentro de una configuración de IBM SPSS Collaboration and Deployment Services - Scoring.

Nodo Lista de archivos: pestaña Configuración

En esta pestaña, defina los directorios, las extensiones de archivo y la entrada para este nodo.

Nota: La extracción de minería de textos no puede procesar archivos Microsoft Office y Adobe PDF en plataformas que no sean Microsoft Windows. No obstante, siempre se pueden procesar archivos XML, HTML o texto.

No se admite ningún nombre de directorio ni de archivo que contenga caracteres no incluidos en la codificación local de la máquina. Al intentar ejecutar una ruta que contiene un nodo de lista de archivos, los nombres de archivo o directorio que contienen estos caracteres harán que la ejecución de la ruta falle. Esto se puede producir con nombres de directorio o nombres de archivo en un idioma extranjero como, por ejemplo, un nombre de archivo en alemán en un entorno local francés.

Directorio. Especifica la carpeta raíz que contiene los documentos que desea listar.

• **Incluir subdirectorios.** Especifica que los subdirectorios también deben explorarse.

Tipos de archivo a incluir en la lista: Puede seleccionar o deseleccionar los tipos de archivo y extensiones que desee utilizar. Si deselecciona una extensión de archivo, se ignoran los archivos con esa extensión. Puede filtrar por las siguientes extensiones:

Tabla 1. Filtros de tipo de archivo por extensión de archivo			
• .rtf, .doc, .docx, .doc m	• .xls, .xlsx, .xls m	• .ppt, .pptx, .ppt m	• .txt, .text
• .htm, .html, .shtml	• .xml	• .pdf	• . \$

Nota: Para obtener más información, consulte “Nodo Lista de archivos” en la página 11.

Si tiene archivos sin extensión o con una extensión de punto final (por ejemplo, File01 o File01.), utilice la opción **Sin extensión** para seleccionarlos.

Solo genera nombres de vía de acceso de documentos. Seleccione esta opción si el campo de salida contendrá uno o más nombres de vías de acceso para la ubicación donde residen los documentos.

Codificación de entrada. Si el campo de salida contendrá texto exacto, elija el valor relevante de la lista siguiente:

- Automático (europeo)
- UTF-8
- UTF-16
- ISO-8859-1
- ISO-8859-2
- Windows-1250
- US ASCII

La salida se muestra como texto de documento UTF-8.

Nodo Lista de archivos: Otras pestañas

La pestaña Tipos es una pestaña estándar en nodos de IBM SPSS Modeler, como lo es la pestaña Anotaciones.

Utilización del nodo Lista de archivos en minería de textos

El nodo Lista de archivos se utiliza cuando los datos de texto residen en documentos no estructurados externos en formatos como, por ejemplo, Microsoft Word, Microsoft Excel y Microsoft PowerPoint, así como Adobe PDF, XML, HTML, y otros.

Como un ejemplo, supongamos que hemos conectado un nodo Lista de archivos a un nodo Minería de textos para proporcionar texto que reside en documentos externos:

1. **Nodo Lista de archivos (pestaña Configuración).** En primer lugar, añadimos este nodo a la ruta para especificar dónde están almacenados los documentos de texto. Seleccionamos este directorio que contiene todos los documentos en los que deseamos realizar la minería de textos.
2. **Nodo Minería de textos (pestaña Campos).** A continuación, añadimos y conectamos un nodo Minería de textos al nodo Lista de archivos. En este nodo, definimos nuestro formato de entrada, plantilla de recursos y formato de salida. Hemos seleccionado el nombre de campo generado en el nodo Lista de archivos, el campo de texto y otros valores. Para obtener más información, consulte el tema [“Utilización del nodo Minería de texto en una ruta”](#) en la página 29.

Para obtener más información sobre la utilización del nodo Minería de textos, consulte [“Nodo de modelado de minería de textos”](#) en la página 20.

Nodo canal de información web

El nodo de canal de información web puede utilizarse para preparar datos de texto desde canales de información web para el proceso de minería de textos. Este nodo acepta canales de información web en dos formatos:

- **Formato RSS.** RSS es un formato simple estandarizado basado en XML para contenido web. El URL para este formato apunta a una página que tiene un conjunto de artículos enlazados como, por ejemplo, fuentes de noticias sindicadas y blogs. Puesto que RSS es un formato estandarizado, cada artículo enlazado se identifica automáticamente y se lo trata como un registro separado en la ruta de datos resultante. No se necesita ninguna entrada adicional para que pueda identificar los datos de texto importantes y los registros del canal de información a no ser que desee aplicar una técnica de filtrado al texto.
- **Formato HTML.** Puede definir uno o más URL a páginas HTML en la pestaña Entrada. A continuación, en la pestaña Registros, defina la etiqueta de inicio de registro e identifique las etiquetas que delimitan el contenido de destino y asigne esas etiquetas a los campos de salida de su elección (descripción, título, fecha de modificación, etc.). Para obtener más información, consulte el tema [“Nodo Canal de información web: pestaña Registros”](#) en la página 14.

Importante: Si está intentando recuperar información sobre la web a través de un servidor proxy, debe habilitar el servidor proxy en el archivo `net.properties` tanto para el cliente como para el servidor de IBM SPSS Modeler Text Analytics. Siga las instrucciones que se detallan en este archivo. Esto se aplica cuando se accede a la web a través del nodo Canal de información de la web o cuando se recupera una licencia de software como servicio (SaaS) de SDL, ya que estas conexiones pasan por Java™. Este archivo se encuentra en `C:\Program Files\IBM\SPSS\Modeler\18.4.0\jre\lib\net.properties` de forma predeterminada.

La salida de este nodo es un conjunto de campos utilizados para describir los registros. El campo **Descripción** se utiliza más comúnmente ya que contiene la mayor parte del contenido del texto. Sin embargo, también pueden interesarle otros campos, como la descripción corta de un registro (campo **Desc corta**) o el título del registro (campo **Título**). Cualquiera de los campos de salida pueden seleccionarse como entrada para un nodo de Minería de textos subsiguiente.

Nota: No se puede utilizar el nodo canal de información web para la puntuación dentro de una configuración de IBM SPSS Collaboration and Deployment Services - Scoring.

Puede encontrar este nodo en la pestaña IBM SPSS Modeler Text Analytics de la paleta de nodos, en la parte inferior de la ventana de IBM SPSS Modeler. Para obtener más información, consulte el tema [“IBM SPSS Modeler Text Analytics nodos”](#) en la página 7.

Nodo canal de información web: pestaña Entrada

La pestaña Entrada se utiliza para especificar una o más direcciones Web o URL, a fin de capturar los datos de texto. En el contexto de la minería de textos, puede especificar los URL para los canales de información que contienen datos de texto.

Importante: Cuando trabaje con datos no RSS, puede preferir utilizar una herramienta de rastreo web, como WebQL®, para automatizar la recopilación de contenido y, a continuación, hacer referencia a la salida de esa herramienta utilizando un nodo fuente distinto.

Puede establecer los siguientes parámetros:

Escriba o pegue los URL. En este campo, puede escribir o pegar uno o más URL. Si está entrando más de uno, especifique sólo uno por línea y utilice la tecla **Intro/Retorno** para separar líneas. Introduzca la vía de acceso de URL completa al archivo. Estos URL pueden ser para canales de información en uno de los dos siguientes formatos:

- **RSS format.** RSS es un formato simple estandarizado basado en XML para contenido web. El URL para este formato apunta a una página que tiene un conjunto de artículos enlazados como, por ejemplo, fuentes de noticias sindicadas y blogs. Puesto que RSS es un formato estandarizado, cada artículo enlazado se identifica automáticamente y se lo trata como un registro separado en la ruta de datos resultante. No se necesita ninguna entrada adicional para que pueda identificar los datos de texto importantes y los registros del canal de información a no ser que desee aplicar una técnica de filtrado al texto.
- **HTML format.** Puede definir uno o más URL a páginas HTML en la pestaña Entrada. A continuación, en la pestaña Registros, defina la etiqueta de inicio de registro e identifique las etiquetas que delimitan el contenido de destino y asigne esas etiquetas a los campos de salida de su elección (descripción, título, fecha de modificación, etc.). Cuando trabaje con datos no RSS, puede preferir utilizar una herramienta de rastreo web, como WebQL®, para automatizar la recopilación de contenido y, a continuación, hacer referencia a la salida de esa herramienta utilizando un nodo fuente distinto. Para obtener más información, consulte el tema [“Nodo Canal de información web: pestaña Registros”](#) en la [página 14](#).

Número de entradas más recientes a leer por URL. Este campo especifica el número máximo de registros a leer por cada URL listado en el campo comenzando con el primer registro encontrado en el canal de información. La cantidad de texto afecta a la velocidad de proceso durante la extracción en sentido descendente en un nodo de Minería de textos o de Análisis de enlaces de texto.

Guarde y vuelva a utilizar canales de información web cuando sea posible. Con esta opción, se exploran canales de información web y el resultado procesado es almacenado en la memoria caché. A continuación, tras la ejecución de rutas subsecuentes, si los contenidos de un determinado canal de información no cambiaron o si no se puede acceder al canal de información (interrupción de Internet, por ejemplo), la versión almacenada en la memoria caché se utiliza para acelerar el tiempo de procesamiento. Cualquier contenido nuevo descubierto en estos canales de información también se almacena en la memoria caché para la próxima vez que se ejecute el nodo.

- **Etiqueta.** Si selecciona **Guardar y reutilizar canales de información web previos siempre que sea posible**, debe especificar un nombre de etiqueta para los resultados. Esta etiqueta se utiliza para describir los canales de información almacenados en la memoria caché en el servidor. Si no se especifica ninguna etiqueta o esta no se reconoce, no será posible su reutilización.

Nodo Canal de información web: pestaña Registros

La pestaña Registros se utiliza para especificar el contenido del texto de canales de información no RSS al identificar dónde comienza cada registro nuevo, como también otra información relevante respecto a cada registro. Si desea saber que un canal de información no RSS (HTML) contiene texto que se encuentra en varios registros, debe identificar la etiqueta de inicio del registro aquí o sino el texto se tratará como un solo registro. Mientras que los canales de información RSS están estandarizados y no necesitan ninguna especificación de etiqueta en esta pestaña, aún puede obtener una vista previa del contenido en la pestaña Vista previa.

Importante: Cuando trabaje con datos no RSS, puede preferir utilizar una herramienta de rastreo web, como WebQL®, para automatizar la recopilación de contenido y, a continuación, hacer referencia a la salida de esa herramienta utilizando un nodo fuente distinto.

URL. Esta lista desplegable contiene una lista de los URL especificados en la pestaña Entrada. Están presentes tanto los canales de información con formato HTML como los RSS. Si la dirección URL

es demasiado larga para la lista desplegable, se recortará automáticamente en el medio utilizando una elipsis para sustituir el texto recortado, como por ejemplo <http://www.ibm.com/ejemplo/inicio-de-dirección...resto-de-dirección/víadeacceso.htm>.

- Con **canales de información de formato HTML**, si el canal contiene más de un registro (o entrada), puede definir cuáles etiquetas HTML contienen los datos correspondientes a los campos que se muestran en la tabla. Por ejemplo, puede definir la etiqueta de inicio que indica que un registro nuevo se ha iniciado, una etiqueta de fecha modificada o el nombre de un autor.
- Con **canales de información con formato RSS**, no se le solicitará que especifique ninguna etiqueta ya que RSS es un formato estandarizado. Sin embargo, puede visualizar resultados de ejemplo en la pestaña Vista previa si así lo desea. Todos los canales de información RSS reconocidos van precedidos por la imagen del logotipo de RSS.

Pestaña Fuente. En esta pestaña, puede ver el código fuente de cualquier canal de información HTML. Este código no es editable. Puede utilizar el campo Encontrar para ubicar etiquetas o información específica en esta página, que después puede copiar y pegar en la tabla de abajo. El campo Encontrar no es sensible a las mayúsculas y minúsculas y hará coincidir series parciales.

Pestaña Vista previa. En esta pestaña, puede previsualizar cómo un nodo Canal de información web leerá un registro. Esto es particularmente útil para canales de información HTML ya que puede cambiar la forma en que se lee un registro al definir etiquetas HTML en la tabla debajo de la pestaña Vista previa.

Etiqueta de inicio de registro no RSS. Esta opción sólo es aplicable a canales de información no RSS. Si el canal de información HTML contiene varios textos que desea fraccionar en varios registros, especifique la etiqueta HTML que señala el inicio de un registro (como puede ser un artículo o entrada de blog) aquí. Si no define una para un canal de información no RSS, Modeler intentará adivinar el formato XML y devolver registros correspondientes. Si Modeler no puede adivinar el formato XML, no se devolverá nada. Si su objetivo es importar todo el contenido de una página y procesarlo más tarde, le recomendamos utilizar lectores XML separados con potentes funciones y, después, importar el resultado en Modeler Text Analytics.

Tabla Campo. Esta opción sólo es aplicable a canales de información no RSS. En esta tabla, puede fraccionar el contenido del texto en campos de salida específicos al especificar una etiqueta de inicio para cualquiera de los campos de salida predefinidos. Introduzca sólo la etiqueta de inicio. Todas las coincidencias se realizan al analizar el HTML y hacer coincidir los contenidos de la tabla con los nombres y atributos de las etiquetas encontradas en el HTML. Puede utilizar los botones al final para copiar las etiquetas que definió y volver a utilizarlas para otros canales de información.

Tabla 2. Posibles campos de salida para canales de información no RSS (formatos HTML)	
Nombre de campo de salida	Contenido de etiqueta esperado
Título	La etiqueta que delimita el título del registro. (opcional)
Descripción breve	La etiqueta que delimita la descripción corta o etiqueta. (opcional)
Descripción	La etiqueta que delimita el texto principal. Si se deja en blanco, este campo contendrá el resto del contenido en la etiqueta <body> (si hay un único registro) o el contenido que se encuentra dentro del registro actual (cuando un delimitador de registro se ha especificado).
Author	La etiqueta que delimita al autor del texto. (opcional)
Colaboradores	La etiqueta que delimita los nombres de los colaboradores. (opcional)
Fecha publicada	La etiqueta que delimita la fecha en la que se publicó el texto. Si se deja en blanco, este campo contendrá la fecha en la que el nodo lee los datos.
Fecha de modificación	La etiqueta que delimita la fecha en la que el texto fue modificado. Si se deja en blanco, este campo contendrá la fecha en la que el nodo lee los datos.

Cuando especifica una etiqueta en la tabla, el canal de información se explora mediante esta etiqueta como la etiqueta mínima para hacer coincidir en lugar de una coincidencia exacta. Es decir, si ha especificado `<div>` para el campo Título, esto coincidiría con cualquier código `<div>` del canal de información, incluidos aquellos con atributos especificados (como `<div class="post three">`), de forma que `<div>` sea igual al código raíz (`<div>`) y cualquier derivado que incluya un atributo y utilice dicho contenido para el campo de salida Título. Si especificó una etiqueta raíz, los demás atributos también se incluyen.

Tabla 3. Ejemplos de etiquetas HTML utilizadas identifican el texto para los campos de salida			
Si especifica:	Coincidiría con:	Y también con:	Pero no coincidiría con:
<code><div></code>	<code><div></code>	<code><div class="post"></code>	cualquier otra etiqueta
<code><p class="auth"></code>	<code><p class="auth"></code>	<code><p color="black" class="auth" id="85643"></code>	<code><p color="black"></code>

Nodo canal de información web: pestaña Filtro de contenido

La pestaña Filtro de contenido se utiliza para aplicar una técnica de filtrado al contenido del canal de información RSS. Esta pestaña no se aplica a canales HTML. Quizá que desee aplicar el filtro si el canal de información contiene una gran cantidad de texto en forma de cabeceras, pies, menús, publicidad, etc. Puede utilizar esta pestaña para eliminar etiquetas HTML no deseadas, JavaScript y palabras o líneas cortas del contenido.

Filtrado de contenido. Si no desea aplicar una técnica de limpieza, seleccione **Ninguna**. De lo contrario, seleccione **Limpiador de contenido RSS**.

Opciones del limpiador de contenidos RSS. Si selecciona **Limpiador de contenido RSS**, puede elegir descartar líneas en base a ciertos criterios. Una línea se delimita mediante una etiqueta HTML como `<p>` y `<li>` pero excluyendo etiquetas en línea como `<span>`, `<b>` y `<font>`. Tenga en cuenta que las etiquetas `<br>` se procesan como saltos de línea.

- **Descartar líneas cortas.** Esta opción ignora líneas que no contienen el **número mínimo de palabras** definido aquí.
- **Descartar líneas con palabras cortas.** Esta opción ignora líneas que tiene más del **largo de palabra promedio mínimo** definido aquí.
- **Descartar líneas con muchas palabras de un solo carácter.** Esta opción ignora líneas que contienen más de cierta **proporción de palabras de un solo carácter**.
- **Descartar líneas que contienen etiquetas específicas.** Esta opción ignora texto en líneas que contienen cualquiera de las etiquetas especificadas en el campo.
- **Descartar líneas que contienen texto específico.** Esta opción ignora líneas que contienen cualquier texto especificado en el campo.

Utilización del nodo Canal de información web en Minería de textos

El nodo Canal de información web puede utilizarse para preparar datos de texto de canales de información web en Internet para el proceso de minado de texto. Este nodo acepta canales de información web en formato HTML o RSS. Estos canales de información sirven como entrada al proceso de minado de textos (un nodo subsiguiente de Minería de textos o Análisis de enlaces de texto).

Si utiliza el nodo Canal de información web, debe asegurarse de especificar que el campo Texto representa **texto real** en el nodo Minería de textos o Análisis de enlaces de texto para indicar que estos canales de información se enlazan directamente con cada artículo o entrada de blog.

Importante: Si está intentando recuperar información sobre la web a través de un servidor proxy, debe habilitar el servidor proxy en el archivo `net.properties` tanto para el cliente como para el servidor de IBM SPSS Modeler Text Analytics. Siga las instrucciones que se detallan en este archivo. Esto se

aplica cuando se accede a la web a través del nodo Canal de información de la web o cuando se recupera una licencia de software como servicio (SaaS) de SDL, ya que estas conexiones pasan por Java. Este archivo se encuentra en `C:\Program Files\IBM\SPSS\Modeler\18.4.0\jre\lib\net.properties` de forma predeterminada.

Ejemplo: nodo Canal de información web (canal de información RSS) con el nodo de modelado de Minería de textos

Como ejemplo, suponga que conectamos un nodo Canal de información web a un nodo Minería de textos para suministrar datos de texto desde un canal de información RSS al proceso de minería de textos.

1. **Nodo Canal de información web (pestaña Entrada).** Primero, añadimos este nodo a la ruta para especificar donde se ubican los contenidos del canal de información y para verificar la estructura del contenido. En la primera pestaña, proporcionamos el URL al canal de información RSS. Ya que nuestro ejemplo es para un canal de información RSS, el formato ya está definido y no necesitamos hacer cambios en la pestaña Registros. Hay disponible un algoritmo de filtrado de contenido opcional para canales de información RSS, sin embargo, en este caso no se aplicó.
2. **Nodo Minería de textos (pestaña Campos).** A continuación, añadimos y conectamos un nodo Minería de textos al nodo Canal de información web. En esta pestaña, definimos la salida del campo de texto por el nodo Canal de información web. En este caso, queríamos utilizar el campo **Descripción**. También seleccionamos la opción campo Texto representa **texto real**, como también otros valores.
3. **Nodo Minería de textos (pestaña Modelo).** A continuación, en la pestaña Modelo, seleccionamos el modo compilación y los recursos. En este ejemplo, elegimos compilar un modelo de concepto directamente desde este nodo utilizando la plantilla de recursos predeterminada.

Para obtener más información sobre la utilización del nodo Minería de textos, consulte [“Nodo de modelado de minería de textos”](#) en la página 20.

Nodo de idioma

Puede utilizar el nodo de idioma para identificar el idioma natural de un campo de texto en los datos de origen.

La salida de este nodo es un campo derivado que contiene el código de idioma detectado.

Nota: No se puede utilizar el nodo de idioma para la puntuación dentro de una configuración de IBM SPSS Collaboration and Deployment Services - Scoring.

Puede encontrar este nodo en la pestaña IBM SPSS Modeler Text Analytics de la paleta de nodos, en la parte inferior de la ventana de IBM SPSS Modeler. Para obtener más información, consulte el tema [“IBM SPSS Modeler Text Analytics nodos”](#) en la página 7.

Nodo de idioma: Pestaña Valores

En esta pestaña puede especificar cómo generar la salida de detalles de idioma para un campo de texto seleccionado.

Campo de texto Seleccione el campo de texto para el que desee identificar el idioma.

Nombre del campo Derivar Especifique un nombre para el campo derivado que contendrá el código del idioma detectado. El valor predeterminado es *Idioma*.

Valor predeterminado para cuando no pueda identificarse el idioma Especifique el nombre del campo que deba crearse si el idioma no puede identificarse. Las opciones disponibles son:

- **Undefined** Si se selecciona, el campo derivado contiene valores nulos.
- **Supported** Si se selecciona, puede elegir uno de los siguientes idiomas ISO soportados:
 - Inglés (EN)
 - Alemán (DE)
 - Español (ES)

- Francés (FR)
- Italiano (IT)
- Holandés (NL)
- Portugués (PT)
- **Custom** Si no hay ningún idioma soportado que resulte adecuado, utilice esta opción para especificar que debe utilizarse un valor personalizado. Normalmente, puede ser un código de idioma ISO, de dos letras, pero puede ser cualquier cadena de texto que necesite.

Capítulo 3. Minería para conceptos y categorías

El nodo de modelado de minería de textos se utiliza para generar uno de dos nuggets de modelo de minería de textos:

- Los *nuggets de modelo de concepto* descubren y extraen conceptos destacados de sus datos de texto estructurados o no estructurados.
- Los *nuggets de modelo de categoría* puntúan y asignan documentos y registros a categorías, que se componen de los conceptos extraídos (y patrones).

Los conceptos y patrones extraídos, así como las categorías de sus nuggets de modelo, pueden combinarse con datos estructurados existentes como, por ejemplo, demográficos, y aplicarse utilizando el suite completo de herramientas de IBM SPSS Modeler para tomar decisiones mejores y más certeras. Por ejemplo, si los clientes listan con frecuencia problemas de inicio de sesión como el impedimento primario para completar tareas de gestión de cuentas en línea, puede que desee incorporar “problemas de inicio de sesión” a sus modelos.

Adicionalmente, el nodo de modelado de minería de textos está totalmente integrado en IBM SPSS Modeler para que pueda desplegar rutas de minería de textos a través de IBM SPSS Modeler Solution Publisher para la puntuación en tiempo real de datos no estructurados en aplicaciones como PredictiveCallCenter. La posibilidad de desplegar estas rutas garantiza implementaciones correctas de minería de textos de circuito cerrado. Por ejemplo, su organización ahora puede analizar anotaciones de llamadas entrantes o salientes al aplicar los modelos predictivos para incrementar la precisión del mensaje de marketing en tiempo real. Se ha demostrado que la utilización de resultados de modelo de minería de textos en rutas mejora la precisión de modelos de datos predictivos.

Para ejecutar IBM SPSS Modeler Text Analytics con IBM SPSS Modeler Solution Publisher, añada el directorio `<install_directory>/ext/bin/spss.TMWBServer` a la variable de entorno `$LD_LIBRARY_PATH`.

En IBM SPSS Modeler Text Analytics, a menudo se hace referencia a conceptos y categorías extraídos. Es importante comprender el significado de los conceptos y categorías ya que pueden ayudar a tomar decisiones más informadas durante su trabajo de exploración y creación de modelos.

Conceptos y nuggets de modelo de concepto

Durante el proceso de extracción, los datos de texto se exploran y analizan para identificar palabras individuales interesantes o relevantes, como `election` o `peace`, y frases como `presidential election`, `election of the president` o `peace treaties`. Estas palabras y frases se conocen colectivamente con el nombre de *términos*. Mediante los recursos lingüísticos, los términos relevantes se extraen y los términos similares se agrupan bajo un término principal llamado **concepto**.

De esta forma, un concepto podría representar varios términos subyacentes en función de su texto y del conjunto de recursos lingüísticos que esté utilizando. Por ejemplo, supongamos que tenemos una encuesta de satisfacción de empleados y que el concepto `salary` se ha extraído. Supongamos también que cuando revisó los registros asociados con `salary`, notó que `salary` no siempre está presente en el texto pero que, en su lugar, determinados registros contenían algo similar como, por ejemplo, los términos `wage`, `wages` y `salaries`. Estos términos se agrupan bajo `salary` ya que el motor de extracción consideró que eran similares o determinó que eran sinónimos basándose en reglas de procesamiento o recursos lingüísticos. En este caso, cualquier documento o registro que contuviese uno de estos términos sería considerado como si contuviese la palabra `salary`.

Si desea ver los términos agrupados bajo un concepto, puede explorar el concepto dentro de un área de trabajo interactiva o buscar los sinónimos que se muestran en el modelo de concepto. Para obtener más información, consulte el tema [“Términos subyacentes en modelos de concepto”](#) en la página 33.

Un **nugget de modelo de concepto** contiene un conjunto de conceptos que puede utilizarse para identificar registros o documentos que también contengan el concepto (incluido cualquiera de sus sinónimos o términos agrupados). Un modelo de concepto puede utilizarse de dos formas. La primera

sería explorar y analizar los conceptos descubiertos en el texto fuente original o identificar rápidamente documentos de interés. La segunda sería aplicar este modelo a documentos o registros de texto nuevos para identificar rápidamente los mismos conceptos clave en los nuevos documentos/registros como, por ejemplo, el descubrimiento en tiempo real de conceptos clave en datos de anotación de un centro de asistencia telefónica.

Para obtener más información, consulte el tema [“Nugget de minería de textos: Modelo de concepto”](#) en la página 30.

Categorías y nuggets de modelo de categoría

Puede crear **categorías** que representen, en esencia, conceptos de nivel superior o temas para capturar las ideas clave, conocimientos y actitudes expresadas en el texto. Las categorías se componen de conjuntos de descriptores como, por ejemplo, *conceptos*, *tipos* y *reglas*. En conjunto, estos descriptores se utilizan para identificar si un registro o documento pertenece a una determinada categoría. Un documento o registro puede explorarse para ver si su texto coincide con un descriptor. Si se encuentra una coincidencia, el documento/registro se asigna a esa categoría. Este proceso se denomina **categorización**.

Las categorías pueden crearse automáticamente utilizando el conjunto sólido de técnicas automatizadas del producto, utilizando manualmente conocimientos adicionales relacionados con los datos, o una combinación de ambos. También puede cargar un conjunto de categorías creadas previamente desde un paquete de análisis de texto a través de la pestaña Modelo de este nodo. La creación manual de categorías o el perfeccionamiento de categorías sólo puede llevarse a cabo a través del área de trabajo interactiva. Para obtener más información, consulte el tema [“Nodo Minería de textos: pestaña Modelo”](#) en la página 23.

Un **nugget de modelo de categoría** contiene un conjunto de categorías junto con sus descriptores. El modelo puede utilizarse para categorizar un conjunto de documentos o registros en función del texto de cada documento/registro. Se lee cada documento o registro y, a continuación, se asigna a cada categoría para la que se haya encontrado una coincidencia de descriptor. De este modo, un documento o registro podría asignarse a más de una categoría. Puede utilizar nuggets de modelo de categoría para ver las ideas esenciales en respuestas de encuestas abiertas o en un conjunto de entradas de blog, por ejemplo.

Para obtener más información, consulte el tema [“Nugget de minería de textos: Modelo de categoría”](#) en la página 38.

Nodo de modelado de minería de textos

El nodo de Minería de textos utiliza técnicas lingüísticas y de frecuencia para extraer conceptos clave del texto y crear categorías con estos conceptos y otros datos. El nodo se puede utilizar para explorar los contenidos de datos de texto o para producir ya sea un nugget de modelo de concepto o de categoría. Cuando ejecuta este nodo de modelado, un motor de extracción lingüística interno extrae y organiza los conceptos, patrones o categorías utilizando métodos de procesamiento del lenguaje natural.

Puede ejecutar el nodo de Minería de textos y producir automáticamente un nugget de modelo concepto o categoría utilizando la opción **Generar directamente**. De forma alternativa, puede utilizar un método de exploración más práctico al utilizar la modalidad **Compilación interactiva** en la que no sólo puede extraer conceptos, crear categorías y refinar los recursos lingüísticos, sino que también puede realizar análisis de enlaces de texto y explorar clústeres. Para obtener más información, consulte el tema [“Nodo Minería de textos: pestaña Modelo”](#) en la página 23.

Puede encontrar este nodo en la pestaña IBM SPSS Modeler Text Analytics de la paleta de nodos, en la parte inferior de la ventana de IBM SPSS Modeler. Para obtener más información, consulte el tema [“IBM SPSS Modeler Text Analytics nodos”](#) en la página 7.

Requisitos. Los nodos de modelado de Minería de textos aceptan datos de texto de un nodo de Canal de información web, de un nodo Lista de archivos o cualquiera de los nodos fuente estándares. Este nodo se instala con IBM SPSS Modeler Text Analytics y se puede acceder a él a la paleta IBM SPSS Modeler Text Analytics.

Nota: Este nodo sustituye el nodo Extracción de texto, que se ofrecía en versiones antiguas del producto. Si tiene rutas más antiguas que utilizan nodos o nuggets de modelo antiguos, debe volver a crear las rutas utilizando el nodo Minería de textos.

Nodo Minería de textos: pestaña Campos

Utilice la pestaña Campos para especificar los valores de campo para los datos de los que se van a extraer conceptos. Considere la posibilidad de utilizar un nodo de Muestra en sentido ascendente desde este nodo cuando trabaje con conjuntos de datos más grandes para acelerar los tiempos de procesamiento. Para obtener más información, consulte el tema [“Muestreo en sentido ascendente para ahorrar tiempo” en la página 29.](#)

Puede establecer los siguientes parámetros:

Campo de ID Seleccione el campo que contiene el identificador para los registros de texto. Los identificadores deben ser enteros. El campo de ID funciona como un índice para los registros de texto individuales. Utilice un campo de ID si el campo de texto representa el texto a ser minado.

Campo de texto. Seleccione el campo que contiene el texto donde se va a realizar la minería. Este campo depende del origen de datos.

Campo de idioma Seleccione el campo que contenga el identificador de idioma ISO, de dos letras. Si no selecciona ningún campo, se presupone que el idioma de cada documento es de la plantilla proporcionada.

Tipo de documento. El tipo de documento especifica la estructura del texto. Seleccione uno de los siguientes tipos:

- **Texto completo.** Se utiliza para la mayoría de los documentos o fuentes de texto. El conjunto de texto completo se explora para su extracción. A diferencia de las otras opciones, no hay valores adicionales para esta opción.
- **Texto estructurado.** Se utiliza para formularios bibliográficos, patentes y cualquier campo que contenga estructuras regulares que puedan identificarse y analizarse. Este tipo de documento se utiliza para omitir todo o parte del proceso de extracción. Le permite definir separadores de términos, asignar tipos e imponer un valor de frecuencia mínima. Si selecciona esta opción, debe pulsar el botón **Valores** y especificar separadores de texto en **Formato de texto estructurado.** del cuadro de diálogo Valores del documento. Para obtener más información, consulte el tema [“Pestaña Valores de documento para campos” en la página 22.](#)

Unidad textual. Seleccione la modalidad de extracción de las siguientes:

- **Modalidad de documento.** Se utiliza para documentos cortos y homogéneos semánticamente, como los artículos de las agencias de noticias.
- **Modalidad de párrafo.** Se utiliza para las páginas web y para los documentos sin etiquetar. El proceso de extracción divide semánticamente los documentos, tomando ventaja de características como etiquetas internas y sintaxis. Si se selecciona esta modalidad, la puntuación se aplica párrafo a párrafo. Por lo tanto, por ejemplo, la regla `apple & orange` es verdadera sólo si `apple` y `orange` se encuentran en el mismo párrafo.

Nota: Debido a la forma de extraer el texto de los documentos en PDF, la **Modalidad de párrafo** no funciona en estos documentos. Esto se debe a que la extracción suprime el marcador de retorno de carro.

Configuración modalidad párrafo. Esta opción solo está disponible si establece la opción de unidad textual en **Modalidad de párrafo.** Especifique el umbral de carácter a utilizar en cualquier extracción. El tamaño actual se redondea hacia arriba o hacia abajo hacia el punto más próximo. Para asegurar que las asociaciones de palabras producidas desde el texto de la colección de documentos son representativas, evite especificar un tamaño de extracción que sea demasiado pequeño.

- **Mínimo.** Especifique el número mínimo de caracteres a utilizar en cualquier extracción.
- **Máximo.** Especifique el número máximo de caracteres a utilizar en cualquier extracción.

Modalidad de partición Utilice la modalidad de partición para elegir si particionar basado en la configuración del nodo de tipo o si seleccionar otra partición. El particionamiento separa los datos en muestras de entrenamiento y prueba.

Pestaña Valores de documento para campos

Formateo de texto estructurado

Si desea omitir todas o parte del proceso de extracción debido a que tiene datos estructurados o desea imponer reglas sobre cómo manejar el texto, utilice la opción de tipo de documento **Texto estructurado** y declare los campos o etiquetas que contiene el texto en la sección **Formateo de texto estructurado** del cuadro de diálogo Configuración de documentos. Los términos extraídos se derivan sólo del texto contenido dentro de los campos o etiquetas declaradas (y etiquetas hijo). Cualquier campo o etiqueta no declarada será ignorada.

En determinados contextos, el procesamiento lingüístico no es necesario y el motor de extracción lingüística puede sustituirse por declaraciones explícitas. En un archivo bibliográfico donde los campos de palabras claves se separan con separadores como punto y coma (;) o coma (,), resulta insuficiente extraer la serie entre dos separadores. Por este motivo, puede suspender el proceso de extracción completo y en su lugar definir reglas de manejo especiales para declarar separadores de términos, asignar tipos al texto extraído o imponer un recuento de frecuencia mínima para extracciones.

Utilice las siguientes reglas al declarar elementos de texto estructurado:

- Sólo se puede declarar un campo, etiqueta o elemento por línea. No tienen por qué estar presentes en los datos.
- Las declaraciones son sensibles a las mayúsculas y minúsculas.
- Si declara una etiqueta que tiene atributos, como por ejemplo `<title id="1234">`, y desea incluir todas las variaciones o, en este caso, todos los ID, añada la etiqueta sin el atributo o el corchete angular de cierre (`>`), como por ejemplo `<title`
- Añada una coma después del nombre de campo o etiqueta para indicar que se trata de texto estructurado. Añada esta coma directamente después del campo o etiqueta pero antes de cualquier separador, tipo o valores de frecuencia, como `author:` o `<place>:`.
- Para indicar que varios términos están contenidos en el campo o etiqueta en el que se utiliza un separador para designar los términos individuales, declare el separador después de la coma, como en `author: ,` o `<section>;`.
- Para asignar un tipo al contenido que se encuentra en la etiqueta, declare el nombre del tipo después de la coma y un separador, como `author: , Person` o `<place>; Location`. Declare el tipo utilizando los nombres como aparecen en el Editor de recursos.
- Para definir un recuento de frecuencia mínima para un campo o etiqueta, declare un número al final de la línea como en `author: , Person1` o `<place>; Location5`. Donde `n` es el recuento de frecuencia que definió, términos encontrados en un campo o etiqueta deben producirse al menos `n` veces en el conjunto entero de documentos o registros a extraerse. Esto también requiere que defina un separador.
- Si tiene una etiqueta que contiene una coma, debe preceder la coma con un carácter de barra inclinada invertida para que no se ignore la declaración. Por ejemplo, si tiene un campo denominado `<topic:source>`, especifíquelo como `<topic\:source>`.

Para ilustrar la sintaxis, asumamos que tiene los siguientes campos bibliográficos recurrentes:

```
author:Morel, Kawashima
abstract:This article describes how fields are declared.
publication:Text Mining Documentation
datepub:March 2010
```

Para este ejemplo, si deseamos que el proceso de extracción se enfoque en el autor y sustrato pero que ignore el resto del contenido, declararíamos sólo los siguientes campos:

```
author: , Person1
abstract:
```

En este ejemplo, la declaración de campo `author: , Person1` dice que el proceso lingüístico se suspendió en los contenidos de campo. En cambio, declara que el campo autor contiene más de un nombre, que está separado del siguiente por un separador coma y que estos nombres deberían asignarse al tipo Persona y que si el nombre sucede al menos una vez en el conjunto entero de documentos o registros, debería extraerse. Puesto que el campo `abstract:` se lista sin ninguna otra declaración, el campo se examinará durante la extracción y se aplicará el proceso y escritura lingüística estándar.

Formateo de texto XML

Si desea limitar el proceso de extracción a sólo el texto dentro de etiquetas XML específicas, utilice la opción de tipo de documento de **texto XML** y declare las etiquetas que contienen el texto en la sección **Formateo de texto XML** del cuadro de diálogo Configuración de documento. Los términos extraídos se derivan sólo del texto contenido dentro de estas etiquetas o de sus etiquetas hijo.

Importante: Si desea pasar por alto el proceso de extracción e imponer reglas en separadores de términos, asignar tipos al texto extraído o imponer un recuento de frecuencia para términos extraídos, utilice la opción **Texto estructurado** que se describe a continuación.

Utilice las siguientes reglas al declarar etiquetas para el formateo de texto XML:

- Sólo se puede declarar una etiqueta XML por línea.
- Los elementos de etiqueta son sensibles a las mayúsculas y minúsculas.
- Si una etiqueta tiene atributos, como `<title id="1234">` y desea incluir todas las variaciones o, en este caso, todos los ID, añada la etiqueta sin el atributo o el corchete final (`>`), como `<title`

Para ilustrar la sintaxis, asumamos que tiene el siguiente documento XML:

```
<section>Rules of the Road
  <title id="01234">Traffic Signals</title>
  <p>Road signs are helpful.</p>
</section>
<p>Learning the rules is important.</p>
```

En este ejemplo, declararemos las siguientes etiquetas:

```
<section>
<title
```

En este ejemplo, puesto que ha declarado la etiqueta `<section>`, el texto de este código y sus etiquetas anidadas, `Traffic Signals` y `Road signs are helpful`, se exploran durante el proceso de extracción. Sin embargo, `Learning the rules is important` es ignorada ya que la etiqueta `<p>` no se declaró explícitamente ni se anidó dentro de una etiqueta declarada.

Nodo Minería de textos: pestaña Modelo

Utilice la pestaña Modelo para especificar el método de compilación y configuración del modelo general del nodo de salida.

Puede establecer los siguientes parámetros:

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilice datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

Modo de creación. Especifica cómo se producirán las nuggets de modelo cuando se ejecute una ruta con este nodo de Minería de textos. De manera alternativa, puede utilizar un método de exploración más práctico al utilizar la modalidad **Compilación interactiva** en la cual no sólo puede extraer conceptos, crear categorías y refinar los recursos lingüísticos, sino que también puede realizar análisis de enlaces de texto y explorar clústeres.

- **Compilación interactiva.** Cuando se ejecuta una ruta, esta opción lanza una interfaz interactiva en la que puede extraer conceptos y patrones, explorar y ajustar los resultados extraídos, crear y refinar categorías, ajustar los recursos lingüísticos (plantillas, sinónimos, tipos, bibliotecas, etc.), y crear nuggets de modelo de categoría. Si desea más información, consulte [“Generar interactivamente”](#) en la página 24.
- **Generar directamente.** Esta opción indica que, cuando se ejecuta la ruta, se debería crear un modelo automáticamente y añadirse a la paleta de Modelos. A diferencia del área de trabajo interactiva, no se necesita manipulación adicional en el tiempo de ejecución aparte de la configuración definida en el nodo. Si selecciona esta opción, aparecerán opciones específicas del modelo con las que puede definir el tipo de modelo que desea producir. Si desea más información, consulte [“Generar directamente”](#) en la página 26.

Almacenar modelos grandes en AS. Si tiene una conexión con IBM SPSS Analytic Server, seleccione esta opción para almacenar Los modelos remotamente en el servidor.

Nota: Los modelos que se creen y almacenen en un servidor sólo podrán puntuarse en dicho servidor. Para reanudar una sesión de área de trabajo interactiva que contenga ese modelo, necesitará una conexión con el servidor original que se haya utilizado para crear la sesión.

Copiar recursos de. Cuando se extrae texto, la extracción no se basa sólo en los valores de la pestaña Experto, sino también en los recursos lingüísticos. Estos recursos sirven como la base de cómo manejar y procesar el texto durante una extracción para obtener los conceptos, tipos y, en algunas ocasiones, los patrones. Puede copiar recursos en este nodo desde una plantilla de recursos, un paquete de análisis de texto (.tap) o un archivo de proyecto SPSS Text Analytics for Surveys (.tas). Realice la selección y, a continuación, pulse **Carga** para definir la plantilla, el paquete o el proyecto desde el que se copiarán los recursos. En el momento de la carga, se almacena una copia del recurso en el nodo. Por lo tanto, si alguna vez desea utilizar un recurso actualizado, debe volver a cargarlo aquí o en una sesión de área de trabajo interactiva. Para su comodidad, la fecha y hora en la que se copiaron y cargaron los recursos se muestra en el nodo. Si desea más información, consulte [“Copiar recursos de plantillas y TAP”](#) en la página 26.

Idioma del texto. Identifica el idioma del texto que se está minando. Los recursos copiados en el nodo controlan las opciones de idioma presentadas. Seleccione el idioma para el que se hayan ajustado los recursos.

Generar interactivamente

En la pestaña Modelo del nodo de modelado de minería de textos, puede elegir un modo de creación para sus nuggets de modelo. Si elige **Crear de forma interactiva**, se abre una interfaz interactiva cuando ejecuta la ruta. En esta área de trabajo interactiva, puede:

- Extraer y explorar los resultados de extracción, incluidos los conceptos y los tipos para descubrir las ideas destacadas en sus datos de texto.
- Utilizar una variedad de métodos para crear y ampliar categorías a partir de conceptos, tipos, patrones de TLA y reglas, para poder puntuar sus documentos y registros en estas categorías.
- Perfeccionar sus recursos lingüísticos (plantillas de recursos, bibliotecas, diccionarios, sinónimos, y más) para poder mejorar sus resultados a través de un proceso iterativo en el que los conceptos se extraen, examinan y perfeccionan.
- Realizar análisis de enlaces de texto (TLA) y utilizar los patrones de TLA descubiertos para crear mejores nuggets de modelo de categoría. El nodo de análisis de enlaces de texto no ofrece las mismas opciones de exploración o prestaciones de modelado.
- Generar clústeres para descubrir relaciones nuevas y explorar relaciones entre concepto, tipos, patrones y categorías en el panel Visualización.

- Generar nuggets de modelo de categoría refinados en la paleta de modelos en IBM SPSS Modeler y utilizarlos en otras rutas.

Nota: No puede crear un modelo interactivo si está creando un trabajo de IBM SPSS Collaboration and Deployment Services.

Utilizar el trabajo de sesión (categorías, TLA, recursos, etc.) de la última actualización de nodo.

Cuando trabaja en una sesión de entorno de trabajo interactivo, puede actualizar el nodo con datos de sesión (parámetros de extracción, recursos, definiciones de categoría, etc.). La opción **Utilizar trabajo de sesión** le permite volver a iniciar el área de trabajo interactiva utilizando los datos de sesión guardados. Esta opción está inhabilitada la primera vez que utiliza este nodo, ya que no es posible que se hayan guardado datos de sesión. Para obtener información sobre cómo actualizar el nodo con datos de sesión para poder utilizar esta opción, consulte [“Guardar y actualizar nodos de modelado”](#) en la página 77.

Si inicia una sesión *con* esta opción, los valores de extracción, categorías, recursos y cualquier otro trabajo de la última vez que realizó una actualización de nodo desde una sesión de área de trabajo interactiva están disponibles cuando vuelve a iniciar una sesión. Puesto que los datos de sesión guardados se utilizan con esta opción, determinado contenido como, por ejemplo, los recursos copiados desde la plantilla de abajo, y otras pestañas, se inhabilitan o ignoran. Pero si inicia una sesión *sin* esta opción, sólo se utiliza el contenido del nodo tal como está definido en el momento, lo que significa que cualquier trabajo anterior que haya realizado en el entorno de trabajo no estará disponible.

Nota:: si cambia el nodo de origen para la ruta después de que los resultados de la extracción se hayan almacenado en la memoria caché con la opción **Utilizar trabajo de sesión...**, deberá ejecutar una nueva extracción una vez que se inicie la sesión interactiva del entorno de trabajo si desea obtener resultados de extracción actualizados.

Omitir extracción y volver a utilizar datos y resultados de la memoria caché. Puede volver a utilizar cualquier dato y resultado de extracción de la memoria caché en la sesión de área de trabajo interactiva. Esta opción es particularmente útil cuando desea ahorrar tiempo y volver a utilizar resultados de extracción en lugar de esperar a que se realice una extracción completamente nueva cuando se inicie la sesión. Para utilizar esta opción, debe haber actualizado este nodo anteriormente desde dentro de una sesión de área de trabajo interactiva y haber elegido la opción de **Mantener el trabajo de sesión y almacenar datos de texto en memoria caché con resultados de extracción para reutilización**. Para obtener información sobre cómo actualizar el nodo con datos de sesión para poder utilizar esta opción, consulte [“Guardar y actualizar nodos de modelado”](#) en la página 77.

Comenzar sesión por. Seleccione la opción indicando la vista y acción que desea que se lleve a cabo en primer lugar tras iniciar la sesión de área de trabajo interactiva. Independientemente de la vista en la que comience, puede cambiar a cualquier vista una vez que esté en la sesión.

- **Utilizar resultados de extracción para crear categorías.** Esta opción inicia el área de trabajo interactiva en la vista Categorías y Conceptos y, si es aplicable, realiza una extracción. En esta vista, puede crear categorías y generar un modelo de categoría. También puede cambiar a otra vista. Para obtener más información, consulte el tema [Capítulo 7, “Modalidad de área de trabajo interactiva”](#), en la página 67.
- **Explorar resultados de análisis de enlaces de texto (TLA).** Esta opción se inicia y comienza extrayendo e identificando relaciones entre conceptos dentro del texto como, por ejemplo, opiniones u otros enlaces en la vista de análisis de enlaces de texto. Debe seleccionar una plantilla o paquete de análisis de texto que contenga reglas de patrón de TLA para poder utilizar esta opción y obtener resultados. Si está trabajando con conjuntos de datos más grandes, la extracción de TLA puede demorar. En este caso, puede que desee considerar la posibilidad de utilizar un nodo de muestra en sentido ascendente. Para obtener más información, consulte el tema [Capítulo 11, “Exploración del Análisis de enlaces de texto”](#), en la página 149.
- **Analizar clústeres co-word.** Esta opción se inicia en la vista Clústeres y actualiza cualquier resultado de extracción obsoleto. En esta vista, puede llevar a cabo el análisis de clúster co-word, que produce un conjunto de clústeres. La agrupación en clúster co-word es un proceso que comienza evaluando la potencia del valor de enlace entre dos conceptos en función de su co-ocurrencia en un determinado registro o documento y finaliza con la agrupación de conceptos fuertemente enlazados en clústeres.

Para obtener más información, consulte el tema [Capítulo 7, “Modalidad de área de trabajo interactiva”](#), en la [página 67](#).

Generar directamente

En la pestaña Modelo del nodo de modelado de minería de textos, puede elegir un modo de creación para sus nuggets de modelo. Si elige **Generar directamente**, puede establecer las opciones en el nodo y, a continuación, solamente ejecutar la ruta. La salida es un nugget de modelo de concepto, que se ha colocado directamente en la paleta de modelos. A diferencia del área de trabajo interactiva, no es necesario ningún tipo de manipulación adicional por parte del usuario en el momento de la ejecución además de los valores de frecuencia definidos para esta opción en el nodo.

Número máximo de conceptos a incluir en el modelo. Esta opción, que se aplica sólo cuando crea un modelo automáticamente (no interactivo), indican que desea crear un modelo de concepto. También indica que este modelo no debe contener más del número especificado de conceptos.

- **Seleccionar conceptos en función de la frecuencia más alta. Número especificado de conceptos.** Empezando por el concepto con la frecuencia más alta, este es el número de conceptos que se seleccionará. Aquí, frecuencia hace referencia al número de veces que aparece un concepto (y todos sus términos subyacentes) en todo el conjunto de los documentos/registros. Este número puede ser mayor al recuento de registros, ya que un concepto puede aparecer varias veces en un registro.
- **Deseleccionar conceptos que aparecen en demasiados registros. Porcentaje de registros.** Deselecciona conceptos con un porcentaje de recuento de registros mayor al número que ha especificado. Esta opción es útil para excluir conceptos que aparecen frecuentemente en el texto o en cada registro pero que no tienen ningún significado en su análisis.

Optimizar para velocidad de puntuación. Seleccionada de forma predeterminada, esta opción asegura que el modelo creado sea compacto y que puntúe a alta velocidad. Si se deselecciona esta opción se crea un modelo mucho más grande que puntúa más lentamente. Sin embargo, el modelo más grande asegura que las puntuaciones visualizadas inicialmente en el modelo de concepto generado sean las mismas que las obtenidas al puntuar el mismo texto con el nugget de modelo.

Copiar recursos de plantillas y TAP

Cuando se extrae texto, la extracción no se basa sólo en los valores de la pestaña Experto, sino también en los recursos lingüísticos. Estos recursos sirven como base para manejar y procesar el texto durante la extracción para obtener los conceptos, tipos, y a veces patrones. Puede copiar recursos en este nodo desde una *plantilla de recursos*, y si se encuentra en el nodo de minería de textos, también puede seleccionar un *paquete de análisis de texto* (TAP) o un proyecto de SPSS Text Analytics for Surveys (.tas).

De forma predeterminada, los recursos se copian en el nodo de la plantilla básica para idiomas con licencia para su producto cuando añade el nodo al lienzo. Si tiene licencias para varios idiomas, se utiliza el primer idioma seleccionado para determinar la plantilla que se va a cargar automáticamente.

En el momento de la carga, se almacena una copia de los recursos seleccionados en el nodo. Sólo se copia el contenido de la plantilla, TAP o recurso de proyecto de SPSS Text Analytics for Surveys siempre que la plantilla, TAP o el propio SPSS Text Analytics for Surveys no esté vinculado al nodo. Esto significa que si los recursos se actualizan posteriormente, estas actualizaciones no estarán disponibles en el nodo automáticamente. En resumen, los recursos cargados en el nodo siempre se utilizan, a menos que actualice un nodo de minería de textos y seleccione la opción **Utilizar trabajo de sesión**. Para obtener más información sobre **Utilizar trabajo de sesión**, continúe leyendo esta sección.

Cuando seleccione un recurso, elija una con el mismo idioma que sus datos de texto. Sólo puede utilizar recursos en los idiomas para los que tiene licencia. Si desea llevar a cabo un análisis de enlaces de texto, debe seleccionar una plantilla que contenga patrones de TLA. Si una plantilla contiene patrones de TLA, aparecerá un icono en la columna TLA del recuadro de diálogo Cargar plantilla de recursos.

Nota: No puede cargar TAP o proyectos de SPSS Text Analytics for Surveys en el nodo de análisis de enlaces de texto.

Plantillas de recurso

Una plantilla de recursos es un conjunto predefinido de bibliotecas y recursos lingüísticos y no lingüísticos avanzados que se han afinado para un determinado dominio o uso. En el nodo de modelado de minería de textos, ya está cargada en el nodo una copia de los recursos de una plantilla básica cuando añade el nodo a la ruta, pero puede cambiar plantillas o cargar un paquete de análisis de texto seleccionando **Plantilla de recursos** o **Paquete de análisis de texto** y, a continuación, pulsando **Cargar**. Para las plantillas, puede seleccionar la plantilla en el recuadro de diálogo Cargar plantilla de recursos.

Nota: Si no ve las plantillas que desea en la lista, pero tiene una copia exportada en la máquina, puede importarla ahora. También puede exportar desde este recuadro de diálogo para compartir con otros usuarios. Si desea más información, consulte [“Importación y exportación de plantillas”](#) en la página 175.

Proyectos de Paquetes de análisis de texto (TAP) y Text Analysis for Surveys (TAS)

Un paquete de análisis de texto (TAP) es un conjunto predefinido de bibliotecas y recursos lingüísticos y no lingüísticos avanzados incorporado con uno o más conjuntos de categorías predefinidas. IBM SPSS Modeler Text Analytics ofrece varios TAP creados previamente, que están especialmente adaptados para un dominio específico. Puede editar estos TAP y guardarlos en un directorio distinto para utilizarlos para acelerar la creación de modelos de categoría. También puede crear sus propios TAP en la sesión interactiva. Si desea más información, consulte [“Carga de los paquetes de análisis de texto”](#) en la página 136.

Si selecciona importar un proyecto de SPSS Text Analytics for Surveys (.tas), se convertirá en un TAP.

Nota: No puede cargar TAP o proyectos de SPSS Text Analytics for Surveys en el nodo de análisis de enlaces de texto.

Utilización de la opción "Utilizar trabajo de sesión" (Pestaña Modelo)

Mientras que los recursos se copian en el nodo en la pestaña Modelo, también puede hacer cambios posteriores en los recursos en una sesión interactiva y puede que desee actualizar el nodo de modelado de minería de textos con estos últimos cambios. En este caso, tendría que seleccionar la opción **Utilizar trabajo de sesión** en la pestaña Modelo del nodo de modelado de minería de texto.

Si selecciona **Utilizar trabajo de sesión**, el botón **Cargar** se inhabilita en el nodo para indicar que se utilizarán los recursos que provienen del área de trabajo interactiva en lugar de los recursos cargados aquí anteriormente.

Para hacer cambios en los recursos una vez que ha seleccionado la opción **Utilizar trabajo de sesión**, puede editar o cambiar los recursos directamente dentro de la sesión del área de trabajo interactiva a través de la vista Editor de recursos. Si desea más información, consulte [“Actualización de los recursos en el nodo después de cargar”](#) en la página 174.

Nodo Minería de textos: pestaña Experto

La pestaña Experto contiene ciertos parámetros avanzados que afectan a la extracción y al manejo del texto. Los parámetros en este recuadro de diálogo controlan el comportamiento básico, así como unos pocos comportamientos avanzados, del proceso de extracción. Sin embargo, representan solamente una porción de las opciones disponibles. También hay un número de recursos lingüísticos y opciones que afectan a los resultados de la extracción, que son controlados por la plantilla de recursos que selecciona en la pestaña Modelo. Para obtener más información, consulte el tema [“Nodo Minería de textos: pestaña Modelo”](#) en la página 23.

Nota: Toda esta pestaña está inhabilitada si ha seleccionado la modalidad **Crear de forma interactiva** utilizando información del área de trabajo interactiva guardada en la pestaña Modelo, en cuyo caso los valores de extracción se toman de la última sesión de área de trabajo guardada.

Puede establecer los parámetros siguientes siempre que realice una extracción:

Limite la extracción a conceptos con una frecuencia global de al menos [n]. Especifica el número mínimo de veces que debe aparecer una palabra o frase en el texto para que pueda extraerse. De esta

forma, un valor de 5 limita la extracción a aquellas palabras o frases que aparecen al menos cinco veces en todos los registros o documentos.

En algunos casos, el hecho de cambiar este límite puede suponer una diferencia sustancial en los resultados extraídos y, por consiguiente, en las categorías. Digamos que trabaja con datos de un restaurante y no incrementa el límite por encima de 1 para esta opción. En ese caso, podría encontrar *pizza* (1), *pizza fina* (2), *pizza vegetal* (2) y *pizza favorita* (2) en los resultados de extracción. Sin embargo, si limitara la extracción con una frecuencia global de 5 o más y volviera a extraer los datos, dejarían de aparecer tres de estos conceptos. Así pues, obtendría *pizza* (7), puesto que *pizza* es la forma más simple y esta palabra se encontraba también como posible candidata. Y en función del resto del texto, podría de hecho tener una frecuencia mayor que siete en función de si hay otras frases con la palabra *pizza* en su interior. Además, si *pizza vegetal* ya era un descriptor de categoría, podría añadir *pizza* como descriptor en lugar de capturar todos los registros. Por esta razón, se recomienda que cambie este límite con precaución siempre que ya se hayan creado categorías.

Tenga en cuenta que esta es una característica sólo de extracción; si la plantilla contiene términos (que usualmente sucede) y se encuentra el término para la plantilla en el texto, este se indexará independientemente de su frecuencia.

Por ejemplo, supongamos que utiliza una plantilla de Recursos básicos que incluye "los angeles" en el tipo <Location> en la biblioteca núcleo; si el documento contiene Los Angeles sólo una vez, Los Angeles formará parte de la lista de conceptos. Para prevenirlo necesitará establecer un filtro que muestre conceptos que sucedan al menos la misma cantidad de veces que el valor especificado en el campo **Limitar extracción a conceptos con una frecuencia global de al menos [n]**.

Areglar errores de puntuación. Esta opción normaliza temporalmente el texto que contiene errores de puntuación (por ejemplo, el uso inapropiado) durante la extracción para mejorar la capacidad de extracción de los conceptos. Esta opción es extremadamente útil cuando el texto es breve y de calidad mediocre (como, por ejemplo, en respuestas de encuestas con final abierto, correo electrónico y datos CRM) o cuando el texto contiene muchas abreviaturas.

Acomodar la ortografía a una longitud mínima de caracteres de palabra de [n] Esta opción aplica una técnica de agrupación difusa que ayuda a agrupar bajo un concepto las palabras que suelen escribirse mal o que tienen una ortografía parecida. El algoritmo de agrupación difusa elimina temporalmente todas las vocales (excepto la primera) y elimina las consonantes dobles/triples de las palabras extraídas y luego las compara para ver si son iguales, de forma que *modeling* y *modelling* se agruparían. Sin embargo, si a cada término se le asigna un tipo diferente, excluyendo el tipo <Unknown>, la técnica de agrupación difusa no se aplicará.

También puede definir el número mínimo de caracteres *raíz* necesarios para poder utilizar la agrupación difusa. El número de caracteres raíz de un término se calcula sumando todos los caracteres y restando los que forman los sufijos de las declinaciones, y en el caso de términos de palabras compuestas, también los determinantes y las preposiciones. Por ejemplo, el término *exercises* se contaría como 8 caracteres raíz en la forma "exercise," ya que la letra *s* al final de la palabra es una inflexión (forma plural. De forma similar, *apple sauce* se cuenta como 10 caracteres raíz ("apple sauce") y *manufacturing of cars* se cuenta como 16 caracteres raíz ("manufacturing car"). Este método de recuento de caracteres solo se utiliza para comprobar si debe aplicarse la agrupación difusa, pero no influye en la forma de coincidencia de las palabras.

Nota: Si encuentra que ciertas palabras posteriormente se agrupan incorrectamente, puede excluir pares de palabras de esta técnica declarándolos explícitamente en la sección **Agrupación difusa: Excepciones** en la pestaña Recursos avanzados. Para obtener más información, consulte el tema ["Agrupación difusa"](#) en la [página 201](#).

Extraer unitérminos Esta opción extrae palabras simples (unitérminos) siempre que la palabra no forme parte de una palabra compuesta, y si es un sustantivo o una categoría léxica no reconocida.

Extraer entidades no lingüísticas Esta opción extrae entidades no lingüísticas, como números de teléfono, números de la seguridad social, horas, fechas, monedas, dígitos, porcentajes, direcciones de correo electrónico y direcciones de HTTP. Puede incluir o excluir ciertos tipos de entidades no lingüísticas en la sección **Entidades no lingüísticas: Configuración** de la pestaña Recursos avanzados. Si

se desactivan las entidades innecesarias, el motor de extracción no malgastará tiempo de proceso. Para obtener más información, consulte el tema [“Configuración”](#) en la página 205.

Algoritmo de mayúsculas Esta opción extrae términos simples y compuestos que no están en los diccionarios incorporados, siempre que la primera letra del término esté en mayúscula. Esta opción supone un buen método para extraer la mayoría de los nombres propios.

Agrupar nombres de persona parciales y completos juntos cuando sea posible Esta opción agrupa nombres que aparecen de diferente manera juntos en el texto. Esta característica es útil porque a menudo se hace referencia a los nombres completos al principio del texto, y más adelante se utiliza la versión abreviada. Esta opción intenta hacer coincidir cualquier unitérmino que tenga el tipo <Unknown> con la última palabra de cualquier término compuesto que se haya tipificado como <Person>. Por ejemplo, si se encuentra *garcía*, que inicialmente se tipificó como <Unknown>, el motor de extracción comprobará si hay algún término compuesto en el tipo <Person> con el término *garcía* como la última palabra, como en *juan garcía*. Esta opción no se aplica a los nombres propios, porque muchos de ellos no se extraen nunca como unitérminos.

Permutación máxima de palabra no funcional Esta opción especifica el número máximo de palabras no funcionales que debe haber para poder aplicar la técnica de permutación. Esta técnica de permutación agrupa frases similares que difieren entre sí solo en las palabras no funcionales (por ejemplo, de y el), independientemente de la flexión. Por ejemplo, digamos que ha establecido este valor en dos palabras como máximo y tanto *company officials* como *officials of the company* se han extraído. En este caso, ambos términos extraídos se agruparían en la lista de conceptos final, ya que ambos términos se consideran iguales cuando se omite *of the*.

Utilizar la derivación al agrupar multitérminos Cuando procese grandes volúmenes de datos, seleccione esta opción para agrupar multitérminos mediante las reglas de derivación.

Nota: Para habilitar la extracción de los resultados del análisis de enlaces de texto, debe iniciar la sesión con la opción **Explorar resultados de análisis de enlaces de texto** y también elegir recursos que contengan definiciones de TLA. Siempre puede extraer resultados de TLA posteriormente durante una sesión de área de trabajo interactiva a través del diálogo Valores de extracción. Para obtener más información, consulte el tema [“Extraer datos”](#) en la página 82.

Muestreo en sentido ascendente para ahorrar tiempo

Cuando tiene una gran cantidad de datos, los tiempos de procesamiento pueden tomar minutos u horas, especialmente cuando se utiliza la sesión de área de trabajo interactiva. Cuando más grande es la cantidad de datos, más tiempo llevará el proceso de categorización y de extracción. Para trabajar de forma más eficiente, puede añadir Nodos de muestra de IBM SPSS Modeler en sentido ascendente desde el nodo Minería de texto. Utilice este nodo de Muestra para tomar una muestra aleatoria utilizando un subconjunto más pequeño de documentos o registros para realizar los primeros pasos.

Una muestra más pequeña es a menudo perfectamente adecuada para decidir cómo editar los recursos y hasta crear la mayoría de las categorías (sino todas). Una vez que lo ejecute en el conjunto de datos más pequeño y esté satisfecho con los resultados, puede aplicar la misma técnica para crear categorías al conjunto de datos entero. A continuación, puede buscar documentos o registros que no se ajustan a las categorías creadas y hacer los ajustes necesarios.

Nota: EL nodo Muestra es un nodo estándar de IBM SPSS Modeler.

Utilización del nodo Minería de texto en una ruta

El nodo de modelado Minería de textos se utiliza para acceder datos y extraer conceptos en una ruta. Puede utilizar cualquier nodo fuente para acceder a datos, como el nodo Base de datos, Var. Nodo Archivo, nodo Canal de información web o nodo Archivo fijo. Para el texto que reside en documentos externos, se puede utilizar un nodo Lista de archivos.

Ejemplo 1: el nodo Lista de archivos y el nodo Minería de textos para compilar un nugget de modelo de concepto directamente

El siguiente ejemplo muestra cómo utilizar el nodo Lista de archivos junto con el nodo de modelado de Minería de textos para generar el nugget de modelo de concepto. Para obtener más información sobre cómo utilizar el nodo Lista de archivos, consulte [“Nodo Lista de archivos” en la página 11](#).

1. **Nodo Lista de archivos (pestaña Configuración).** En primer lugar, añadimos este nodo a la ruta para especificar dónde están almacenados los documentos de texto. Seleccionamos este directorio que contiene todos los documentos en los que deseamos realizar la minería de textos.
2. **Nodo Minería de textos (pestaña Campos).** A continuación, añadimos y conectamos un nodo Minería de textos al nodo Lista de archivos. En este nodo, definimos nuestro formato de entrada, plantilla de recursos y formato de salida. Hemos seleccionado el nombre de campo generado a partir del nodo Lista de archivos y hemos seleccionado el campo de texto, así como otros valores. Para obtener más información, consulte el tema [“Utilización del nodo Minería de texto en una ruta” en la página 29](#).
3. **Nodo Minería de textos (pestaña Modelo).** A continuación, en la pestaña Modelo, seleccionamos el modo compilación para generar un nugget de modelo de concepto directamente desde este nodo. Puede seleccionar una plantilla de recurso distinta o mantener los recursos básicos.

Ejemplo 2: nodos de Archivo Excel y Minería de textos para compilar interactivamente un modelo de categoría

Este ejemplo muestra cómo el nodo de Minería de textos también puede iniciar una sesión de área de trabajo interactiva. Para obtener más información sobre el área de trabajo interactiva, consulte [Capítulo 7, “Modalidad de área de trabajo interactiva”](#), en la página 67.

1. **Nodo fuente Excel (pestaña Datos).** Primero, añadimos este nodo a la ruta para especificar dónde se almacena el texto.
2. **Nodo Minería de textos (pestaña Campos).** A continuación, añadimos y conectamos un nodo Minería de textos. En esta primera pestaña, definimos nuestro formato de entrada. Hemos seleccionado un nombre de campo en el nodo de origen.
3. **Nodo Minería de textos (pestaña Modelo).** A continuación, en la pestaña Modelo, seleccionamos compilar un nugget de modelo de categoría de manera interactiva y para utilizar los resultados de extracción para compilar categorías automáticamente. En este ejemplo, cargamos una copia de los recursos y un conjunto de categorías a partir de un paquete de análisis de texto.
4. **Sesión de área de trabajo interactiva.** A continuación, ejecutamos la ruta y se abrió la interfaz de área de trabajo interactiva. Después de realizar una extracción, comenzamos a explorar los datos y mejorar las categorías.

Nugget de minería de textos: Modelo de concepto

Un nugget de modelo de concepto de minería de textos se crea siempre que ejecuta correctamente un nodo de modelo de minería de textos donde ha seleccionado la opción de **Generar un modelo directamente** en la pestaña Modelo. Un nugget de modelo de concepto de minería de textos se utiliza para el descubrimiento en tiempo real de conceptos clave en otros datos de texto como, por ejemplo, datos de anotación de un centro de asistencia telefónica.

El nugget de modelo de concepto consta de una lista de conceptos, que se han asignado a tipos. Puede seleccionar uno o todos los conceptos de ese modelo para la puntuación con otros datos. Cuando ejecuta una ruta que contiene un nugget de modelo de minería de textos, se añaden campos nuevos a los datos en función de la modalidad de creación seleccionada en la pestaña Modelo del nodo de modelado de minería de textos antes de crear el modelo. Para obtener más información, consulte el tema [“Modelo de concepto: Pestaña Modelo” en la página 31](#).

Si el nugget de modelo se ha generado utilizando documentos traducidos, la puntuación se realizará en el idioma traducido. Del mismo modo, si el nugget de modelo se ha generado utilizando el inglés como idioma, puede especificar un idioma de traducción en el nugget de modelo, ya que los documentos se traducirán al inglés.

Los nugget de modelo de minería de datos de texto se colocan en la paleta de nugget de modelo (ubicada en la pestaña Modelos en la parte superior derecha de la ventana de IBM SPSS Modeler) cuando se generan.

Visualización de los resultados

Para ver información sobre el nugget de modelo, pulse con el botón derecho el nodo en la paleta de nuggets de modelo y elija **Examinar** en el menú contextual(o **Editar** en los nodos de una ruta).

Añadir modelos a una ruta

Para añadir el nugget de modelo a una ruta, pulse el icono en la paleta nuggets de modelo y pulse el lienzo de rutas donde desea ubicar el nodo. O pulse con el botón derecho el icono y elija **Añadir a ruta** en el menú contextual. A continuación, conecte su ruta al nodo y ya está listo para pasar los datos para generar predicciones.

Precaución: si desea utilizar un nugget de puntuación para volver a generar un nodo de modelado que contenga el modelo de categoría y la plantilla utilizada, se recomienda crear un TAP y utilizarlo en una sesión interactiva, en lugar del nodo de modelado, antes de generar el nugget de puntuación.

Modelo de concepto: Pestaña Modelo

En modelos de concepto, la pestaña Modelo muestra el conjunto de conceptos extraídos. Los conceptos se presentan en un formato de tabla con una fila para cada concepto. El objetivo de esta pestaña es seleccionar los conceptos que se utilizarán para la puntuación.

Nota: Si ha generado un nugget de modelo de categoría en su lugar, esta pestaña presentará información diferente. Para obtener más información, consulte el tema [“Nugget de modelo de categoría: Pestaña Modelo”](#) en la página 39.

Todos los conceptos están seleccionados para la puntuación de forma predeterminada, tal como se muestra en los recuadros de selección de la columna del extremo izquierdo. Un recuadro seleccionado significa que el concepto se utilizará para la puntuación. Un recuadro no seleccionado significa que el concepto se excluirá de la puntuación. Puede seleccionar varias filas seleccionándolas y pulsando uno de los recuadros de selección en su selección.

Para obtener más información sobre cada concepto, puede consultar la información adicional proporcionada en cada una de las siguientes columnas:

Concepto. Esta es la palabra o frase principal que se ha extraído. En algunos casos, este concepto representa el nombre del concepto así como otros términos subyacentes asociados con este concepto. Para conocer los términos subyacentes que son parte de un concepto, visualice el panel Términos subyacentes dentro de esta pestaña y seleccione el concepto para ver los términos correspondientes en la parte inferior del recuadro de diálogo. Para obtener más información, consulte el tema [“Términos subyacentes en modelos de concepto”](#) en la página 33.

Global. Aquí, global (frecuencia) hace referencia al número de veces que aparece un concepto (y todos sus términos subyacentes) en todo el conjunto de los documentos/registros.

- **Gráfico de barras.** La frecuencia global de este concepto en los datos de texto presentados como un gráfico de barras. La barra toma el color del tipo al que está asignado el concepto para poder distinguir los tipos visualmente.
- **%.** La frecuencia global de este concepto en los datos de texto presentada como un porcentaje.
- **N.** El número real de apariciones de este concepto en los datos de texto.

Docs. Aquí, Docs hace referencia al recuento de documentos, lo que significa el número de documentos o registros en los que aparece el concepto (y todos sus términos subyacentes).

- **Gráfico de barras.** El recuento de documentos para este concepto presentado como un gráfico de barras. La barra toma el color del tipo al que está asignado el concepto para poder distinguir los tipos visualmente.
- **%.** El recuento de documentos para este concepto presentado como un porcentaje.
- **N.** El número real de documentos o registros que contienen este concepto.

Tipo. El tipo al que está asignado el concepto. Para cada concepto, las columnas Global y Docs aparecen en un color para indicar el tipo al que está asignado este concepto. Un **tipo** es una agrupación semántica de conceptos. Para obtener más información, consulte el tema [“Diccionarios de tipo” en la página 187.](#)

Trabajo con conceptos

Pulsando una celda de la tabla con el botón derecho del ratón, puede mostrar un menú contextual en el que puede:

- **Seleccionar todo.** Se seleccionarán todas las filas de la tabla.
- **Copiar.** Los conceptos seleccionados se copian en el portapapeles.
- **Copiar con campos** Los conceptos seleccionados se copian en el portapapeles junto con la cabecera de la columna.
- **Seleccionar seleccionados.** Selecciona todos los recuadros de selección para las filas seleccionadas en la tabla, incluyendo así esos conceptos para la puntuación.
- **Deseleccionar seleccionados.** Deselecciona todos los recuadros de selección para las filas seleccionadas en la tabla.
- **Seleccionar todo.** Selecciona todos los recuadros de selección en la tabla. Esto da como resultado la utilización de todos los conceptos en la salida final.
- **Deseleccionar todo.** Deselecciona todos los recuadros de selección en la tabla. La deselección de un concepto significa que no se utilizará en la salida final.
- **Incluir conceptos.** Muestra el recuadro de diálogo Incluir conceptos. Para obtener más información, consulte el tema [“Opciones para incluir conceptos para su puntuación” en la página 32.](#)

Opciones para incluir conceptos para su puntuación

Para marcar o desmarcar rápidamente esos conceptos que se utilizarán para la puntuación, pulse el botón de la barra de herramientas para **Incluir conceptos**.



Figura 1. Botón de la barra de herramientas Incluir conceptos

Al pulsar este botón de la barra de herramientas se abrirá el cuadro de diálogo Incluir conceptos para permitirle seleccionar conceptos basados en reglas. Todos los conceptos que tienen una marca de selección en la pestaña Modelo, se incluirán para la puntuación. Aplique una regla en este subdiálogo para cambiar cuáles conceptos se utilizarán para la puntuación.

Puede elegir entre las siguientes opciones:

Seleccionar conceptos en función de la frecuencia más alta. Número especificado de conceptos.

Comenzando con el concepto con la frecuencia global más alta, este es el número de conceptos que se comprobarán. Aquí, frecuencia hace referencia al número de veces que aparece un concepto (y todos sus términos subyacentes) en todo el conjunto de los documentos/registros. Este número puede ser mayor al recuento de registros, ya que un concepto puede aparecer varias veces en un registro.

Comprobar conceptos en base al recuentos de documentos. Recuento mínimo. Se trata del número de documentos más bajo necesario para comprobar los conceptos. En este caso, el recuento de documentos hace referencia al número de documentos/registros en los que aparece el concepto (y todos sus términos subyacentes).

Comprobar conceptos asignados al tipo. Seleccione un tipo de la lista desplegable para comprobar todos los conceptos asignados a este tipo. Los conceptos se asignan automáticamente a los tipos durante el proceso de extracción. Un **tipo** es una agrupación semántica de conceptos. Los tipos incluyen elementos como conceptos de nivel más alto, palabras positivas y negativas, y calificadores, calificadores contextuales, nombres propios, lugares, organizaciones, etc. Para obtener más información, consulte el tema [“Diccionarios de tipo” en la página 187.](#)

Deseleccionar conceptos que aparecen en demasiados registros. Porcentaje de registros.

Deselecciona conceptos con un porcentaje de recuento de registros mayor al número que ha

especificado. Esta opción es útil para excluir conceptos que aparecen frecuentemente en el texto o en cada registro pero que no tienen ningún significado en su análisis.

Deseleccionar conceptos asignados al tipo. Deselecciona conceptos que coinciden con el tipo que seleccionó en la lista desplegable.

Términos subyacentes en modelos de concepto

Puede ver los términos subyacentes definidos para los conceptos seleccionados en la tabla. Al pulsar el botón de conmutador de los términos subyacentes en la barra de herramientas, puede visualizar la tabla de términos subyacentes en un panel dividido en la parte inferior del diálogo.

Estos términos subyacentes incluyen los sinónimos definidos en los recursos lingüísticos (independientemente de si se encuentran en el texto o no) como también cualquier forma plural/singular encontrada en el texto utilizado para generar el nugget de modelo, términos permutados, términos de agrupación difusa, etc.



Figura 2. Botón de la barra de herramientas Mostrar términos subyacentes

Nota: no se puede editar la lista de términos subyacentes. Esta lista se genera a través de sustituciones, definiciones de sinónimos (en el diccionario de sustituciones), agrupación difusa y otros, todos los que se definen en los recursos lingüísticos. Para realizar cambios en la forma en cómo se agrupan los términos bajo un concepto o cómo se manejan, debe realizar cambios directamente en los recursos (editable en Editor de recursos en el área de trabajo interactiva o en Editor de plantillas y después volver a cargarlos en el nodo) y, a continuación, volver a ejecutar la ruta para obtener un nuevo nugget de modelo con los resultados actualizados.

Al pulsar con el botón derecho la celda que contiene un término o concepto subyacente, puede visualizar un menú contextual en el que puede:

- **Copiar.** La celda seleccionada se copia al portapapeles.
- **Copiar con campos.** La celda seleccionada se copia al portapapeles junto con las cabeceras de columna.
- **Seleccionar todo.** Se seleccionarán todas las celdas en la tabla.

Modelo de concepto: pestaña Configuración

La pestaña Valores se utiliza para definir el valor de campo de texto para los nuevos datos de entrada, de ser necesario. También es el lugar donde define el modelo de datos para su salida (modalidad de puntuación).

Nota: Esta pestaña aparece sólo cuando el nugget de modelo está colocado en el lienzo. No existe cuando accede a este recuadro de diálogo directamente en la paleta de modelos.

Modalidad de puntuación: Conceptos como registros

Con esta modalidad de puntuación, se crea un registro nuevo para cada par concept/document. Generalmente, hay más registros en la salida que los que había en la entrada.

Además de los campos de entrada, se añaden los siguientes campos nuevos a los datos:

Tabla 4. Campos de salida para "Conceptos como registros"	
Campo	Descripción
Concept	Contiene el nombre de concepto extraído encontrado en el campo de datos de texto.
Type	Almacena el tipo de concepto como un nombre de tipo completo como, por ejemplo, <i>Ubicación</i> o <i>Persona</i> . Un tipo es una agrupación semántica de conceptos. Para obtener más información, consulte el tema "Diccionarios de tipo" en la página 187.

Tabla 4. Campos de salida para "Conceptos como registros" (continuación)	
Campo	Descripción
Count	Muestra el número de apariciones para ese concepto (y sus términos subyacentes) en el cuerpo del texto (registro/documento).

Cuando selecciona esta opción, se inhabilitan las otras opciones, a excepción de **Adaptar errores de puntuación**.

Modalidad de puntuación: Conceptos como campos

En modelos de concepto, para cada registro de entrada, se crea un registro nuevo para cada concepto encontrado en un determinado documento. Por lo tanto, existe la misma cantidad de registros de salida que existía en la entrada. Sin embargo, ahora cada registro (fila) contiene un campo nuevo (columna) para cada concepto que se ha seleccionado (utilizando la marca de selección) en la pestaña Modelo. El valor para cada campo de concepto depende de si selecciona **Distintivos** o **Recuentos** como su valor de campo en esta pestaña.

Nota: Si está utilizando conjuntos de datos muy grandes, por ejemplo, con una base de datos Db2, el uso de **Conceptos como campos** puede encontrar problemas de proceso debido a la cantidad de datos. En este caso, se recomienda utilizar **Conceptos como registros** en su lugar.

Valores de campo. Elija si el campo nuevo para cada concepto contendrá un valor de recuento o de distintivo.

- **Distintivos.** Esta opción se utiliza para obtener distintivos con dos valores distintos en la salida, como por ejemplo *Sí/No*, *True/False*, *T/Fo 1 y 2*. Los tipos de almacenamiento se establecen automáticamente para reflejar los valores elegidos. Por ejemplo, si especifica valores numéricos para los distintivos, se manejarán automáticamente como un valor entero. Los tipos de almacenamiento de las marcas pueden ser una cadena, un número entero, un número real o la fecha/hora. Especifique un valor de distintivo para **True** y para **False**.
- **Recuentos.** Se utiliza para obtener un recuento de la cantidad de apariciones del concepto en un determinado registro.

Extensión de nombre de campo. Especifique una extensión para el nombre de campo. Los nombres de campo se generan al utilizar el nombre de concepto más esta extensión.

- **Añadir como.** Especifique dónde debe añadirse la extensión del nombre de campo. Elija **Prefijo** para añadir la extensión al inicio de la cadena. Elija **Sufijo** para añadir la extensión al final de la cadena.

Arreglar errores de puntuación. Esta opción normaliza temporalmente el texto que contiene errores de puntuación (por ejemplo, el uso inapropiado) durante la extracción para mejorar la capacidad de extracción de los conceptos. Esta opción es extremadamente útil cuando el texto es breve y de calidad mediocre (como, por ejemplo, en respuestas de encuestas con final abierto, correo electrónico y datos CRM) o cuando el texto contiene muchas abreviaturas.

Modelo de concepto: pestaña Campos

La pestaña Campos define el valor del campo de texto para los nuevos datos de entrada, si es necesario.

Nota: Esta pestaña solo aparece cuando el nugget de modelo está colocado en la ruta. No existe cuando accede a esta salida directamente en la paleta Modelos.

Campo de texto. Seleccione el campo que contiene el texto donde se va a realizar la minería. Este campo depende del origen de datos.

Tipo de documento. El tipo de documento especifica la estructura del texto. Seleccione uno de los siguientes tipos:

- **Texto completo.** Se utiliza para la mayoría de los documentos o fuentes de texto. El conjunto de texto completo se explora para su extracción. A diferencia de las otras opciones, no hay valores adicionales para esta opción.

- **Texto estructurado.** Se utiliza para formularios bibliográficos, patentes y cualquier campo que contenga estructuras regulares que puedan identificarse y analizarse. Este tipo de documento se utiliza para omitir todo o parte del proceso de extracción. Le permite definir separadores de términos, asignar tipos e imponer un valor de frecuencia mínima. Si selecciona esta opción, debe pulsar el botón **Valores** y especificar separadores de texto en **Formato de texto estructurado.** del cuadro de diálogo Valores del documento. Para obtener más información, consulte el tema [“Pestaña Valores de documento para campos”](#) en la página 22.

Codificación de entrada. Esta opción está disponible sólo si ha indicado que el campo de texto representa **Nombres de vía de acceso a documentos.** Especifica la codificación de texto predeterminada. Una conversión se realiza desde la codificación especificada o reconocida a ISO-8859-1. Por lo que si especificó otra codificación, el motor de extracción la convertirá a ISO-8859-1 antes de que sea procesada. Los caracteres que no se ajusten a la definición de codificación de ISO-8859-1 se convertirán en espacios.

Idioma del texto. Identifica el idioma del texto que se está minando; este es el idioma principal detectado durante la extracción. Póngase en contacto con su representante de ventas si le interesa comprar una licencia para un idioma admitido para el que no tiene acceso de momento.

Modelo de concepto: pestaña Resumen

La pestaña Resumen presenta información sobre el modelo mismo (carpeta *Análisis*), campos utilizados en el modelo (carpeta *Campos*), valores utilizados al compilar el modelo (carpeta *Configuración de compilación*) y entrenamiento de modelo (carpeta *Resumen de entrenamiento*).

Cuando primero examina un nodo de modelado, las carpetas en la pestaña Resumen se colapsan. Para ver los resultados del interés, utilice el control de expansión situado a la izquierda de la carpeta para mostrar los resultados, o pulse el botón **Expandir todo** para mostrar los resultados. Para ocultar los resultados después de visualizarlos, utilice el control de expansión para colapsar el archivo específico que desea ocultar, o pulse el botón **Colapsar todo** para colapsar todas las carpetas.

Utilización de nuggets de modelo de concepto en una ruta

Cuando utiliza un nodo de modelado de minería de textos, puede generar un nugget de modelo de concepto o un nugget de modelo de categoría (a través de una sesión de área de trabajo interactiva). El siguiente ejemplo muestra cómo utilizar un modelo de concepto en una ruta simple.

Ejemplo: Nodo de archivo de estadísticas con el nugget de modelo de concepto

El siguiente ejemplo muestra cómo utilizar el nugget de modelo de concepto de minería de textos.



Figura 3. Ruta de ejemplo: Nodo de archivo de estadísticas con un nugget de modelo de concepto de minería de textos

1. **Nodo de archivo de estadísticas (Pestaña Datos).** En primer lugar, añadimos este nodo a la ruta para especificar dónde están almacenados los documentos de texto.

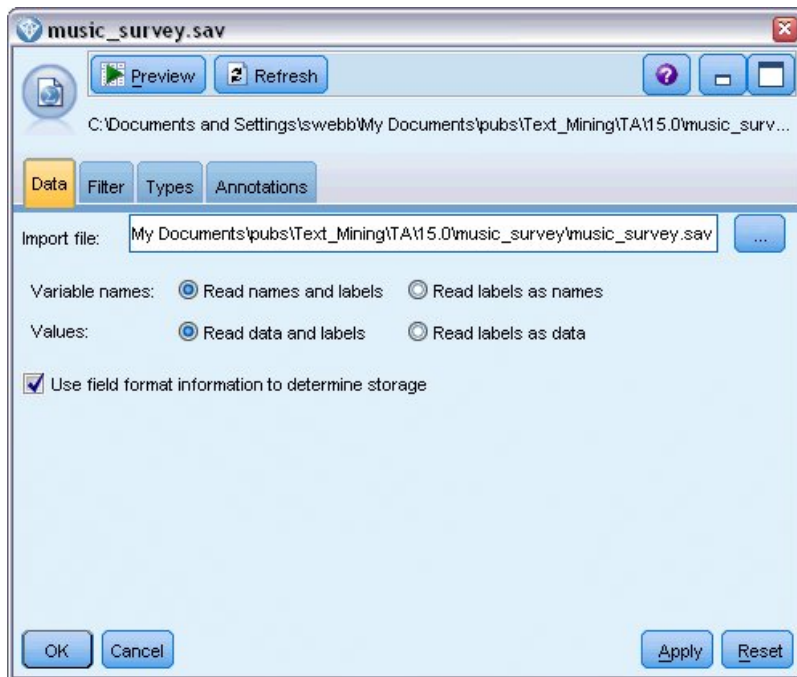


Figura 4. Cuadro de diálogo nodo Archivo de estadísticas: pestaña Datos

2. **Nugget de modelo de concepto de minería de textos (Pestaña Modelo).** A continuación, hemos añadido y conectado un nugget de modelo de concepto al nodo de archivo de estadísticas. Hemos seleccionado los conceptos que deseábamos utilizar para puntuar nuestros datos.

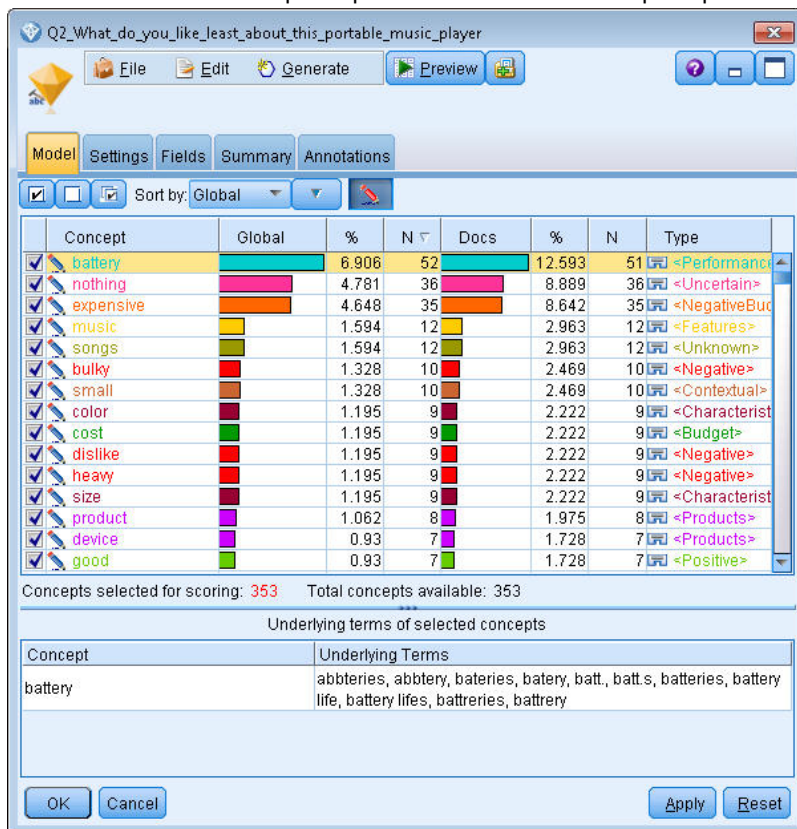


Figura 5. Recuadro de diálogo de nugget de modelo de minería de textos: Pestaña Modelo

3. **Nugget de modelo de concepto de minería de textos (Pestaña Valores).** A continuación, hemos definido el formato de salida y seleccionado *Conceptos como campos*. Se creará un campo nuevo

en la salida para cada concepto seleccionado en la pestaña Modelo. Cada nombre de campo estará constituido por el nombre del concepto y el prefijo "Concept_"

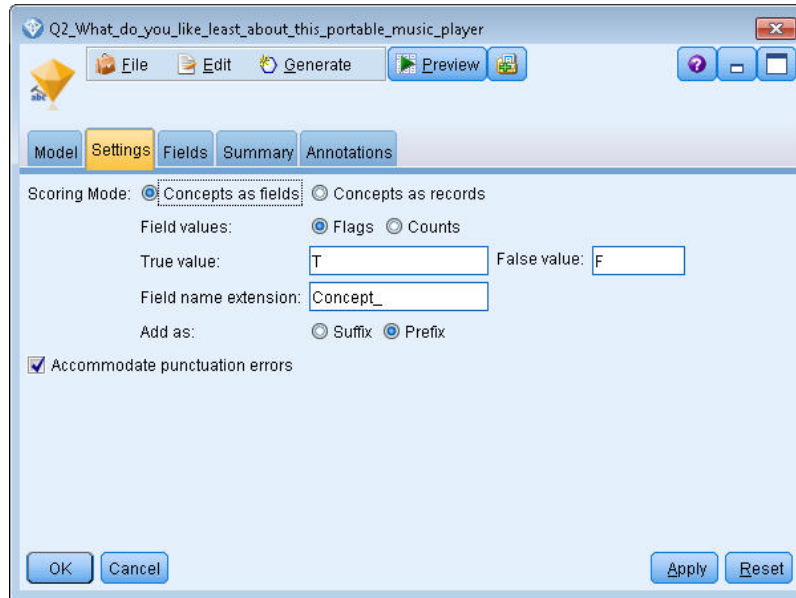


Figura 6. Recuadro de diálogo de nugget de modelo de concepto de minería de textos: Pestaña Valores

4. **Nugget de modelo de concepto de minería de textos (Pestaña Campos).** A continuación, hemos seleccionado el campo de texto, **Q2_What_do_you_like_least_about_this_portable_music_player**, que es el nombre de campo que proviene del nodo de archivo de estadísticas. También hemos seleccionado la opción **El campo de texto representa a: Texto real**.

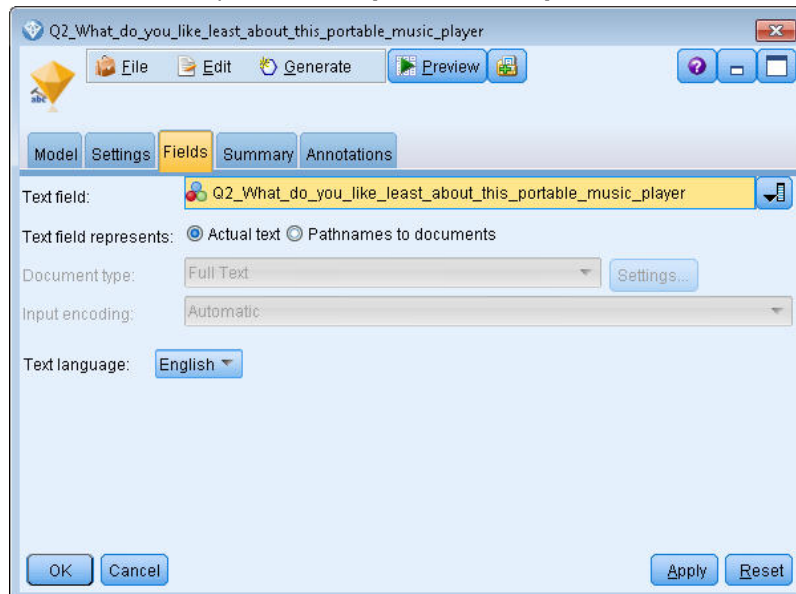


Figura 7. Recuadro de diálogo de nugget de modelo de concepto de minería de textos: Pestaña Campos

5. **Nodo de tabla.** A continuación, hemos adjuntado un nodo de tabla para ver los resultados y hemos ejecutado la ruta. La salida de la tabla se abre en la pantalla.

	Respondent_ID	Q1_What do you like most about this portable music player	Q2_What do you like least about this portable music player	Concept_reliable	Concept_downloading...	Concept_white color	Concept_limited
1	1	little, li...	expensive	F	F	F	F
2	2	The ba...	The screen is hard to see when outside.	F	F	F	F
3	3	cost a...	difficult software	F	F	F	F
4	4	Having...	Nothing, I love it!	F	F	F	F
5	5	The sh...	Battery life seems shorter than advertised.	F	F	F	F
6	6	Batter...	Ubiquitousness; everyone has one.	F	F	F	F
7	7	I like it...	I wish the 40GB model was still available. I have a 20GB model and need more memory.	F	F	F	F
8	8	portabi...	it doesn't have a light.	F	F	F	F
9	9	Small, ...	Nothing, I love it.	F	F	F	F
10	10	Able t...	it is in the shop due to a hardware failure.	F	F	F	F
11	11	It's por...	smudges on the display	F	F	F	F
12	12	Living i...	Battery life	F	F	F	F
13	13	mobility	Technical difficulties setting it up initially and managing the library of songs on my PC.	F	F	F	F
14	14	I like th...	it is a little heavy, and the battery life isn't long enough.	F	F	F	F
15	15	it hold...	Battery life.	F	F	F	F
16	16	It's fun...	nothing	F	F	F	F
17	17	its cool	battery	F	F	F	F
18	18	lots of ...	it was very expensive	F	F	F	F
19	19	Others...	I find the controls hard to use.	F	F	F	F
20	20	lightw...	so small afraid I'll lose it easily	F	F	F	F

Figura 8. Salida de tabla desplazada para mostrar los distintivos del concepto

Nugget de minería de textos: Modelo de categoría

Un nugget de modelo de categoría de minería de textos se crea siempre que genera un modelo de categoría desde dentro del área de trabajo interactiva. Este nugget de modelado contiene un conjunto de categorías, cuya definición se compone de conceptos, tipos, patrones de TLA y/o reglas de categoría. El nugget se utiliza para categorizar respuestas de encuestas, entradas de blog, otros canales de información de la web, y cualquier otro dato de texto.

Si inicia una sesión de área de trabajo interactiva en el nodo de modelado, puede explorar los resultados de la extracción, perfeccionar los recursos, afinar las categorías antes de generar modelos de categoría. Cuando ejecuta una ruta que contiene un nugget de modelo de minería de textos, se añaden campos nuevos a los datos en función de la modalidad de creación seleccionada en la pestaña Modelo del nodo de modelado de minería de textos antes de crear el modelo. Para obtener más información, consulte el tema “Nugget de modelo de categoría: Pestaña Modelo” en la página 39.

Si el nugget de modelo se ha generado utilizando documentos traducidos, la puntuación se realizará en el idioma traducido. Del mismo modo, si el nugget de modelo se ha generado utilizando el inglés como idioma, puede especificar un idioma de traducción en el nugget de modelo, ya que los documentos se traducirán al inglés.

Los nugget de modelo de minería de datos de texto se colocan en la paleta de nugget de modelo (ubicada en la pestaña Modelos en la parte superior derecha de la ventana de IBM SPSS Modeler) cuando se generan.

Visualización de los resultados

Para ver información sobre el nugget de modelo, pulse con el botón derecho el nodo en la paleta de nuggets de modelo y elija **Examinar** en el menú contextual(o **Editar** en los nodos de una ruta).

Añadir modelos a una ruta

Para añadir el nugget de modelo a una ruta, pulse el icono en la paleta nuggets de modelo y pulse el lienzo de rutas donde desea ubicar el nodo. O pulse con el botón derecho el icono y elija **Añadir a ruta** en el menú contextual. A continuación, conecte su ruta al nodo y ya está listo para pasar los datos para generar predicciones.

Precaución: si desea utilizar un nugget de puntuación para volver a generar un nodo de modelado que contenga el modelo de categoría y la plantilla utilizada, se recomienda crear un TAP y utilizarlo en una sesión interactiva, en lugar del nodo de modelado, antes de generar el nugget de puntuación.

Nugget de modelo de categoría: Pestaña Modelo

Para modelos de categoría, la pestaña modelo muestra la lista de categorías en el modelo de categoría en la izquierda y los descriptores para una categoría seleccionada en la derecha. Cada categoría se compone de un número de descriptores. Para cada categoría que seleccione, aparecen los descriptores asociados en la tabla. Estos descriptores pueden incluir conceptos, reglas de categoría, tipos y patrones de TLA. También se muestra el tipo de cada descriptor, así como algunos ejemplos de lo que cada descriptor representa.

En esta pestaña, el objetivo es seleccionar las categorías que desea utilizar para la puntuación. Para un modelo de categoría, los documentos y registros se puntúan en categorías. Si un documento o registro contiene uno o varios de los descriptores en su texto o algún término subyacente, ese documento o registro se asigna a la categoría a la que pertenece el descriptor. Estos términos subyacentes incluyen los sinónimos definidos en los recursos lingüísticos (independientemente de si se encontraron en el texto o no), así como cualquier término en plural/singular extraído encontrado en el texto utilizado para generar el nugget de modelo, términos permutados, términos de agrupación difusa, etc.

Nota: Si ha generado un nugget de modelo de concepto en su lugar, esta pestaña contendrá resultados diferentes. Para obtener más información, consulte el tema [“Modelo de concepto: Pestaña Modelo”](#) en la [página 31](#).

Árbol de categorías

Para conocer más sobre cada categoría, seleccione la categoría y revise la información que aparece para los descriptores de esa categoría. Para cada descriptor, puede revisar la siguiente información:

- Nombre de **descriptor**. Este campo contiene un icono que representa el tipo de descriptor que es, así como el nombre del descriptor.

Tabla 5. Iconos de descriptor

	Conceptos		Patrones de TLA
	Tipos		Reglas de categorías

- **Tipo.** Este campo contiene el nombre de tipo para el descriptor. Los tipos son recopilaciones de conceptos similares (agrupaciones semánticas) como, por ejemplo, nombres de organización, productos, u opiniones positivas. Las reglas no se asignan a tipos.
- **Detalles.** Este campo contiene una lista de lo que se incluye en el descriptor. En función del número de coincidencias, puede que no vea la lista completa para cada descriptor debido a limitaciones de tamaño en el recuadro de diálogo.

Selección y copia de categorías

Todas las categorías superiores están seleccionadas para la puntuación de forma predeterminada, tal como se muestra en los recuadros de selección del panel izquierdo. Un recuadro seleccionado significa que la categoría se utilizará para la puntuación. Un recuadro no seleccionado significa que la categoría se excluirá de la puntuación. Puede seleccionar varias filas seleccionándolas y pulsando uno de los recuadros de selección en su selección. Además, si se selecciona una categoría o subcategoría pero no se selecciona una de sus subcategorías, el recuadro de selección muestra un fondo azul para indicar que sólo hay una selección parcial en el hijo de la categoría seleccionada.

Pulsando una categoría del árbol con el botón derecho del ratón, puede mostrar un menú contextual desde el que puede:

- **Seleccionar seleccionados.** Selecciona todos los recuadros de selección para las filas seleccionadas en la tabla.
- **Deseleccionar seleccionados.** Deselecciona todos los recuadros de selección para las filas seleccionadas en la tabla.
- **Seleccionar todo.** Selecciona todos los recuadros de selección en la tabla. Esto da como resultado la utilización de todas las categorías en la salida final. También puede utilizar el icono de recuadro de selección correspondiente en la barra de herramientas.

- **Deseleccionar todo.** Deselecciona todos los recuadros de selección en la tabla. La desección de una categoría significa que no se utilizará en la salida final. También puede utilizar el icono de recuadro de selección vacío correspondiente en la barra de herramientas.

Pulsando una celda de la tabla de descriptor con el botón derecho del ratón, puede mostrar un menú contextual en el que puede:

- **Copiar.** Los conceptos seleccionados se copian en el portapapeles.
- **Copiar con campos.** El descriptor seleccionado se copia en el portapapeles junto con las cabeceras de columna.
- **Seleccionar todo.** Se seleccionarán todas las filas de la tabla.

Nugget de modelo de categoría: pestaña Configuración

La pestaña Valores se utiliza para definir el valor de campo de texto para los nuevos datos de entrada, de ser necesario. También es el lugar donde define el modelo de datos para su salida (modalidad de puntuación).

Nota: Esta pestaña aparece en el recuadro de diálogo del nodo sólo cuando el nugget de modelo se encuentra en el lienzo o en una ruta. No existe cuando accede a este nugget directamente en la paleta de modelos.

Modalidad de puntuación: Categorías como campos

Con esta opción, existe la misma cantidad de registros de salida que existía en la entrada. Sin embargo, ahora cada registro contiene un campo nuevo para cada categoría que se ha seleccionado (utilizando la marca de selección) en la pestaña Modelo. Para cada campo, especifique un valor de distintivo para **True** y para **False**, como *Sí/No*, *True/False*, *T/F*, o *1 y 2*. Los tipos de almacenamiento se establecen automáticamente para reflejar los valores elegidos. Por ejemplo, si especifica valores numéricos para los distintivos, se manejarán automáticamente como un valor entero. Los tipos de almacenamiento de las marcas pueden ser una cadena, un número entero, un número real o la fecha/hora.

Nota: Si está utilizando conjuntos de datos de gran tamaño, por ejemplo, con una base de datos Db2, el uso de **Categorías como campos** puede encontrar problemas de proceso debido a la cantidad de datos. En este caso, se recomienda utilizar **Categorías como registros** en su lugar.

Extensión de nombre de campo. Puede elegir especificar un prefijo/sufijo de extensión para el nombre de campo o puede elegir utilizar los códigos de categoría. Los nombres de campo se generan al utilizar el nombre de categoría más esta extensión.

- **Añadir como.** Especifique dónde debe añadirse la extensión del nombre de campo. Elija **Prefijo** para añadir la extensión al inicio de la cadena. Elija **Sufijo** para añadir la extensión al final de la cadena.

Si una subcategoría no está seleccionada. Esta opción le permite especificar cómo se manejarán los descriptores que pertenecen a subcategorías que no fueron seleccionadas para la puntuación. Existen dos opciones.

- La opción **Excluir sus descriptores completamente de puntuar** provocará que los descriptores de subcategorías que no tienen marcas de selección (no seleccionadas) sean ignoradas y no utilizadas durante la puntuación.
- La opción **Agregar descriptores con aquellos en la categoría padre** hará que los descriptores de subcategorías que no tienen marca de selección (no seleccionados) se utilicen como descriptores para la categoría padre (la categoría por encima de esta subcategoría). Si varios niveles de subcategorías y no seleccionados, los descriptores se desplegarán en la primera categoría padre disponible.

Puntuar solo la categoría que coincida con el nivel inferior. Utilice esta opción para generar la salida solo en una línea (por ejemplo, si la categoría es `GeneralSatisfaction/Pos`, la selección de esta opción genera `GeneralSatisfaction/Pos`. Sin esta opción, obtendrá dos líneas: `GeneralSatisfaction` y `GeneralSatisfaction/Pos`).

Arreglar errores de puntuación. Esta opción normaliza temporalmente el texto que contiene errores de puntuación (por ejemplo, el uso inapropiado) durante la extracción para mejorar la capacidad de

extracción de los conceptos. Esta opción es extremadamente útil cuando el texto es breve y de calidad mediocre (como, por ejemplo, en respuestas de encuestas con final abierto, correo electrónico y datos CRM) o cuando el texto contiene muchas abreviaturas.

Modalidad de puntuación: Categorías como registros

Con esta opción, se crea un registro nuevo para cada par de category y document. Generalmente, hay más registros en la salida que los que había en la entrada. Además de los campos de entrada, también se añaden campos nuevos a los datos en función del tipo de modelo que sea.

Tabla 6. Campos de salida para "Categorías como registros"	
Campo de salida nuevo	Descripción
Category	Contiene el nombre de la categoría a la que se asignó el documento de texto. Si la categoría es una subcategoría de otra, la vía de acceso completa al nombre de la categoría está controlada por el valor que elija en este diálogo.

Valores para categorías jerárquicas. Esta opción controla cómo se muestran los nombres de las subcategorías en la salida.

- **Ruta de categoría completa.** Esta opción mostrará el nombre de la categoría y la vía de acceso completa de las categorías principales mediante la utilización de barras diagonales para separar los nombres de las categorías de los nombres de las subcategorías.
- **Ruta de categoría abreviada.** Esta opción mostrará únicamente el nombre de la categoría utilizando puntos suspensivos para mostrar el número de categorías principales para esa categoría en cuestión.
- **Categoría de nivel inferior.** Esta opción mostrará únicamente el nombre de la categoría sin mostrar la vía de acceso completa o las categorías principales.

Si una subcategoría no está seleccionada. Esta opción le permite especificar cómo se manejarán los descriptores que pertenecen a subcategorías que no fueron seleccionadas para la puntuación. Existen dos opciones.

- La opción **Excluir sus descriptores completamente de puntuar** provocará que los descriptores de subcategorías que no tienen marcas de selección (no seleccionadas) sean ignoradas y no utilizadas durante la puntuación.
- La opción **Agregar descriptores con aquellos en la categoría padre** hará que los descriptores de subcategorías que no tienen marca de selección (no seleccionados) se utilicen como descriptores para la categoría padre (la categoría por encima de esta subcategoría). Si varios niveles de subcategorías y no seleccionados, los descriptores se desplegarán en la primera categoría padre disponible.

Arreglar errores de puntuación. Esta opción normaliza temporalmente el texto que contiene errores de puntuación (por ejemplo, el uso inapropiado) durante la extracción para mejorar la capacidad de extracción de los conceptos. Esta opción es extremadamente útil cuando el texto es breve y de calidad mediocre (como, por ejemplo, en respuestas de encuestas con final abierto, correo electrónico y datos CRM) o cuando el texto contiene muchas abreviaturas.

Nugget de modelo de categoría: Otras pestañas

La pestaña Campos y la pestaña Valores para el nugget de modelo de categoría son las mismas que para el nugget de modelo de concepto.

- Pestaña Campos. Para obtener más información, consulte el tema [“Modelo de concepto: pestaña Campos”](#) en la página 34.
- Pestaña Resumen. Para obtener más información, consulte el tema [“Modelo de concepto: pestaña Resumen”](#) en la página 35.

Utilización de nuggets de modelo de categoría en una ruta

El nugget de modelo de categoría de minería de textos se genera a partir de una sesión de área de trabajo interactiva. Puede utilizar este nugget de modelo en una ruta.

Ejemplo: Nodo de archivo de estadísticas con el nugget de modelo de categoría

El siguiente ejemplo muestra cómo utilizar el nugget de modelo de minería de textos.



Figura 9. Ruta de ejemplo: Nodo de archivo de estadísticas con un nugget de modelo de categoría de minería de textos

1. **Nodo de archivo de estadísticas (Pestaña Datos).** En primer lugar, añadimos este nodo a la ruta para especificar dónde están almacenados los documentos de texto.

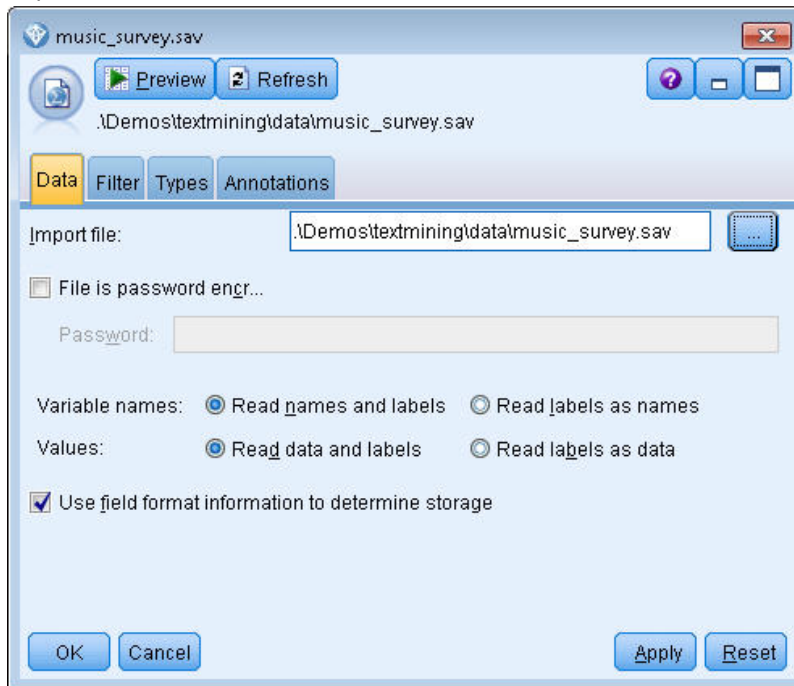


Figura 10. Cuadro de diálogo nodo Archivo de estadísticas: pestaña Datos

2. **Nugget de modelo de categoría de minería de textos (Pestaña Modelo).** A continuación, hemos añadido y conectado un nugget de modelo de categoría al nodo de archivo de estadísticas. Hemos seleccionado las categorías que deseábamos utilizar para puntuar nuestros datos.

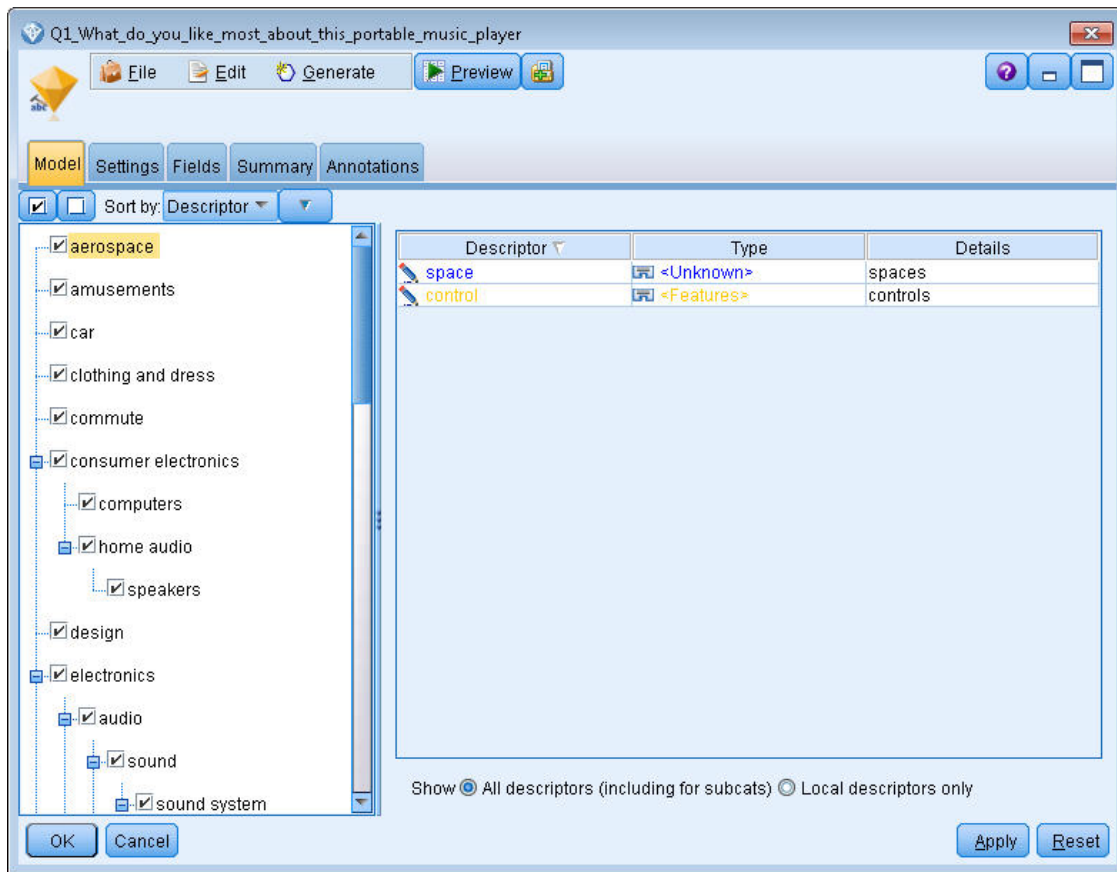


Figura 11. Recuadro de diálogo de nugget de modelo de minería de textos: Pestaña Modelo

3. **Nugget de modelo de minería de textos (Pestaña Valores).** A continuación, hemos definido el formato de salida **Categorías como campos**.

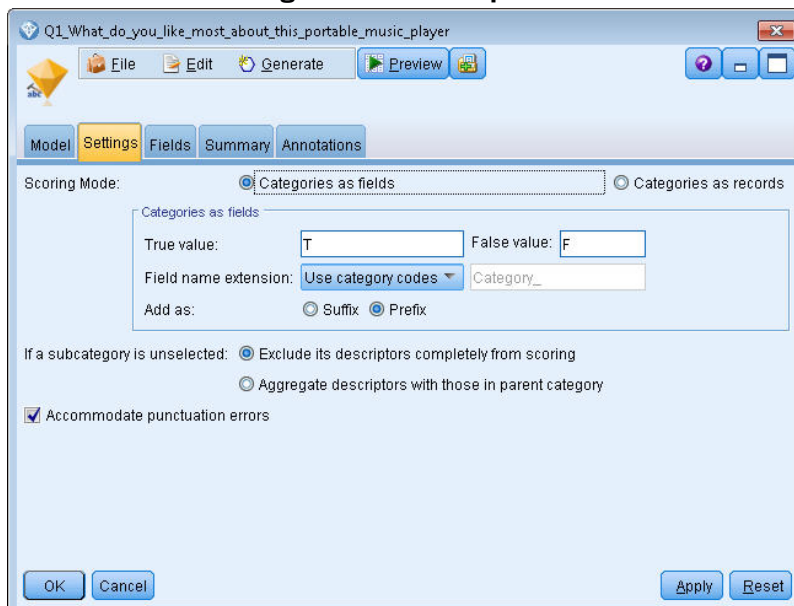


Figura 12. Recuadro de diálogo de nugget de modelo de categoría: Pestaña Valores

4. **Nugget de modelo de categoría de minería de textos (Pestaña Campos).** A continuación, hemos seleccionado la variable de campo de texto, que es el nombre de campo que proviene del nodo de archivo de estadísticas, y hemos seleccionado la opción El campo de texto representa a **Texto real**, así como otros valores.

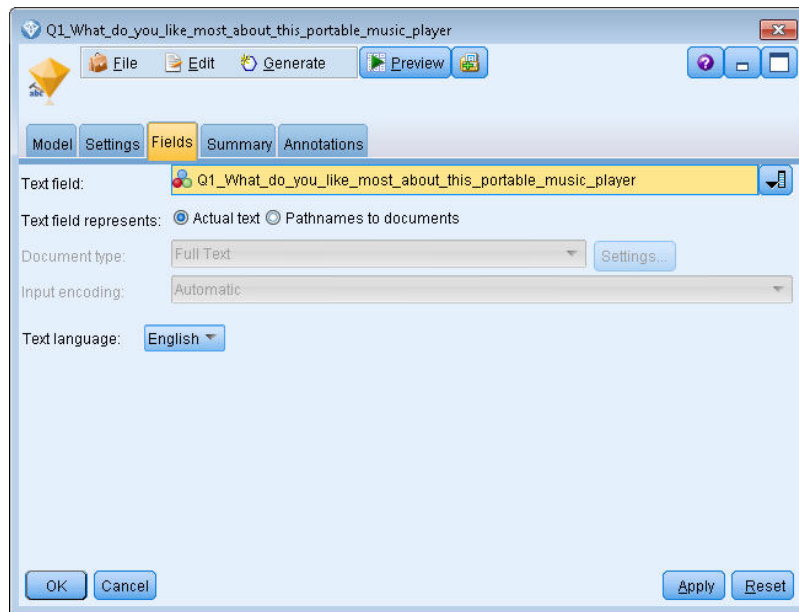


Figura 13. Recuadro de diálogo de nugget de modelo de minería de textos: Pestaña Campos

5. **Nodo de tabla.** A continuación, hemos adjuntado un nodo de tabla para ver los resultados y hemos ejecutado la ruta.

Table (3 fields, 844 records)

	ID	Q1_What_do_you_like_most_about_this_portable_music_player	Category
1	1	little, light	light
2	2	The battery power is great.	light
3	2	The battery power is great.	electronics/battery
4	2	The battery power is great.	electronics
5	3	cost and size	size
6	6	Battery life. Portability. Accessories. Style.	light
7	6	Battery life. Portability. Accessories. Style.	electronics/battery
8	6	Battery life. Portability. Accessories. Style.	electronics
9	7	I like its ability to store all of my music. I also like the ability to create playlists.	playlists
10	7	I like its ability to store all of my music. I also like the ability to create playlists.	light
11	7	I like its ability to store all of my music. I also like the ability to create playlists.	music
12	8	portability, capacity, sound quality, durability	light
13	8	portability, capacity, sound quality, durability	electronics/audio/sound
14	8	portability, capacity, sound quality, durability	electronics/audio

Figura 14. Salida de tabla

Capítulo 4. Minería para enlaces de texto

Nodo Análisis para los enlaces de texto

El nodo de análisis de enlaces de texto (TLA) añade una tecnología de coincidencia de patrón a la extracción de conceptos de minería de textos para poder identificar relaciones entre los conceptos en los datos de texto basadas en patrones conocidos. Estas relaciones pueden describir como se siente un cliente respecto a un producto, cuáles son las empresas que están haciendo negocios juntas o hasta las relaciones entre genes y agentes farmacéuticos.

Por ejemplo, puede ser que no le interese demasiado conseguir el nombre de producto de la competencia. Al utilizar este nodo, también puede enterarse lo que piensa la gente de este producto, si es que existen este tipo de opiniones en los datos. Las relaciones y asociaciones se identifican y extraen al hacer coincidir patrones conocidos con los datos de texto.

Puede utilizar las reglas de patrón TLA dentro de ciertas plantillas de recurso que vienen con IBM SPSS Modeler Text Analytics o crear/editar las propias. Las reglas de patrones están hechas de macros, listas de palabras y espacios entre palabras para formar una consulta Booleana, o regla, que se compara al texto de entrada. Siempre que una regla de patrón TLA coincida con un texto, este texto puede indicarse como un resultado TLA y reestructurado como datos de salida. Para obtener más información, consulte el tema [Capítulo 18, “Sobre las reglas de enlaces de texto”, en la página 211](#).

El nodo de Análisis de enlaces de texto ofrece una forma más directa de identificar y extraer resultados de patrones TLA desde el texto y después añadir los resultados al conjunto de datos en la ruta. Pero el nodo de Análisis de enlaces de texto no es la única forma en la que puede realizar el análisis de enlaces de texto. También puede utilizar una sesión de área de trabajo interactiva en el nodo de modelado de la Minería de textos.

En el área de trabajo interactiva, puede explorar los resultados del patrón TLA y utilizarlos como descriptores de categoría y/o para obtener más información sobre los resultados utilizando desplegables y gráficos. Para obtener más información, consulte el tema [Capítulo 11, “Exploración del Análisis de enlaces de texto”, en la página 149](#). De hecho, la utilización del nodo de Minería de textos para extraer resultados de TLA es una buena manera de explorar y refinar las plantillas de sus datos para ser utilizadas posteriormente directamente en el nodo de TLA.

La salida se puede representar hasta en 6 ranuras o partes. Para obtener más información, consulte el tema [“Salida de nodo TLA” en la página 48](#).

Puede encontrar este nodo en la pestaña IBM SPSS Modeler Text Analytics de la paleta de nodos, en la parte inferior de la ventana de IBM SPSS Modeler. Para obtener más información, consulte el tema [“IBM SPSS Modeler Text Analytics nodos” en la página 7](#).

Requisitos. El nodo de Análisis de enlaces de texto acepta datos de texto que se leen en un campo utilizando cualquiera de los nodos fuente estándar (nodo Base de datos, nodo Archivo sin formato, etc.) o se leen en un campo que lista vías de acceso a documentos externos generados por un nodo de Lista de archivos o un nodo Canal de información web.

Puntos fuertes. El nodo de Análisis de enlaces de texto va más allá de la extracción de conceptos básica para proporcionar información sobre las relaciones *entre* conceptos, así como también opiniones relacionadas o calificadores que pueden ser revelados en los datos.

Nodo Análisis para los enlaces de texto: pestaña Campos

Utilice la pestaña Campos para especificar los valores de campo para los datos de los que se van a extraer conceptos. Puede establecer los siguientes parámetros:

Campo de ID. Seleccione el campo que contiene el identificador para los registros de texto. Los identificadores deben ser enteros. El campo de ID funciona como un índice para los registros de texto individuales. Utilice un campo de ID si el campo de texto representa el texto a ser minado.

Campo de texto. Seleccione el campo que contiene el texto donde se va a realizar la minería. Este campo depende del origen de datos.

Campo de idioma. Seleccione el campo que contenga el identificador del idioma ISO de dos letras. Si no selecciona ningún campo, se presupone que el idioma de cada documento es de la plantilla proporcionada.

Tipo de documento. El tipo de documento especifica la estructura del texto. Seleccione uno de los siguientes tipos:

- **Texto completo.** Se utiliza para la mayoría de los documentos o fuentes de texto. El conjunto de texto completo se explora para su extracción. A diferencia de las otras opciones, no hay valores adicionales para esta opción.
- **Texto estructurado.** Se utiliza para formularios bibliográficos, patentes y cualquier campo que contenga estructuras regulares que puedan identificarse y analizarse. Este tipo de documento se utiliza para omitir todo o parte del proceso de extracción. Le permite definir separadores de términos, asignar tipos e imponer un valor de frecuencia mínima. Si selecciona esta opción, debe pulsar el botón **Valores** y especificar separadores de texto en **Formato de texto estructurado**, del cuadro de diálogo Valores del documento. Para obtener más información, consulte el tema [“Pestaña Valores de documento para campos” en la página 22](#).

Unidad textual. Seleccione la modalidad de extracción de las siguientes:

- **Modalidad de documento.** Se utiliza para documentos cortos y homogéneos semánticamente, como los artículos de las agencias de noticias.
- **Modalidad de párrafo.** Se utiliza para las páginas web y para los documentos sin etiquetar. El proceso de extracción divide semánticamente los documentos, tomando ventaja de características como etiquetas internas y sintaxis. Si se selecciona esta modalidad, la puntuación se aplica párrafo a párrafo. Por lo tanto, por ejemplo, la regla `apple & orange` es verdadera sólo si `apple` y `orange` se encuentran en el mismo párrafo.

Nota: Debido a la forma de extraer el texto de los documentos en PDF, la **Modalidad de párrafo** no funciona en estos documentos. Esto se debe a que la extracción suprime el marcador de retorno de carro.

Configuración modalidad párrafo. Esta opción solo está disponible si establece la opción de unidad textual en **Modalidad de párrafo**. Especifique el umbral de carácter a utilizar en cualquier extracción. El tamaño actual se redondea hacia arriba o hacia abajo hacia el punto más próximo. Para asegurar que las asociaciones de palabras producidas desde el texto de la colección de documentos son representativas, evite especificar un tamaño de extracción que sea demasiado pequeño.

- **Mínimo.** Especifique el número mínimo de caracteres a utilizar en cualquier extracción.
- **Máximo.** Especifique el número máximo de caracteres a utilizar en cualquier extracción.

Copiar recursos de. Cuando se extrae texto, la extracción no se basa sólo en los valores de la pestaña Experto, sino también en los recursos lingüísticos. Estos recursos sirven como la base de cómo manejar y procesar el texto durante una extracción para obtener los conceptos, tipos y, en algunas ocasiones, los patrones TLA. Puede copiar recursos en este nodo desde una plantilla de recursos.

Una plantilla de recursos es un conjunto predefinido de bibliotecas y recursos lingüísticos y no lingüísticos avanzados que se han afinado para un determinado dominio o uso. Estos recursos sirven como la base de cómo manejar y procesar datos durante la extracción. Pulsar **Cargar** y seleccionar la plantilla desde la que se van a copiar los recursos.

Las plantillas se cargan cuando las selecciona y no cuando se ejecuta la ruta. En el momento de la carga, una copia de los recursos se almacena en el nodo. Por lo tanto, si alguna vez desea utilizar una plantilla actualizada, la tendrá que volver a cargar aquí. Para obtener más información, consulte el tema [“Copiar recursos de plantillas y TAP” en la página 26](#).

Idioma del texto. Identifica el idioma del texto que se está minando. Los recursos copiados en el nodo controlan las opciones de idioma presentadas. Seleccione el idioma para el que se hayan ajustado los recursos.

Nodo Análisis de enlaces de texto: pestaña Experto

En este nodo, la extracción de resultados de patrón de análisis de enlaces de texto (TLA) se habilita automáticamente. La pestaña Experto contiene ciertos parámetros adicionales que afectan cómo se extrae y maneja el texto. Los parámetros en este recuadro de diálogo controlan el comportamiento básico, así como unos pocos comportamientos avanzados, del proceso de extracción. También hay un número de opciones y recursos lingüísticos que también afectan los resultados de extracción, que son controlados por la plantilla de recursos que seleccione.

Limite la extracción a conceptos con una frecuencia global de al menos [n]. Especifica el número mínimo de veces que debe aparecer una palabra o frase en el texto para que pueda extraerse. De esta forma, un valor de 5 limita la extracción a aquellas palabras o frases que aparecen al menos cinco veces en todos los registros o documentos.

En algunos casos, el hecho de cambiar este límite puede suponer una diferencia sustancial en los resultados extraídos y, por consiguiente, en las categorías. Digamos que trabaja con datos de un restaurante y no incrementa el límite por encima de 1 para esta opción. En ese caso, podría encontrar *pizza* (1), *pizza fina* (2), *pizza vegetal* (2) y *pizza favorita* (2) en los resultados de extracción. Sin embargo, si limitara la extracción con una frecuencia global de 5 o más y volviera a extraer los datos, dejarían de aparecer tres de estos conceptos. Así pues, obtendría *pizza* (7), puesto que *pizza* es la forma más simple y esta palabra se encontraba también como posible candidata. Y en función del resto del texto, podría de hecho tener una frecuencia mayor que siete en función de si hay otras frases con la palabra *pizza* en su interior. Además, si *pizza vegetal* ya era un descriptor de categoría, podría añadir *pizza* como descriptor en lugar de capturar todos los registros. Por esta razón, se recomienda que cambie este límite con precaución siempre que ya se hayan creado categorías.

Tenga en cuenta que esta es una característica sólo de extracción; si la plantilla contiene términos (que usualmente sucede) y se encuentra el término para la plantilla en el texto, este se indexará independientemente de su frecuencia.

Por ejemplo, supongamos que utiliza una plantilla de Recursos básicos que incluye "los angeles" en el tipo <Location> en la biblioteca núcleo; si el documento contiene Los Angeles sólo una vez, Los Angeles formará parte de la lista de conceptos. Para prevenirlo necesitará establecer un filtro que muestre conceptos que sucedan al menos la misma cantidad de veces que el valor especificado en el campo **Limitar extracción a conceptos con una frecuencia global de al menos [n].**

Arreglar errores de puntuación. Esta opción normaliza temporalmente el texto que contiene errores de puntuación (por ejemplo, el uso inapropiado) durante la extracción para mejorar la capacidad de extracción de los conceptos. Esta opción es extremadamente útil cuando el texto es breve y de calidad mediocre (como, por ejemplo, en respuestas de encuestas con final abierto, correo electrónico y datos CRM) o cuando el texto contiene muchas abreviaturas.

Acomodar la ortografía a una longitud mínima de caracteres de palabra de [n] Esta opción aplica una técnica de agrupación difusa que ayuda a agrupar bajo un concepto las palabras que suelen escribirse mal o que tienen una ortografía parecida. El algoritmo de agrupación difusa elimina temporalmente todas las vocales (excepto la primera) y elimina las consonantes dobles/triples de las palabras extraídas y luego las compara para ver si son iguales, de forma que *modeling* y *modelling* se agruparían. Sin embargo, si a cada término se le asigna un tipo diferente, excluyendo el tipo <Unknown>, la técnica de agrupación difusa no se aplicará.

También puede definir el número mínimo de caracteres *raíz* necesarios para poder utilizar la agrupación difusa. El número de caracteres raíz de un término se calcula sumando todos los caracteres y restando los que forman los sufijos de las declinaciones, y en el caso de términos de palabras compuestas, también los determinantes y las preposiciones. Por ejemplo, el término *exercises* se contaría como 8 caracteres raíz en la forma "exercise," ya que la letra *s* al final de la palabra es una inflexión (forma plural. De forma similar, *apple sauce* se cuenta como 10 caracteres raíz ("apple sauce") y *manufacturing of cars* se cuenta como 16 caracteres raíz ("manufacturing car"). Este método de recuento de caracteres solo se utiliza para comprobar si debe aplicarse la agrupación difusa, pero no influye en la forma de coincidencia de las palabras.

Nota: Si encuentra que ciertas palabras posteriormente se agrupan incorrectamente, puede excluir pares de palabras de esta técnica declarándolos explícitamente en la sección **Agrupación difusa: Excepciones**

en la pestaña Recursos avanzados. Para obtener más información, consulte el tema [“Agrupación difusa”](#) en la página 201.

Extraer unitérminos Esta opción extrae palabras simples (unitérminos) siempre que la palabra no forme parte de una palabra compuesta, y si es un sustantivo o una categoría léxica no reconocida.

Extraer entidades no lingüísticas Esta opción extrae entidades no lingüísticas, como números de teléfono, números de la seguridad social, horas, fechas, monedas, dígitos, porcentajes, direcciones de correo electrónico y direcciones de HTTP. Puede incluir o excluir ciertos tipos de entidades no lingüísticas en la sección **Entidades no lingüísticas: Configuración** de la pestaña Recursos avanzados. Si se desactivan las entidades innecesarias, el motor de extracción no malgastará tiempo de proceso. Para obtener más información, consulte el tema [“Configuración”](#) en la página 205.

Algoritmo de mayúsculas Esta opción extrae términos simples y compuestos que no están en los diccionarios incorporados, siempre que la primera letra del término esté en mayúscula. Esta opción supone un buen método para extraer la mayoría de los nombres propios.

Agrupar nombres de persona parciales y completos juntos cuando sea posible Esta opción agrupa nombres que aparecen de diferente manera juntos en el texto. Esta característica es útil porque a menudo se hace referencia a los nombres completos al principio del texto, y más adelante se utiliza la versión abreviada. Esta opción intenta hacer coincidir cualquier unitérmino que tenga el tipo <Unknown> con la última palabra de cualquier término compuesto que se haya tipificado como <Person>. Por ejemplo, si se encuentra *garcía*, que inicialmente se tipificó como <Unknown>, el motor de extracción comprobará si hay algún término compuesto en el tipo <Person> con el término *garcía* como la última palabra, como en *juan garcía*. Esta opción no se aplica a los nombres propios, porque muchos de ellos no se extraen nunca como unitérminos.

Permutación máxima de palabra no funcional Esta opción especifica el número máximo de palabras no funcionales que debe haber para poder aplicar la técnica de permutación. Esta técnica de permutación agrupa frases similares que difieren entre sí solo en las palabras no funcionales (por ejemplo, de y el), independientemente de la flexión. Por ejemplo, digamos que ha establecido este valor en dos palabras como máximo y tanto *company officials* como *officials of the company* se han extraído. En este caso, ambos términos extraídos se agruparían en la lista de conceptos final, ya que ambos términos se consideran iguales cuando se omite *of the*.

Utilizar la derivación al agrupar multitérminos Cuando procese grandes volúmenes de datos, seleccione esta opción para agrupar multitérminos mediante las reglas de derivación.

Salida de nodo TLA

Después de ejecutar el nodo Análisis de enlaces de texto, los datos se reestructuran. Es importante entender la forma en que la minería de textos reestructura los datos. Si desea una estructura diferente para la minería de datos, puede utilizar nodos en la paleta Operaciones de campo para lograrlo. Por ejemplo, si trabaja con datos en los que cada fila representa un registro de texto, entonces una fila se crea para cada patrón descubierto en los datos de texto de origen. Por cada fila en la salida, hay 15 campos:

- Seis campos (**Concept#**, como **Concept1**, **Concept2**, ..., y **Concept6**) representan los conceptos encontrados en la coincidencia de patrón.
- Seis campos (**Type#**, como **Type1**, **Type2**, ..., y **Type6**) representan el tipo de cada concepto.
- **Nombre de regla** representa el nombre de la regla de vínculo de texto que se utiliza en la coincidencia del texto y que produce la salida.
- Un campo que utiliza el nombre del campo ID que especificó en el nodo y que representa el registro o ID de documento como estaba en los datos de entrada
- **Texto coincidente** representa la porción de datos de texto en el registro o documento original que coincidió con el patrón TLA.

Nota: Cualquier ruta existente previamente que contenga un nodo Análisis de enlaces de texto de un release anterior a 5.0 puede que no sea totalmente ejecutable hasta que actualice los nodos. Algunas mejoras en versiones posteriores de IBM SPSS Modeler requieren que los nodos más antiguos se

reemplacen por versiones más nuevas, las cuales tienen una mayor capacidad de despliegue y una mayor eficacia.

También es posible realizar una traducción automática de ciertos idiomas. Esta característica le permite minar documentos en un idioma que no pueda hablar o leer. Si desea utilizar esta función de traducción, debe tener acceso a Software como servicio SDL (SaaS).

Almacenamiento en caché de resultados TLA

Si almacena en la memoria caché, los resultados del análisis de enlaces de texto están en la ruta. Para evitar repetir la extracción de los resultados de análisis de enlaces de texto cada vez que se ejecuta la ruta, seleccione el nodo Análisis de enlace de texto y, en los menús, seleccione **Editar > Nodo > Memoria caché > Habilitar**. La próxima vez que se ejecute la ruta, la salida se almacena en memoria caché en el nodo. El icono del nodo muestra una pequeña gráfica de "documento" que cambia de blanca a verde cuando se llena la memoria caché. La memoria caché se conserva durante la duración de la sesión. Para conservar la memoria caché para otro día (después de que la ruta se cierre y se vuelva a abrir), seleccione el nodo y, en los menús, seleccione **Editar > Nodo > Memoria caché > Guardar memoria caché**. La próxima vez que abra la ruta, puede volver a cargar la memoria caché que guardó en vez de ejecutar la traducción nuevamente.

De manera alternativa, puede guardar o habilitar la memoria caché de un nodo al pulsar con el botón derecho el nodo y elegir **Caché** en el menú contextual.

Utilización del nodo Análisis para los enlaces de texto en una ruta

El nodo Análisis de enlaces de texto se utiliza para acceder a los datos y extraer conceptos en una ruta. Puede utilizar cualquier nodo fuente para acceder a los datos.

Ejemplo: nodo Archivo de estadísticas con el nodo Análisis de enlaces de texto

El siguiente ejemplo muestra cómo utilizar el nodo Análisis de enlaces de texto.



Figura 15. Ejemplo: nodo Archivo de estadísticas con el nodo Análisis de enlaces de texto

1. **Nodo de archivo de estadísticas (Pestaña Datos).** Primero, añadimos este nodo a la ruta para especificar dónde se almacena el texto.

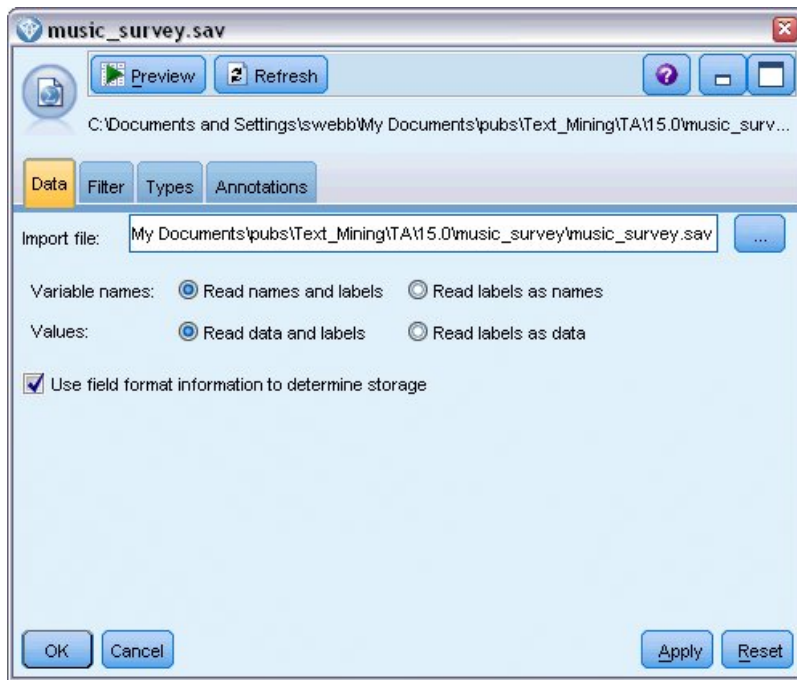


Figura 16. Cuadro de diálogo nodo Archivo de estadísticas: pestaña Datos

2. **Nodo Análisis de enlaces de texto (pestaña Campos).** A continuación, adjuntamos este nodo a la ruta para extraer conceptos para el modelado en sentido descendiente o su visualización. Especificamos el campo de ID y el nombre del campo de texto que contiene los datos, como también otros valores.

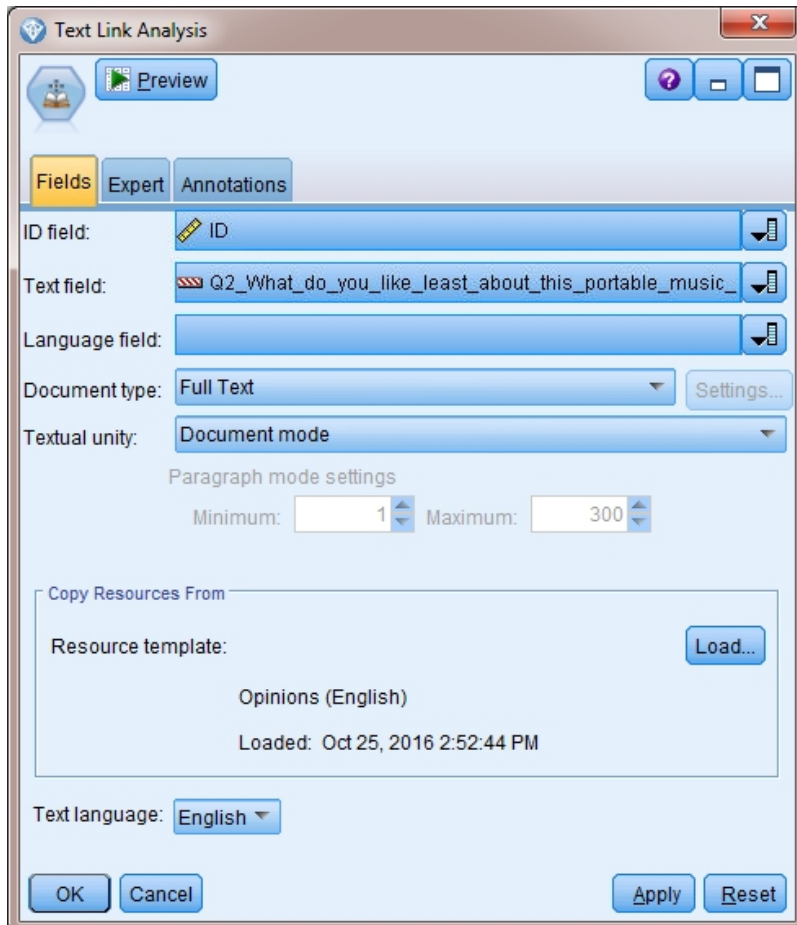


Figura 17. Cuadro de diálogo de nodo Análisis de enlaces de texto: pestaña Campos

3. **Nodo de tabla.** Por último, adjuntamos un nodo Tabla para ver los conceptos extraídos de nuestros documentos de texto. En la salida de tabla que se muestra, puede ver los resultados de patrón TLA que se encontraron en los datos después de ejecutarse esta ruta con un nodo Análisis de enlaces de texto. Algunos resultados muestran sólo que la coincidencia fue un concepto/tipo. En otros, los resultados son más complejos y contienen varios tipos y conceptos. Además, como resultado de ejecutar datos a través del nodo Análisis de enlaces de texto y extraer conceptos, se cambian varios aspectos de los datos. Los datos originales en el ejemplo contenían 8 campos y 405 registros. Después de ejecutar el nodo Análisis de enlaces de texto, ahora hay 15 campos y 640 registros. Ahora hay una fila para cada resultado de patrón TLA encontrado. Por ejemplo, ID 7 quedó a tres filas del original porque tres resultados de patrón TLA fueron extraídos. Puede utilizar un nodo Fusión si desea fusionar estos datos de salida con los datos originales.

	Concept1	Type1	Concept2	Type2	Conc...	Type3	Con...	Type4	Conc...	Type5	Con...	Type6	Rule Number	ID	Matched Text
1	expensive	NegativeBudget	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	1	<*"expensive">
2	screen	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	2	The <*"screen"> is <*"hard"> to see when outside
3	software	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0211_opinion + topic	3	<*"difficult"> <*"software">
4	nothing	Uncertain	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0153_topic/opinion	4	<*"Nothing"> <*"I love it">
5	like	Positive	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	4	Nothing , <*"I love it">
6	battery life	Unknown	too long	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	5	<*"Battery life"> seems <*"shorter"> than advertised
7	ubiquitousness	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0500_topic	6	<*"Ubiquitousness">
8	40gb model	Unknown	available	Posti...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	7	I wish the <*"40GB model"> was still <*"available">
9	20gb model	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*"20GB model"> and <*"need more"> <*"memory">
10	memory	Unknown	need more	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*"20GB model"> and <*"need more"> <*"memory">

Figura 18. Nodo de salida Tabla

Capítulo 5. Examinar texto de origen externo

Nodo Visor de archivos

Cuando esté realizando minería a una colección de documentos, puede especificar los nombres de vía de acceso completos de archivos directamente en los nodos de modelado de Minería de textos. Sin embargo, cuando realiza la salida a un nodo Tabla, sólo verá el nombre de vía de acceso completo de un documento en vez de el texto dentro de él. El nodo Visor de archivo se puede utilizar como análogo del nodo Tabla y permite acceder al texto dentro de cada uno de los documentos sin necesidad de fusionarlos en un solo archivo.

El nodo Visor de archivo puede ayudarle a entender mejor los resultados de una extracción de texto al proporcionarle acceso al texto fuente, o sin traducir, del que se extrajeron los conceptos ya que de otra forma sería inaccesible en la ruta. Este nodo se añade a la ruta después de un nodo Lista de archivo para obtener una lista de enlaces a todos los archivos.

El resultado de este nodo es una ventana que muestra todos los elementos del documento que se leyeron y usaron para extraer conceptos. Desde esta venta, puede presionar el icono de la barra de herramientas para iniciar el informe en un navegador externo listando los nombres de documentos como hiperenlaces. Puede pulsar un enlace para abrir el documento correspondiente en la colección. Para obtener más información, consulte el tema [“Utilización del nodo Visor de archivo”](#) en la página 53.

Puede encontrar este nodo en la pestaña IBM SPSS Modeler Text Analytics de la paleta de nodos, en la parte inferior de la ventana de IBM SPSS Modeler. Para obtener más información, consulte el tema [“IBM SPSS Modeler Text Analytics nodos”](#) en la página 7.

Nota: Cuando trabaja en modalidad cliente-servidor y los nodos Visor de archivo son parte de la ruta, las colecciones de documentos deben almacenarse en un directorio de servidor web en el servidor. Dado que el nodo de salida de Minería de textos produce una lista de documentos que se almacena en el directorio del servidor web, la configuración de seguridad del servidor web gestiona los permisos para estos documentos.

Configuración del nodo Visor de archivos

Puede especificar los siguientes valores para el nodo Visor de archivos.

Campo de documentos. Seleccione el campo en los datos que contienen el nombre completo y la vía de acceso de los documentos a visualizar.

Título para página HTML generada. Cree un título que aparezca en la parte superior de la página que contenga la lista de documentos.

Utilización del nodo Visor de archivo

El siguiente ejemplo muestra cómo utilizar el nodo Visor de archivo.

Ejemplo: nodos Lista de archivos y Visor de archivos



Figura 19. Ruta que ilustra la utilización del nodo Visor de archivos

1. **Nodo Lista de archivos (pestaña Configuración).** En primer lugar, añadimos este nodo para especificar dónde se ubican los documentos.

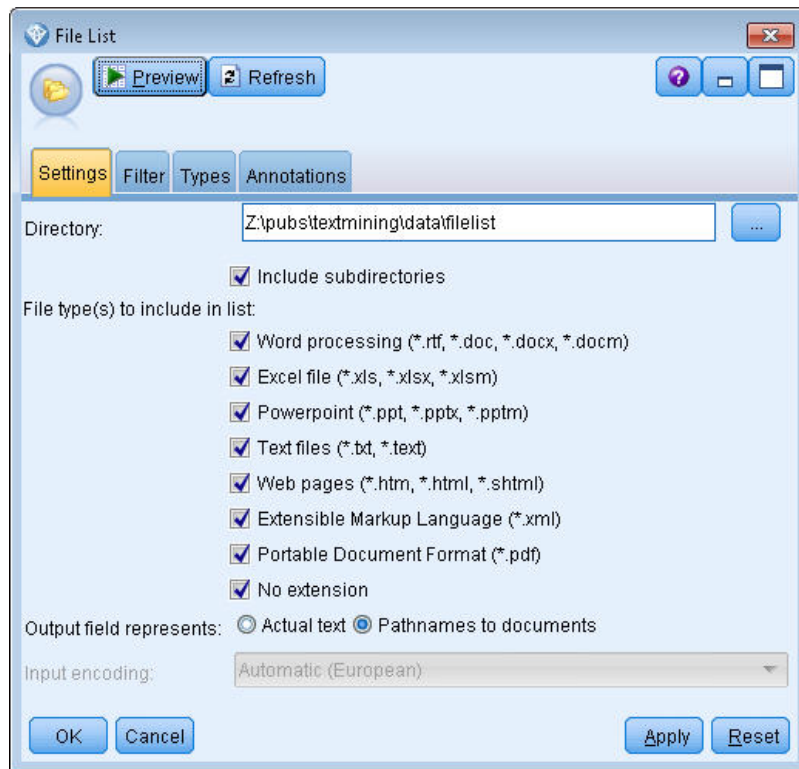


Figura 20. Cuadro de diálogo Lista de archivos: pestaña Configuración

2. **Nodo Visor de archivos (pestaña Configuración).** A continuación, adjuntamos el nodo Visor de archivos para producir una lista de documentos HTML.

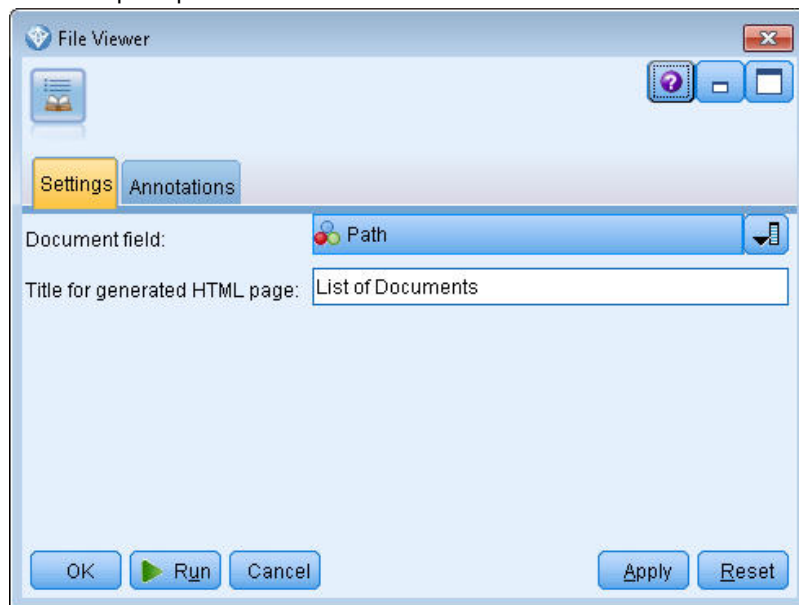


Figura 21. Cuadro de diálogo nodo Visor de archivos: pestaña Configuración

3. **Diálogo salida de visor de archivos.** A continuación, ejecutamos la ruta que saca la lista de documentos en una ventana nueva.

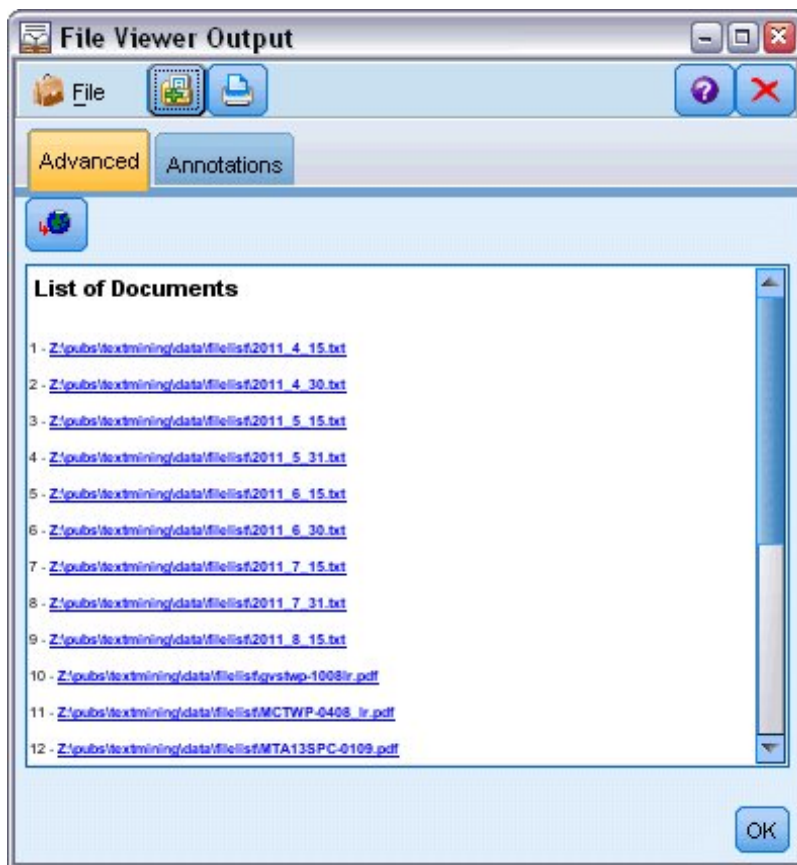


Figura 22. Salida del visor de archivos

4. Para ver los documentos, pulsamos el botón de la barra de herramientas que muestra un globo con una flecha roja. Esto abre una lista de hipervínculos de documento en el navegador.

Capítulo 6. Propiedades de nodo para los scripts

IBM SPSS Modeler tiene un lenguaje de scripts para permitirle ejecutar rutas desde la línea de comandos. Aquí, puede aprender sobre las propiedades del nodo que son específicas a cada uno de los nodos que se entregan con IBM SPSS Modeler Text Analytics. Para obtener más información sobre el conjunto estándar de nodos que se entregan con IBM SPSS Modeler, consulte la Guía de scripts y automatización.

Nodo Lista de archivos: filelistnode

Puede utilizar las propiedades en la siguiente tabla para los scripts. El nodo se denomina `filelistnode`.

Tabla 7. Propiedades de script de nodo Lista de archivos	
Propiedades de los scripts	Tipo de datos
path	serie
recurse	Flag
word_processing	Flag
excel_file	Flag
powerpoint_file	Flag
text_file	Flag
web_page	Flag
xml_file	Flag
pdf_file	Flag
no_extension	Flag

Nota: El parámetro 'Crear lista' ya no está disponible y los scripts que contengan la opción se convertirán automáticamente en una salida de 'Archivos'.

Nodo Canal de información web: webfeednode

Puede utilizar las propiedades en la siguiente tabla para los scripts. El nodo en sí se denomina `webfeednode`.

Tabla 8. Propiedades de scripts del nodo Canal de información web		
Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
urls	serie1 serie2 ...serien	Cada URL se especifica en la estructura de lista. Lista URL separada por “\n”
recent_entries	Flag	
limit_entries	integer	Número de las entradas más recientes a leer por URL.
use_previous	Flag	Para guardar y volver a utilizar caché de Canal de información web.
use_previous_label	serie	Nombre para el caché web guardado.
start_record	serie	Etiqueta de inicio no RSS.

Tabla 8. Propiedades de scripts del nodo Canal de información web (continuación)

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
url <i>n</i> .title	serie	Por cada URL en la lista, debe definir uno aquí también. El primero será url1.title, donde el número coincide con su posición en la lista URL. Esta es la etiqueta de inicio que contiene el título del contenido.
url <i>n</i> .short_description	serie	Lo mismo para url <i>n</i> .title.
url <i>n</i> .description	serie	Lo mismo para url <i>n</i> .title.
url <i>n</i> .authors	serie	Lo mismo para url <i>n</i> .title.
url <i>n</i> .contributors	serie	Lo mismo para url <i>n</i> .title.
url <i>n</i> .published_date	serie	Lo mismo para url <i>n</i> .title.
url <i>n</i> .modified_date	serie	Lo mismo para url <i>n</i> .title.
html_alg	None HTMLCleaner	Método de filtrado de contenido.
discard_lines	Flag	Descartar líneas cortas. Se utiliza con min_words
min_words	integer	Número mínimo de palabras.
discard_words	Flag	Descartar líneas cortas. Se utiliza con min_avg_len
min_avg_len	integer	
discard_scw	Flag	Descartar líneas con muchas palabras de un solo carácter. Se utiliza con max_scw
max_scw	integer	Porcentaje de porción máxima 0-100 de palabras de un solo carácter en una línea
discard_tags	Flag	Descartar líneas que contengan determinadas etiquetas.
tags	serie	Los caracteres especiales deben ir precedidos por un carácter de barra inclinada invertida \.
discard_spec_words	Flag	Descartar líneas que contengan series específicas
words	serie	Los caracteres especiales deben ir precedidos por un carácter de barra inclinada invertida \.

Nodo de idioma: languageidentifier

Puede utilizar las propiedades en la siguiente tabla para los scripts. El nodo se denomina languageidentifier.

Tabla 9. Propiedades de script del nodo de idioma

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
text	campo	
language_field_name	serie	El nombre de campo que se genera como salida.

Tabla 9. Propiedades de script del nodo de idioma (continuación)		
Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
unidentified_language_value	Undefined Supported Custom	El valor predeterminado que se utilizará cuando no pueda identificarse el idioma.
unidentified_language_supported	en de es fr it ja nl pt	Código ISO. Sólo está disponible si unidentified_language_value es Supported.
unidentified_language_custom	serie	Sólo está disponible si unidentified_language_value es Custom.

Nodo minería de textos: TextMiningWorkbench

Puede utilizar los siguientes parámetros para definir o actualizar un nodo mediante scripts. El nodo en sí se denomina TextMiningWorkbench.

Importante: No es posible especificar una plantilla de recurso distinta a través de scripts. Si piensa que necesita una plantilla, debe seleccionarla en el cuadro de diálogo del nodo.

Tabla 10. Propiedades de scripts del nodo de modelado de minería de textos		
Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
text	campo	
method	ReadText ReadPath	
docType	integer	Con valores posibles (0,1,2) donde 0 = Full Text, 1 = Structured Text y 2 = XML

Tabla 10. Propiedades de scripts del nodo de modelado de minería de textos (continuación)

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Tenga en cuenta que los valores que tienen caracteres especiales, como "UTF-8", deben ir entre comillas para no confundirlos con un operador matemático.
unity	<i>integer</i>	Con valores posibles (0,1) donde 0 = Paragraph y 1 = Document
para_min	<i>integer</i>	
para_max	<i>integer</i>	
mtag	<i>serie</i>	Contiene todos los valores mtag (en el cuadro de diálogo Valores para archivos XML)
mclef	<i>serie</i>	Contiene todos los valores mclef (en el cuadro de diálogo Valores para archivos de Texto estructurado)
partition	<i>campo</i>	
custom_field	<i>Flag</i>	Indica si un campo de partición se especificará o no.
use_model_name	<i>Flag</i>	
model_name	<i>serie</i>	
use_partitioned_data	<i>Flag</i>	Si se ha definido un campo de partición, sólo se utilizarán los datos de entrenamiento para la generación del modelo.
model_output_type	Interactive Model	Interactive da como resultado un modelo de categoría. Model da como resultado un modelo de concepto.
use_interactive_info	<i>Flag</i>	Sólo para la compilación interactiva en una sesión de entorno de trabajo.
reuse_extraction_results	<i>Flag</i>	Sólo para la compilación interactiva en una sesión de entorno de trabajo.

Tabla 10. Propiedades de scripts del nodo de modelado de minería de textos (continuación)

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
interactive_view	Categories TLA Clusters	Sólo para la compilación interactiva en una sesión de entorno de trabajo.
extract_top	integer	Este parámetro se utiliza cuando model_type = Concept
use_check_top	Flag	
check_top	integer	
use_uncheck_top	Flag	
uncheck_top	integer	
language	de en es fr it ja nl pt	
frequency_limit	integer	En desuso en 14.0.
concept_count_limit	integer	Limite la extracción a conceptos con una frecuencia global de al menos este valor.
fix_punctuation	Flag	
fix_spelling	Flag	
spelling_limit	integer	
extract_uniterm	Flag	
extract_nonlinguistic	Flag	
upper_case	Flag	
group_names	Flag	
permutation	integer	Permutación máxima de palabra no funcional (el valor predeterminado es 3).

Nugget de modelo de minería de textos: TMWBModelApplier

Puede utilizar las propiedades en la siguiente tabla para los scripts. El nugget mismo se denomina TMWBModelApplier.

Tabla 11. Propiedades del nugget de modelo de minería de textos		
Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
scoring_mode	Fields	
	Records	
field_values	Flags	Esta opción no está disponible en el nugget de modelo Categoría. Para Flags, establezca TRUE o FALSE
	Counts	
true_value	serie	Con Flags, defina el valor para true.
false_value	serie	Con Flags, defina el valor para false.
extension_concept	serie	Especifique una extensión para el nombre de campo. Los nombres de campo se generan al utilizar el nombre de concepto más esta extensión. Especifique donde poner esta extensión utilizando el valor add_as.
extension_category	serie	Extensión de nombre de campo. Puede elegir especificar un prefijo/sufijo de extensión para el nombre de campo o puede elegir utilizar los códigos de categoría. Los nombres de campo se generan al utilizar el nombre de categoría más esta extensión. Especifique donde poner esta extensión utilizando el valor add_as.
add_as	Suffix	
	Prefix	
fix_punctuation	Flag	

Tabla 11. Propiedades del nugget de modelo de minería de textos (continuación)

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
excluded_subcategories_descriptors	RollUpToParent Ignore	Sólo para modelos de categoría. Si una subcategoría no está seleccionada. Esta opción le permite especificar cómo se manejarán los descriptores que pertenecen a subcategorías que no fueron seleccionadas para la puntuación. Existen dos opciones. • Ignore. La opción Excluir sus descriptores completamente de la puntuación hará que los descriptores de subcategorías que no están marcados (no seleccionados) se omitan y no se utilicen durante la puntuación. • RollUpToParent. La opción Agregar descriptores a aquellos de la categoría padre hará que los descriptores de subcategorías que no estén marcados (no seleccionados) se utilicen como descriptores de la categoría padre (la categoría por encima de esta subcategoría). Si varios niveles de subcategorías y no seleccionados, los descriptores se desplegarán en la primera categoría padre disponible
check_model	Flag	En desuso en la versión 14
text	campo	
method	ReadText ReadPath	
docType	integer	Con valores posibles (0,1,2) donde 0 = Full Text, 1 = Structured Text y 2 = XML

Tabla 11. Propiedades del nugget de modelo de minería de textos (continuación)

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Tenga en cuenta que los valores que tienen caracteres especiales, como "UTF-8", deben ir entre comillas para no confundirlos con un operador matemático.
language	de en es fr it ja nl pt	

Nodo Análisis de enlaces de texto: textlinkanalysis

Puede utilizar los parámetros en la siguiente tabla para definir o actualizar un nodo a través de scripts. El nodo en sí se denomina textlinkanalysis.

Importante: No es posible especificar una plantilla de recurso mediante scripts. Para seleccionar una plantilla, debe hacerlo desde adentro del cuadro de diálogo del nodo.

Tabla 12. Propiedades de script de nodo de Análisis de enlaces de texto (TLA)

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
id_field	campo	
text	campo	

Tabla 12. Propiedades de script de nodo de Análisis de enlaces de texto (TLA) (continuación)

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
method	ReadText	
	ReadPath	
docType	integer	Con valores posibles (0,1,2) donde 0 = Full Text, 1 = Structured Text y 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Tenga en cuenta que los valores que tienen caracteres especiales, como "UTF-8", deben ir entre comillas para no confundirlos con un operador matemático.
unity	integer	Con valores posibles (0,1) donde 0 = Paragraph y 1 = Document
para_min	integer	
para_max	integer	
mtag	serie	Contiene todos los valores mtag (en el cuadro de diálogo Valores para archivos XML)
mclef	serie	Contiene todos los valores mclef (en el cuadro de diálogo Valores para archivos de Texto estructurado)

Tabla 12. Propiedades de script de nodo de Análisis de enlaces de texto (TLA) (continuación)

Propiedades de los scripts	Tipo de datos	Descripción de la propiedad
language	de en es fr it ja nl pt	
concept_count_limit	<i>integer</i>	Limite la extracción a conceptos con una frecuencia global de al menos este valor.
fix_punctuation	<i>Flag</i>	
fix_spelling	<i>Flag</i>	
spelling_limit	<i>integer</i>	
extract_uniterm	<i>Flag</i>	
extract_nonlinguistic	<i>Flag</i>	
upper_case	<i>Flag</i>	
group_names	<i>Flag</i>	
permutation	<i>integer</i>	Permutación máxima de palabra no funcional (el valor predeterminado es 3).

Capítulo 7. Modalidad de área de trabajo interactiva

Desde un nodo de modelado de minería de texto, puede optar por iniciar una sesión de área de trabajo interactiva durante la ejecución de la ruta. En este entorno de trabajo, puede extraer conceptos clave de los datos de texto, puede construir categorías, y explorar patrones de análisis de enlace de texto y clústeres y generar modelos de categoría. En este capítulo, discutimos la interfaz del entorno de trabajo desde una perspectiva de alto nivel junto con los elementos principales con los que trabaja, incluyendo:

- **Resultados de la extracción.** Después de que se realiza una extracción, estas son las palabras claves y frases identificadas y extraídas de los datos de texto, a las que también se conoce como *conceptos*. Estos conceptos se agrupan en *tipos*. Al utilizar estos tipos y conceptos, puede explorar los datos así como también crear sus categorías. Estos se gestionan en la vista **Categorías y Conceptos**.
- **Categorías.** Utilización de descriptores (como resultados de extracción, patrones, y reglas) como una definición, puede crear de manera manual o automática un conjunto de categorías a las que documentos y registros se asignan basándose en si contienen o no una parte de la definición de la categoría. Estos se gestionan en la vista **Categorías y Conceptos**.
- **Clústeres.** *Clústeres* son un grupo de conceptos entre los que los enlaces se han descubierto que indican una relación entre ellos. Los conceptos se agrupan utilizando un algoritmo complejo que utiliza, entre otros factores, con qué frecuencia dos conceptos aparecen juntos en comparación con qué frecuencia aparecen por separado. Éstos se gestionan en la vista **Clústeres**. También puede añadir los conceptos que componen un clúster en categorías.
- **Patrones de análisis de enlaces de texto.** Si tiene reglas de patrones de análisis de enlace de texto (TLA) en sus recursos lingüísticos o está utilizando una plantilla de recursos que ya tiene algunas reglas TLA, puede extraer patrones de los datos de texto. Estos patrones le pueden ayudar a descubrir relaciones interesantes entre conceptos en sus datos. También puede utilizar estos patrones como descriptores en las categorías. Éstos se gestionan en la vista **Análisis de enlace de texto**.
- **Recursos lingüísticos.** El proceso de extracción se basa en un conjunto de parámetros y definiciones lingüísticos para controlar cómo se extrae el texto y cómo se maneja. Estos son gestionados en forma de plantillas y bibliotecas en la vista **Editor de recursos**.

Posibles problemas de área de trabajo interactiva

- Varias sesiones de área de trabajo interactivas pueden provocar un comportamiento muy lento. SPSS Modeler Text Analytics y SPSS Modeler comparten un motor de tiempo de ejecución Java cuando se inicia una sesión de área de trabajo. En función del número de sesiones de área de trabajo interactivas que invoque durante una sesión de SPSS Modeler, es posible que la memoria del sistema provoque que la aplicación funcione muy lentamente, aunque abra y cierre la misma sesión. Este efecto puede ser especialmente apreciable si está trabajando con una gran cantidad de datos o tiene una máquina que tenga menos RAM que el valor recomendado de 4 GB. Si observa que la máquina responde muy lentamente, es recomendable que guarde todos sus trabajos, cierre SPSS Modeler, y vuelva a iniciar la aplicación. Ejecutar SPSS Modeler Text Analytics en una máquina con menos memoria de la recomendada, especialmente cuando se trabaje con grandes conjuntos de datos o durante prolongados períodos de tiempo, puede provocar que Java quede sin memoria, y se cierre. Se recomienda encarecidamente actualizar al valor al recomendado, o más, de memoria (o utilice el servidor de SPSS Modeler Text Analytics) si trabaja con cantidades de datos grandes.
- SPSS Modeler El cliente puede quedarse sin memoria después de que se ejecuten varias sesiones de trabajo interactivas de SPSS Modeler Text Analytics sin reiniciar la aplicación. Supervise el uso de la memoria en la línea de estado y, si se agota ésta, cierre y vuelva a abrir el cliente de SPSS Modeler.

La vista de Categorías y Conceptos

La interfaz de aplicación se compone de varias vistas. La vista Categorías y Conceptos es la ventana en la que puede crear y explorar las categorías así como explorar y modificar los resultados de la extracción.

Categorías hace referencia a un grupo de ideas y patrones estrechamente relacionados, a los que se asignan documentos y registros a través de un proceso de puntuación. Mientras que *conceptos* se refiere al nivel más básico de resultados de extracción disponibles para utilizar como bloques de creación, denominados descriptores, para sus categorías.

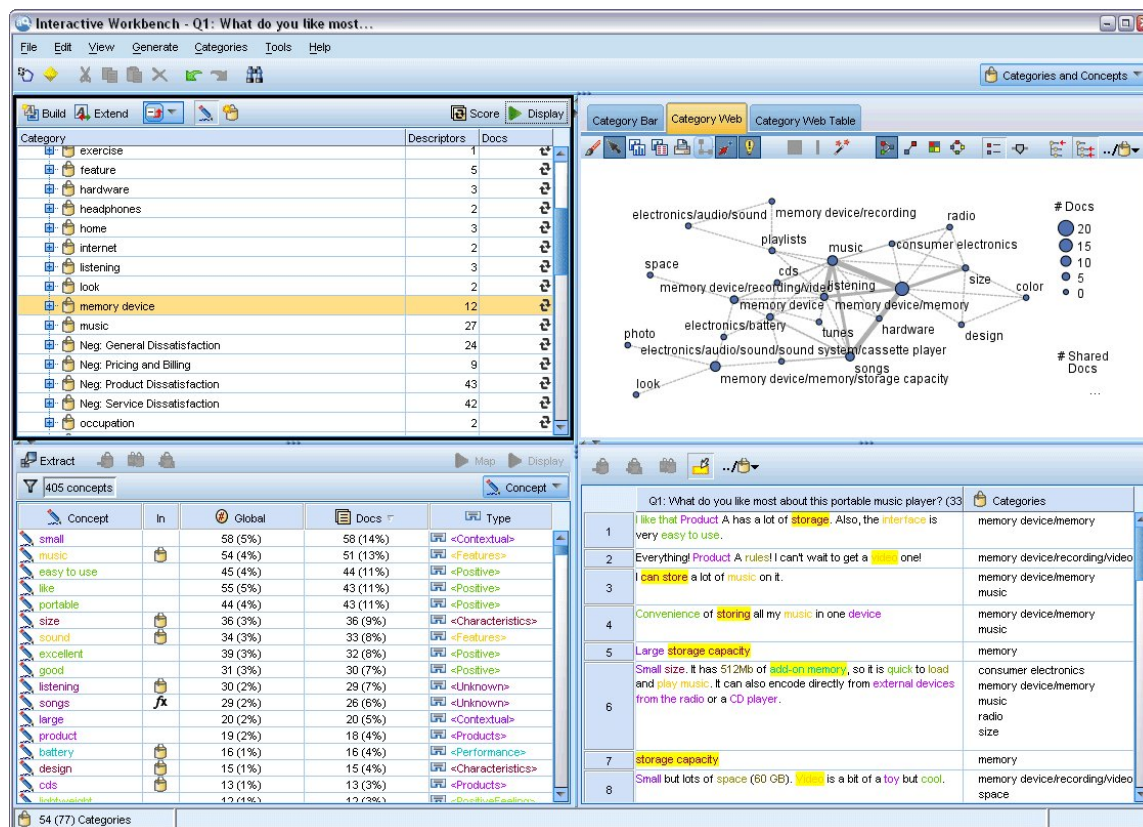


Figura 23. Vistas Conceptos y Categorías

La vista Categorías y Conceptos está organizada en cuatro paneles, cada uno de los cuales se pueden ocultar o mostrar seleccionando su nombre en el menú Ver. Para obtener más información, consulte el tema [Capítulo 9, “Categorización de datos de texto”](#), en la página 95.

Panel Categorías

Situada en la esquina superior izquierda, esta área presenta una tabla en la que puede gestionar las categorías que vaya a crear. Después de extraer los conceptos y los tipos de los datos de texto, puede empezar a crear categorías utilizando técnicas como redes semánticas y la inclusión de conceptos, o puede crearlas manualmente. Si pulsa dos veces sobre el nombre de una categoría, se abre el cuadro de diálogo Definiciones de categoría que muestra todos los descriptores que forman su definición, como conceptos, tipos, y reglas. Para obtener más información, consulte el tema [Capítulo 9, “Categorización de datos de texto”](#), en la página 95. No todas las técnicas automáticas están disponibles para todos los idiomas.

Cuando selecciona una fila en el panel, puede visualizar información sobre documentos/registros correspondientes o descriptores en los paneles de datos y visualización.

Panel de Resultados de extracción

Situada en el ángulo inferior izquierdo, esta área presenta los resultados de la extracción. Cuando ejecuta una extracción, el motor de extracción lee los datos del texto, identifica los conceptos relevantes y asigna un tipo a cada uno. *Conceptos* son palabras o frases extraídas de los datos de texto. *Tipos* son agrupaciones semánticas de conceptos almacenados en forma de diccionarios tipo. Cuando la extracción se completa, conceptos y tipos aparecen con la codificación de color en el panel Resultados extraídos.

Para obtener más información, consulte el tema [“Resultados de extracción: Conceptos y tipos”](#) en la [página 81](#).

Puede ver el conjunto de términos subyacentes de un concepto pasando el ratón por encima del nombre del concepto. Al hacerlo aparecerá una etiqueta con información que muestra el nombre del concepto y varias líneas de términos agrupados bajo dicho concepto. Estos términos subyacentes incluyen los sinónimos definidos en los recursos lingüísticos (independientemente de si se encontraron en el texto o no), así como los términos en plural/singular extraídos, términos permutados, términos de agrupación difusa, etc. Puede copiar estos términos o ver el conjunto completo de términos subyacentes pulsando con el botón derecho del ratón en el nombre del concepto y seleccionando la opción del menú contextual.

Minería de textos es un proceso repetitivo en el que los resultados de la extracción se revisan de acuerdo con el contexto de los datos de texto, se ajustan para generar resultados nuevos y luego se reevalúan. Los resultados de la extracción pueden refinarse modificando los recursos lingüísticos. Este ajuste puede hacerse en parte directamente desde el panel Resultados de extracción o desde el panel de datos, o sino también directamente en la vista Editor de recursos. Para obtener más información, consulte el tema [“La vista Editor de recursos”](#) en la [página 73](#).

Nota: Si hay más resultados que puedan caber en el panel visible, puede utilizar los controles de la parte inferior del panel para desplazarse hacia delante y hacia atrás por los resultados, o escriba un número de página para ir a la misma.

Panel Visualización

Situada en la esquina superior derecha, esta área presenta varias perspectivas sobre las similitudes en el documento/registro de categorización. Cada gráfico proporciona información similar pero la presenta de forma diferente o con un nivel diferente de detalles. Estos gráficos pueden utilizarse para analizar los resultados de la categorización y para ayudar a ajustar las categorías o los informes. Por ejemplo, en un gráfico puede descubrir las categorías que son demasiado diferentes o demasiado similares (por ejemplo, comparten más del 75% de sus registros). El contenido en un gráfico o diagrama corresponden a la selección en los otros paneles. Para obtener más información, consulte el tema [“Gráficos de categoría”](#) en la [página 159](#).

Panel Datos

El panel Datos está situado en la esquina inferior derecha. Este panel presenta una tabla que contiene los documentos o registros correspondientes a una selección en otra área de la vista. Según lo que se ha seleccionado, solo el texto correspondiente aparece en el panel Datos. Una vez que ha realizado una selección, pulse el botón **Mostrar** para llenar el panel Datos con el texto correspondiente.

Si tiene una selección en otro panel, los documentos o registros correspondientes muestran los conceptos resaltados en color para ayudarle a identificarlos fácilmente en el texto. También puede pasar el ratón sobre los elementos con codificación por color para que aparezca una ayuda contextual que muestre el nombre del concepto bajo el que ha sido extraído y el tipo al que fue asignado. Para obtener más información, consulte el tema [“El panel Datos”](#) en la [página 103](#).

Buscar y encontrar en la vista Categorías y Conceptos

En algunos casos, puede que necesite localizar información rápidamente en una sección determinada. Al utilizar barra de herramientas Buscar, puede especificar la serie que desea buscar y definir otros criterios de búsqueda otros como la distinción entre mayúsculas y minúsculas o la dirección de búsqueda. Después puede elegir el panel en el que desea buscar.

Utilizar la característica Buscar

1. En la vista Categorías y conceptos, seleccione **Editar > Buscar** en los menús. La barra de herramientas Buscar aparecerá encima del panel Categorías y paneles de visualización.
2. Escriba la cadena de palabras que desea buscar en el cuadro de texto. Puede utilizar los botones de la barra de herramientas para controlar las mayúsculas/minúsculas, la coincidencia parcial y la dirección de la búsqueda.

3. En la barra de herramientas, pulse en el nombre del panel en el que desea buscar. Si se encuentra una coincidencia, el texto se resalta en la ventana.
4. Para buscar la coincidencia siguiente, pulse en el nombre del panel de nuevo.

La vista de clústeres

En la vista de clústeres puede crear y explorar los resultados de clúster que se han encontrado en los datos de texto. Los *Clústeres* son agrupaciones de conceptos generados por algoritmos de agrupación en clúster basados en la frecuencia con que se producen los conceptos y en la frecuencia con que aparecen juntos. El objetivo de los clústeres es agrupar conceptos que se producen juntos, mientras que el objetivo de las categorías es agrupar documentos o registros basándose en la forma en que el texto que contienen coincide con los descriptores (conceptos, reglas, patrones) para cada categoría.

Cuanto más a menudo los conceptos de un clúster se produzcan juntos en conjunto con cuanto menos frecuentemente se producen con otros conceptos, mejor será el clúster para identificar relaciones interesantes de conceptos. Dos conceptos se producen juntos cuando ambos aparecen (o uno de sus sinónimos o términos aparecen) en el mismo documento o registro. Para obtener más información, consulte el tema [Capítulo 10, “Análisis de clústeres”](#), en la [página 143](#).

Puede crear clústeres y explorarlos en un conjunto de diagramas y gráficos que podrían ayudarle a descubrir relaciones entre conceptos que de lo contrario llevaría demasiado tiempo encontrar. Aunque no puede añadir clústeres completos a las categorías, puede añadir los conceptos de un clúster a una categoría mediante el cuadro de diálogo Definición de clústeres. Para obtener más información, consulte el tema [“Definiciones de clúster”](#) en la [página 147](#).

Puede realizar cambios a los valores de la agrupación en clúster para influenciar los resultados. Para obtener más información, consulte el tema [“Creando clústeres”](#) en la [página 144](#).

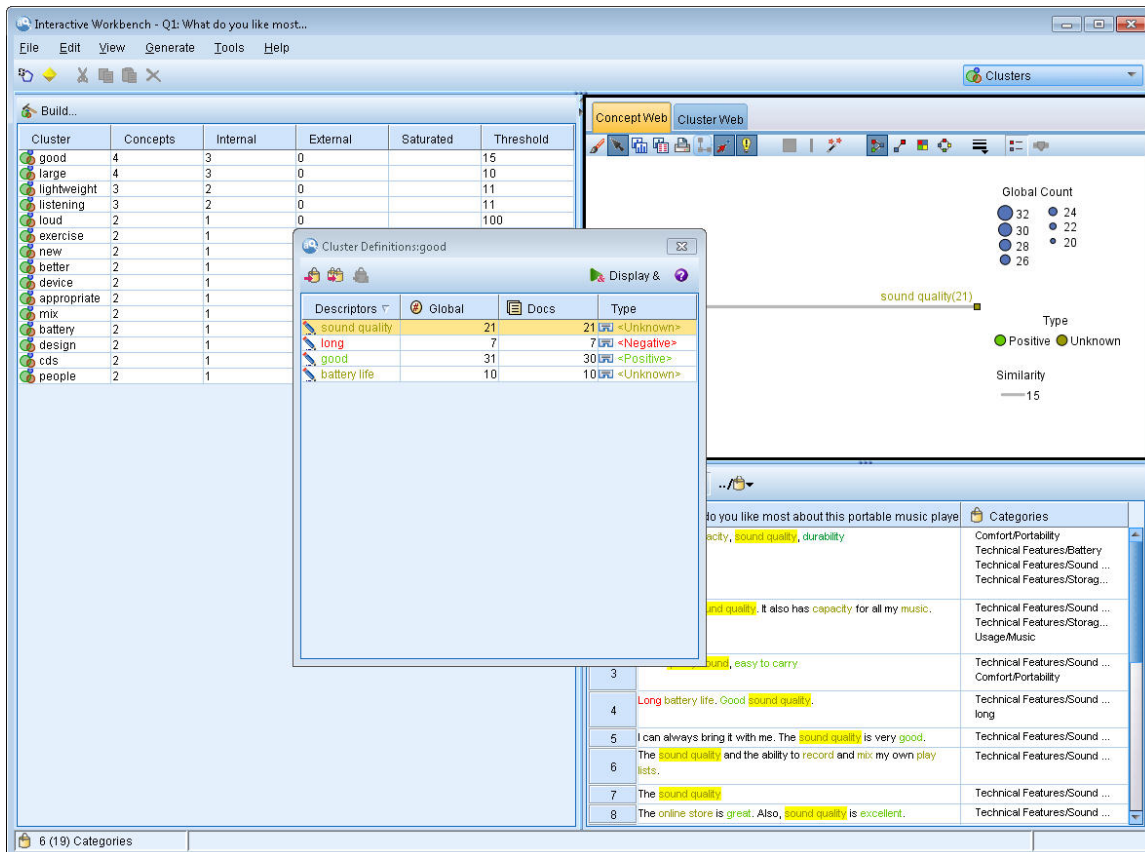


Figura 24. Vista Clústeres

La vista Clústeres está organizada en tres paneles, cada uno de los cuales pueden ocultarse o mostrarse seleccionando su nombre en el menú Ver. Normalmente solo están visibles los paneles Clústeres y Visualización.

Panel Clústeres

Situado en el lado izquierdo, este panel presenta los clústeres que fueron descubiertos en los datos de texto. Puede crear resultados de agrupaciones de clúster pulsando el botón **Crear**. Los clústeres están formados por un algoritmo de agrupación de clúster que intenta identificar conceptos que se producen juntos frecuentemente.

Siempre que tiene lugar una nueva extracción, los resultados de clúster se borran y debe volver a crear los clústeres para obtener los últimos resultados. Al crear las agrupaciones, puede cambiar algunos valores, como el número máximo de clústeres a crear, el número máximo de conceptos que puede contener, o el número máximo de enlaces con conceptos externos que puede tener. Para obtener más información, consulte el tema [“Exploración de los clústeres”](#) en la página 146.

Panel Visualización

Situado en la esquina superior derecha, este panel ofrece dos perspectivas de agrupación de clúster: un gráfico web de concepto y un gráfico web de clúster. Si no es visible, puede acceder a este panel desde el menú Ver (**Vista > Visualización**). Según lo que se seleccione en el panel de clústeres, puede visualizar las interacciones correspondientes entre clústeres o dentro de ellos. Los resultados se presentan en múltiples formatos:

- **Web de concepto.** Gráfico web que muestra todos los conceptos dentro de los clústeres seleccionados, así como los conceptos enlazados fuera del clúster.
- **Web de clúster.** Gráfico web que muestra los enlaces desde los clústeres seleccionados a otros clústeres, así como cualquier enlace entre estos u otros clústeres.

Nota: Para mostrar un gráfico clúster web, debe tener con enlaces externos ya construidos. Los enlaces externos son enlaces entre pares de conceptos de clústeres separados (un concepto dentro de un clúster y un concepto fuera, en otro clúster). Para obtener más información, consulte el tema [“Gráficos de clúster”](#) en la página 161.

Panel Datos

El panel Datos se encuentra en la esquina inferior derecha y está oculto de forma predeterminada. No puede visualizarse ningún resultado del panel de Datos desde el panel Clústeres ya que estos clústeres abarcan varios documentos/registros, haciendo que los resultados sean poco interesantes. No obstante, puede ver los datos correspondientes a una selección en el cuadro de diálogo Definiciones de clúster. Según lo que se ha seleccionado en el cuadro de diálogo, solo el texto correspondiente aparece en el panel Datos. Una vez que haya realizado una selección, pulse el botón **Visualizar** para rellenar el panel Datos con los documentos o registros que contienen todos los conceptos juntos.

Los documentos o registros correspondientes muestran los conceptos resaltados en color para ayudarle a identificarlos fácilmente en el texto. También puede pasar el ratón por encima de los elementos con codificación por color para mostrar el concepto bajo el que se extrajo y el tipo al que se asignó. El panel Datos puede contener varias columnas pero la columna de campo de texto siempre se muestra. Lleva el nombre del campo de texto que se utilizó durante la extracción o un nombre de documento si los datos de texto están en varios archivos diferentes. Hay otras columnas disponibles. Para obtener más información, consulte el tema [“El panel Datos”](#) en la página 103.

La vista Análisis de enlaces de texto

En la vista Análisis de enlace de texto, puede crear y explorar patrones de análisis de enlace de texto encontrados en los datos de texto. El análisis de enlace de texto (TLA) es una tecnología de coincidencia de patrón que le permite definir reglas de TLA y compararlas con conceptos y relaciones extraídos reales encontrados en el texto.

Los patrones son muy útiles cuando se intenta detectar relaciones entre conceptos u opiniones sobre un asunto en particular. Algunos ejemplos incluyen querer extraer opiniones sobre productos a partir de datos de encuesta, relaciones genómicas desde dentro de los documentos de la investigación médica, o relaciones entre personas o lugares a partir de datos de inteligencia.

Una vez que haya extraído algunos patrones TLA, puede explorarlos en los paneles Datos o Visualización e incluso añadirlos a categorías en la vista Categorías y conceptos. Deben haber algunas reglas TLA definidas en la plantilla o biblioteca de recursos que está utilizando para poder extraer resultados TLA. Para obtener más información, consulte el tema [Capítulo 18, “Sobre las reglas de enlaces de texto”, en la página 211.](#)

Si ha elegido extraer resultados de patrones TLA, los resultados se presentan en esta vista. Si no ha optado por hacerlo, deberá utilizar el botón **Extraer** y elegir la opción de habilitar la extracción de patrones.

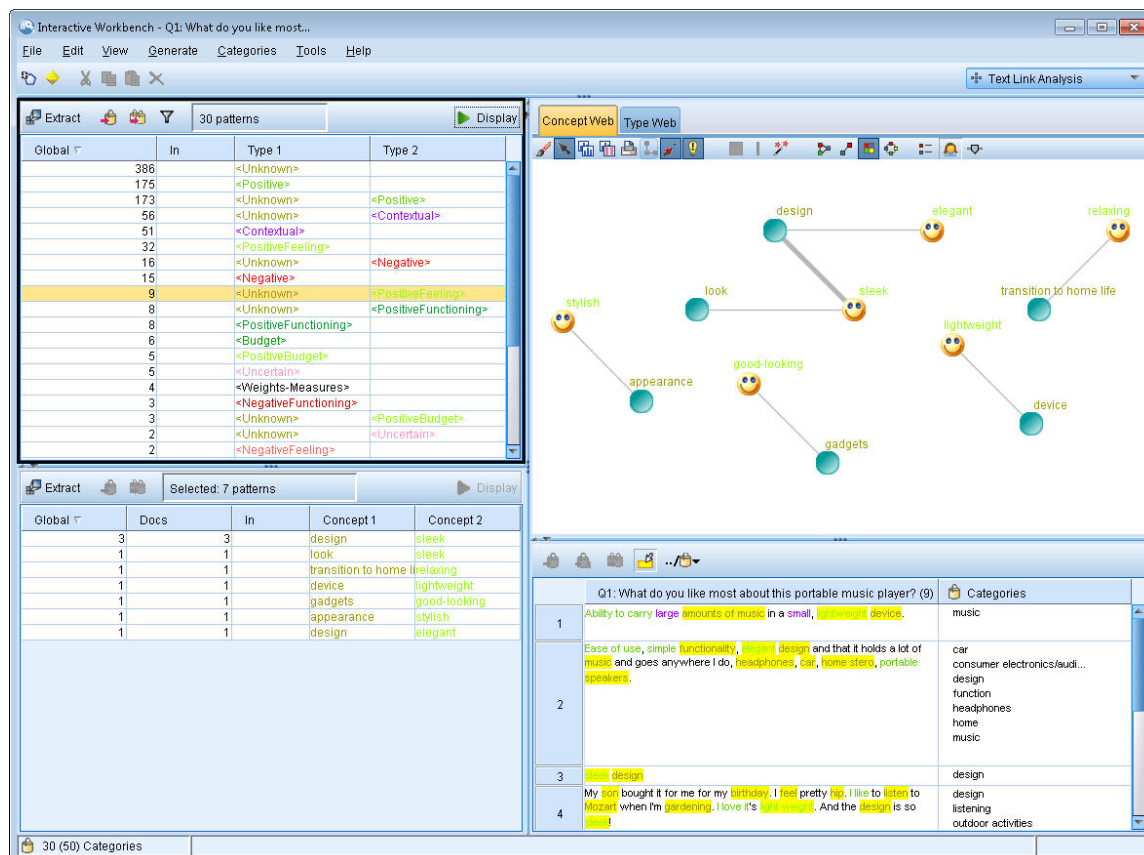


Figura 25. Vista Análisis para los enlaces de texto

La vista Análisis para los enlaces de texto está organizada en cuatro paneles, cada uno de los cuales puede ocultarse o mostrarse seleccionando su nombre en el menú Ver. Para obtener más información, consulte el tema [Capítulo 11, “Exploración del Análisis de enlaces de texto”, en la página 149.](#)

Paneles Patrón de tipo y concepto

Situados en el lado izquierdo, los paneles Patrón de tipo y concepto son dos paneles interconectados en los que puede explorar y seleccionar los resultados de patrones TLA. Los patrones se componen de una serie de hasta seis tipos o seis conceptos. La regla de patrón TLA tal como se definen en los recursos lingüísticos dicta la complejidad de los resultados de patrón. Para obtener más información, consulte el tema [Capítulo 18, “Sobre las reglas de enlaces de texto”, en la página 211.](#)

Los resultados de patrones se agrupan primero a nivel de tipo y a continuación se dividen en patrones de concepto. Por esta razón, hay dos paneles de resultados diferentes: Patrones tipo (superior izquierdo) y Patrones concepto (inferior izquierdo).

- **Patrones tipo.** El panel Patrones de tipo presenta patrones extraídos que constan de dos o más tipos relacionados que coinciden con una regla de patrón TLA. Los patrones tipo se muestran como `<Organization> + <Location> + <Positive>`, que podrían proporcionar comentarios positivos sobre una organización en una ubicación específica.
- **Patrones de concepto.** El panel de Patrones de concepto presenta los patrones extraídos a nivel de concepto para todos los patrones de tipo actualmente seleccionados en el panel Patrones de tipo que está por encima de él. Los patrones de concepto siguen una estructura como `hotel + paris + wonderful`.

Al igual que con los resultados de extracción en la vista Categorías y Conceptos, puede revisar los resultados aquí. Si ve cualquier perfeccionamiento que desee hacer a los tipos y conceptos que componen estos patrones, puede realizarlos en el panel Resultados de extracciones en la vista Categorías y conceptos, o directamente en el Editor de recursos, y volver a extraer los patrones.

Panel Visualización

Situado en la esquina superior derecha de la vista Análisis de enlace de texto, este panel presenta un gráfico web de los patrones seleccionados como patrones de tipo o de concepto. Si no está visible, puede acceder a este panel desde el menú Ver (**Ver > Visualización**). En función de lo que esté seleccionado en los otros paneles, puede ver las interacciones correspondientes entre documentos/registros y los patrones.

Los resultados se presentan en múltiples formatos:

- **Gráfico de concepto.** Este gráfico presenta todos los conceptos de los patrones seleccionados. La anchura de línea y los tamaños de nodo (si no se muestran los iconos de tipo) de un gráfico de concepto muestran el número de apariciones globales en la tabla seleccionada.
- **Gráfico de tipo.** Este gráfico presenta todos los tipos de los patrones seleccionados. La anchura de línea y los tamaños de nodo (si no se muestran los iconos de tipo) del gráfico muestran el número de apariciones globales en la tabla seleccionada. Los nodos se representan ya sea mediante un tipo de color o mediante un icono.

Para obtener más información, consulte el tema [“Gráficos de Análisis de enlace de texto” en la página 162.](#)

Panel Datos

El panel Datos está situado en la esquina inferior derecha. Este panel presenta una tabla que contiene los documentos o registros correspondientes a una selección en otra área de la vista. Según lo que se ha seleccionado, solo el texto correspondiente aparece en el panel Datos. Una vez que ha realizado una selección, pulse el botón **Mostrar** para llenar el panel Datos con el texto correspondiente.

Si tiene una selección en otro panel, los documentos o registros correspondientes muestran los conceptos resaltados en color para ayudarle a identificarlos fácilmente en el texto. También puede pasar el ratón sobre los elementos con codificación por color para que aparezca una ayuda contextual que muestre el nombre del concepto bajo el que ha sido extraído y el tipo al que fue asignado. Para obtener más información, consulte el tema [“El panel Datos” en la página 103.](#)

La vista Editor de recursos

IBM SPSS Modeler Text Analytics captura rápida y precisamente conceptos clave de datos de texto utilizando un motor de extracción sólido. Este motor se basa principalmente en los recursos lingüísticos para dictaminar qué cantidad de datos textuales y sin estructurar deben analizarse e interpretarse.

La vista de Editor de recursos es donde puede ver y afinar los recursos lingüísticos utilizados para extraer conceptos, agruparlos en tipos, descubrir patrones en los datos de texto, y mucho más IBM SPSS Modeler Text Analytics ofrece varias plantillas de recursos preconfiguradas. Además, en algunos idiomas puede utilizar los recursos en paquetes de análisis de texto. Para obtener más información, consulte el tema [“Uso de los paquetes de análisis de texto” en la página 135.](#)

Puesto que estos recursos no siempre se adaptan perfectamente al contexto de sus datos, puede crear, editar y administrar sus propios recursos para un contexto o dominio determinados en el Editor de recursos. Para obtener más información, consulte el tema [Capítulo 15, “Trabajo con bibliotecas”](#), en la página 179.

Para simplificar el proceso de afinado de sus recursos lingüísticos, puede realizar tareas de diccionario comunes directamente desde la vista Categorías y conceptos a través de menús contextuales en los paneles Resultados de extracción y Datos. Para obtener más información, consulte el tema [“Refinar resultados de extracción”](#) en la página 89.

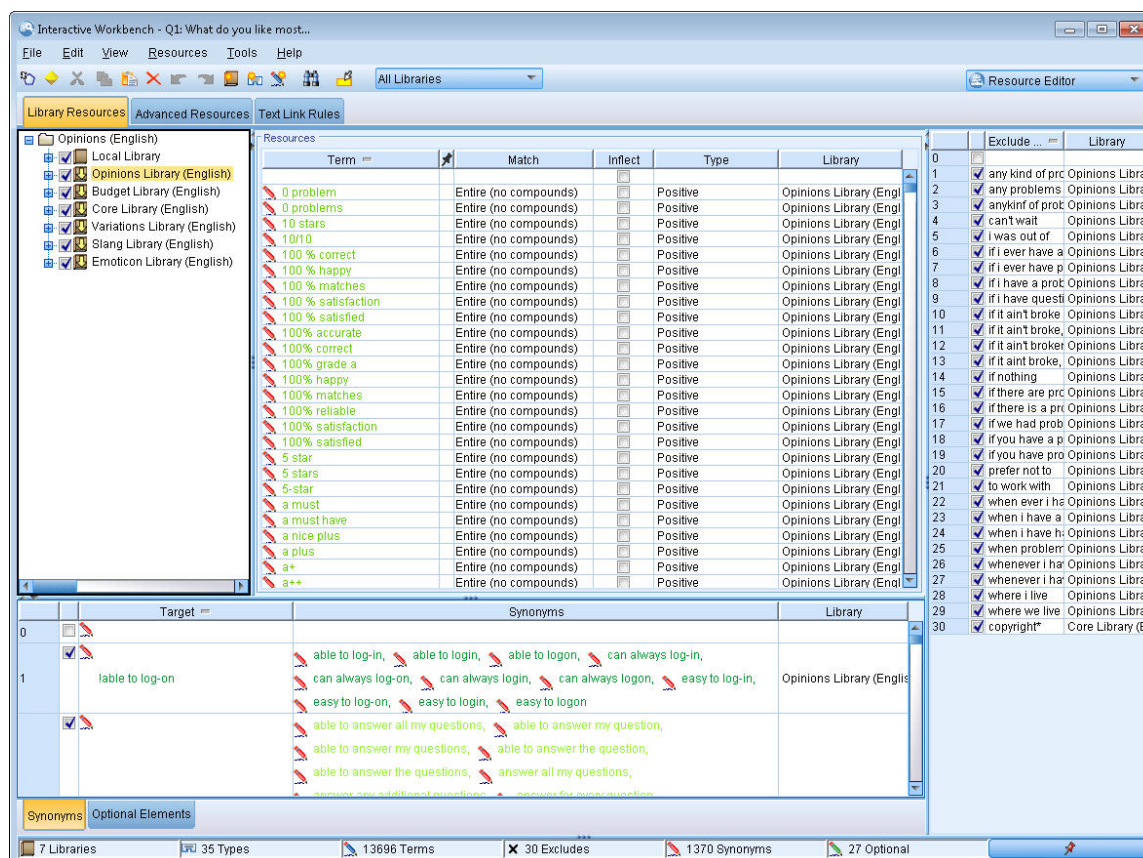


Figura 26. Vista del editor de recursos

Las operaciones que realiza en el Editor de recursos se concentran en torno a la administración y ajuste de los recursos lingüísticos. Estos recursos se almacenan en forma de plantillas y bibliotecas. La vista Editor de recursos se organiza en cuatro partes: panel Árbol de bibliotecas, panel Diccionario de tipos, panel Diccionario de sustituciones y panel Diccionario de exclusiones.

Nota: Para obtener más información, consulte el tema [“La interfaz Editor”](#) en la página 170.

Opciones de configuración

Puede definir opciones generales para IBM SPSS Modeler Text Analytics en el cuadro de diálogo Opciones. Este cuadro de diálogo contiene las siguientes pestañas:

- **Sesión.** Esta pestaña contiene opciones generales y delimitadores.
- **Mostrar.** Esta pestaña contiene opciones para los colores que se utilizan en la interfaz.
- **Sonidos.** Esta pestaña contiene opciones para las pistas de sonido.

Para editar opciones

1. En los menús, seleccione **Herramientas > Opciones**. Se abre el cuadro de diálogo Opciones.
2. Seleccione la pestaña que contiene la información que desee cambiar.

3. Cambie las opciones pertinentes.
4. Pulse **Aceptar** para guardar los cambios.

Opciones: separador Sesión

En este separador, puede definir algunos de los valores básicos.

Visualización de gráfico de categoría y panel de datos. Estas opciones afectan al modo en que los datos se presentan en el panel Datos y en el panel Visualización en la vista Categorías y conceptos.

- **Visualizar el límite para el panel Datos y la Web de categorías** Esta opción establece el número máximo de documentos a mostrar o utilizar para llenar los paneles o gráficos de Datos y los diagramas en la vista Categorías y conceptos.
- **Mostrar las categorías para los documentos/registros en el momento de la visualización.** Si se selecciona, los documentos o registros se anotan cada vez que pulsa Visualizar para que las categorías a las que pertenecen puedan visualizarse en la columna Categorías en el panel Datos así como en los gráficos de categoría. En algunos casos, especialmente con conjuntos de datos más grandes, es posible que desee desactivar esta opción para que los datos y gráficos se visualicen mucho más rápido.

Añadir a Categoría desde el panel Datos. Estas opciones afectan a lo que se añade a las categorías cuando los documentos y registros se añaden desde el panel Datos.

- **En la vista Categorías y conceptos, copia.** Añadir un documento o un registro desde el panel Datos en esta vista copiará en **Solo conceptos** o en ambos **Conceptos y patrones**.
- **En la vista Análisis de enlace de texto, copia.** Añadir un documento o un registro desde el panel Datos en esta vista copiará en **Solo patrones** o en ambos **Conceptos y patrones**.

Delimitador del Editor de recursos. Seleccione el carácter que se va utilizar como delimitador al especificar elementos, como por ejemplo conceptos, sinónimos y elementos opcionales, en la vista Editor de recursos.

Opciones: Visualizar pestaña

En esta pestaña, puede editar las opciones que inciden en el aspecto y el funcionamiento general de la aplicación, y los colores que se utilizan para distinguir los elementos.

Nota: Para cambiar el aspecto y la sensación del producto a un aspecto clásico o a uno de un release anterior, abra el diálogo Opciones de usuario en el menú Herramientas en la ventana principal IBM SPSS Modeler.

Colores personalizados. Edite los colores de los elementos que aparecen en pantalla. Para cada uno de los elementos de la tabla, puede cambiar el color. Para especificar un color personalizado, pulse en el área de color a la derecha del elemento que desea cambiar y elija un color en la lista desplegable de colores.

- **Texto no extraído.** Datos de texto que no se han extraído aún visible en el panel Datos.
- **Fondo resaltado.** Color de fondo del texto seleccionado al seleccionar elementos en los paneles o texto en el panel Datos.
- **Fondo de extracción necesaria.** El color de fondo de los resultados de extracción, patrones y paneles de clústeres que indica que los cambios se han realizado en las bibliotecas y una extracción es necesaria.
- **Fondo de comentarios de Categoría.** Color de fondo de la categoría que aparece después de una operación.
- **Tipo predeterminado.** Color predeterminado para tipos y conceptos que aparecen en el panel Datos y en el panel Resultados de extracción. Este color se aplicará a cualquier tipo personalizado que vaya a crear en el Editor de recursos. Puede sustituir este color predeterminado para sus diccionarios tipo personalizados mediante la edición de las propiedades para estos diccionarios tipo en Editor de recursos. Para obtener más información, consulte el tema [“Crear tipos” en la página 189](#).

- **Tabla a bandas 1.** Primero de los dos colores utilizados de manera alterna en la tabla en el cuadro de diálogo Editar conceptos forzados para diferenciar cada conjunto de líneas.
- **Tabla a bandas 2.** Segundo de los dos colores utilizados de forma alterna en la tabla en el cuadro de diálogo Editar conceptos forzados para diferenciar cada conjunto de líneas.

Nota: Si pulsa el botón **Restablecer a predeterminado**, todas las opciones en este cuadro de diálogo se restablecen a los valores que tenían cuando instaló este producto por primera vez.

Opciones: Pestaña Sonidos

En esta pestaña, puede editar las opciones que inciden en los sonidos. En Eventos de sonido puede especificar que se utilice un sonido para notificar que ha ocurrido un evento. Hay numerosos sonidos disponibles. Utilice el botón de puntos suspensivos (...) para examinar y seleccione un sonido. Los archivos .wav que se utilizan para crear sonidos para IBM SPSS Modeler Text Analytics se almacenan en el subdirectorio *media* del directorio de instalación. Si no desea que se reproduzca ningún sonido, seleccione **Silenciar todos los sonidos**. Los sonidos están silenciados de forma predeterminada.

Nota: Si pulsa el botón **Restablecer a predeterminado**, todas las opciones en este cuadro de diálogo se restablecen a los valores que tenían cuando instaló este producto por primera vez.

Configuración de Microsoft Internet Explorer para obtener ayuda

Configuración de Microsoft Internet Explorer

La mayoría de las características de ayuda de esta aplicación utilizan tecnología basada en Microsoft Internet Explorer. Algunas versiones de Internet Explorer (incluida la versión que se incluyen con Microsoft Windows XP, Service Pack 2) bloquearán de forma predeterminada lo que se consideren "contenidos activos" en las ventanas de Internet Explorer de su ordenador local. Esta configuración predeterminada puede hacer que se bloqueen algunos contenidos de las características de ayuda. Para ver todos los contenidos de ayuda, puede cambiar el comportamiento predeterminado de Internet Explorer.

1. Elija en los menús de Internet Explorer, seleccione:
Herramientas > Opciones de Internet...
2. Pulse en la pestaña **Opciones avanzadas**.
3. Desplácese hacia abajo hasta la sección **Seguridad**.
4. Seleccione **Permitir que el contenido activo se ejecute en archivos de Mi PC**.

Generación de los nuggets de modelo y los nodos de modelado

Cuando está en una sesión interactiva, puede que desee utilizar el trabajo que ha realizado para generar ya sea:

- **Un nodo modelado de minería de textos.** Un nodo modelado generado a partir de una sesión de área de trabajo interactiva es un nodo de Minería de textos cuyos valores y opciones reflejan aquellos almacenados en la sesión interactiva abierta. Esto puede ser útil cuando ya no tiene el nodo minería de textos original o cuando desea crear una nueva versión. Para obtener más información, consulte el tema [Capítulo 3, "Minería para conceptos y categorías", en la página 19](#).
- **Un nugget de modelo de categoría.** Un nugget de modelo generado a partir de una sesión de área de trabajo interactiva es un nugget de modelo de categoría. Debe tener al menos una categoría en la vista Categorías y Conceptos para generar un nugget de modelo de categoría. Para obtener más información, consulte el tema ["Nugget de minería de textos: Modelo de categoría" en la página 38](#).

Para generar un nodo modelado de minería de textos

1. En los menús, seleccione **Generar > Generar nodo de modelado**. Un nodo modelado de minería de textos se añade al lienzo de trabajo utilizando todos los valores que se encuentran en la sesión de entorno de trabajo. El nodo se denomina después del campo de texto.

Para generar un nugget de modelo de categoría

1. En los menús, seleccione **Generar > Generar modelo**. Un nugget de modelo se genera directamente en la paleta de modelos con el nombre predeterminado.

Guardar y actualizar nodos de modelado

Mientras está trabajando en una sesión interactiva, es recomendable que actualice el nodo de modelado de vez en cuando para guardar los cambios. También debería actualizar el nodo de modelado siempre que termine de trabajar en la sesión de área de trabajo interactiva y quiera guardar su trabajo. Al actualizar el nodo de modelado, el contenido de la sesión de trabajo se guarda en el nodo de Minería de textos que originó la sesión de área de trabajo interactiva. Esto no cierra la ventana de salida.

Importante: Esta actualización no guardará su ruta. Para guardar la ruta, hágalo en la ventana principal de IBM SPSS Modeler después de actualizar el nodo de modelado.

Para actualizar el nodo de modelado

1. En los menús, seleccione **Archivo > Actualizar nodo de modelado**. El nodo de modelado se actualiza con los valores de construcción y extracción, junto con las opciones y las categorías que tenga.

Cierre y finalización de sesiones

Cuando haya terminado de trabajar en la sesión, puede dejar la sesión de tres formas:

- **Guardar.** Esta opción le permite guardar el trabajo en el nodo de modelo de origen para futuras sesiones, así como también publicar bibliotecas para reutilizar en otras sesiones. Para obtener más información, consulte el tema “[Compartimiento de bibliotecas](#)” en la [página 184](#). Después de guardar, la ventana de sesión se cierra, y la sesión se suprime del gestor de salida en la ventana IBM SPSS Modeler.
- **Salir.** Esta opción descartará cualquier trabajo que no se haya guardado, cerrará la ventana de sesión y suprimirá la sesión del gestor de salida en la ventana IBM SPSS Modeler. Para liberar la memoria, recomendamos guardar los trabajos importantes y salir de la sesión.
- **Cerrar.** Esta opción no guardará ni descartará ningún trabajo. Esta opción cierra la ventana de sesión pero la sesión continúa ejecutándose. Puede abrir la ventana de sesión nuevamente mediante la selección de esta sesión en el gestor de salida en la ventana IBM SPSS Modeler.

Para cerrar la sesión de entorno de trabajo

1. En los menús, seleccione **Archivo > Cerrar**.

Accesibilidad desde el teclado

La interfaz de área de trabajo interactiva ofrece atajos de teclado para hacer más accesible la funcionalidad del producto. En el nivel más básico, puede pulsar Alt y la tecla adecuada para activar menús de ventana (por ejemplo Alt+F para acceder al menú de Archivo) o presionar la tecla tabulador para desplazarse por los controles del cuadro de diálogo. Esta sección abarca los métodos abreviados del teclado para la navegación alternativa. Hay otros atajos de teclado para la interfaz IBM SPSS Modeler.

Tabla 13. Métodos abreviados genéricos del teclado

Tecla de atajo	Función
Ctrl+1	Mostrar la primera pestaña en un panel con pestañas.
Ctrl+2	Mostrar la segunda pestaña en un panel con pestañas.
Ctrl+A	Seleccionar todos los elementos del panel que tiene el foco.
Ctrl+C	Copiar el texto seleccionado al portapapeles.

Tabla 13. Métodos abreviados genéricos del teclado (continuación)

Tecla de atajo	Función
Ctrl+E	Iniciar extracción en las vistas Categorías y conceptos y Análisis de enlaces de texto.
Ctrl+F	Visualizar la barra de herramientas Buscar en Editor de recursos/Editor de plantillas, si ya no está visible y poner el foco allí.
Ctrl+I	En la vista Categorías y conceptos, iniciar el cuadro de diálogo Definiciones de categoría para la categoría seleccionada. En la vista Clúster, iniciar el cuadro de diálogo Definiciones de Clúster para el clúster seleccionado.
Ctrl+R	Abrir el cuadro de diálogo Añadir términos en Editor de recursos/Editor de plantillas.
Ctrl+T	Abrir el cuadro de diálogo Propiedades de tipo para crear un nuevo tipo en Editor de recursos/Editor de plantillas.
Ctrl+V	Pegar el contenido del portapapeles.
Ctrl+X	Cortar elementos seleccionados en Editor de recursos/Editor de plantillas.
Ctrl + Y	Rehacer la última acción en la vista.
Ctrl+Z	Deshacer la última acción en la vista.
F1	Mostrar la ayuda, o desde un cuadro de diálogo, mostrar la ayuda al contexto de un elemento.
F2	Entrar y salir del modo de edición en las casillas de la tabla.
F6	Mover el foco entre los paneles principales en la vista activa.
F8	Mover el foco a las barras de división del panel para cambiar el tamaño.
F10	Expandir el menú Archivo principal.
flecha arriba, flecha abajo	Cambiar el tamaño del panel verticalmente cuando la barra de división está seleccionada.
flecha izquierda, flecha derecha	Cambiar el tamaño del panel horizontalmente cuando la barra de división está seleccionada.
Inicio, Fin	Cambiar el tamaño de los paneles a su tamaño máximo y mínimo cuando la barra de división está seleccionada.
Pestaña	Avanzar por los elementos en la ventana, panel o cuadro de diálogo.
Shift+F10	Mostrar el menú de contexto de un elemento.
Mayús+Tabulador	Retroceder por los elementos de la ventana o del cuadro de diálogo.
Mayús+flecha	Seleccionar caracteres en el campo de edición cuando se está en el modo de edición (F2).
Ctrl+Tab	Avanzar el foco a la siguiente área principal de la ventana.
Mayús+Control+Tab	Retroceder el foco al área principal anterior de la ventana.

Métodos abreviados de los cuadros de diálogo

Hay varios métodos abreviados y teclas de lectura de pantalla que resultan útiles cuando se trabaja con cuadros de diálogo. Al entrar en un cuadro de diálogo, puede que deba pulsar el tabulador para colocar el foco en el primer control e iniciar el lector de la pantalla. En la tabla siguiente se detalla una lista completa de los métodos abreviados especiales del teclado y el lector de pantallas.

Tabla 14. Atajos para los cuadros de diálogo

Tecla de atajo	Función
Pestaña	Avanzar por los elementos de la ventana o del cuadro de diálogo.
Ctrl+Tab	Avanzar desde un cuadro de texto al siguiente elemento.
Mayús+Tabulador	Retroceder por los elementos de la ventana o del cuadro de diálogo.
Mayús+Control+Tab	Retroceder desde un cuadro de texto al elemento anterior.
barra de espaciado	Seleccionar el control o el botón que contiene el foco.
Esc	Cancelar los cambios y cerrar el cuadro de diálogo.
Intro	Validar cambios y cerrar el cuadro de diálogo (equivalente al botón Aceptar). Si se encuentra en un cuadro de texto, primero debe pulsar Ctrl+Tabulador para salir del mismo.

Capítulo 8. Extracción de conceptos y tipos

Siempre que se ejecuta una ruta que inicia el entorno de trabajo interactivo, una extracción se realiza automáticamente en los datos de texto en la ruta. El resultado final de esta extracción es un conjunto de conceptos, tipos, y, en el caso donde existen patrones TLA en los recursos lingüísticos, patrones. Puede ver y trabajar con conceptos y tipos en el panel Resultados de extracción. Si desea más información, consulte [“Cómo funciona la extracción”](#) en la página 5.

Si desea ajustar los resultados de extracción, puede modificar los recursos lingüísticos y repetir. Si desea más información, consulte [“Refinar resultados de extracción”](#) en la página 89. El proceso de extracción se basa en los recursos y los parámetros en el recuadro de diálogo Extraer para dictar cómo extraer y organizar los resultados. Puede utilizar los resultados de extracción para definir la mejor parte, si no todas, de las definiciones de categoría.

Nota: A partir de la versión 18.2, se han mejorado los resultados de los conceptos extraídos (ahora son similares a los resultados de los conceptos extraídos en IBM SPSS Text Analytics for Surveys).

Resultados de extracción: Conceptos y tipos

Durante el proceso de extracción, todos los datos de texto se exploran y se identifican, se extraen y se asignan a tipos los conceptos relevantes. Cuando la extracción ha finalizado, los resultados aparecen en el panel Resultados extraídos, que se encuentra en el ángulo inferior izquierdo de la vista Conceptos y Categorías. La primera vez que inicia la sesión, la plantilla de recursos lingüísticos que seleccionó en el nodo se utiliza para extraer y organizar estos conceptos y tipos.

Nota: Si hay más resultados que puedan caber en el panel visible, puede utilizar los controles de la parte inferior del panel para desplazarse hacia delante y hacia atrás por los resultados, o escriba un número de página para ir a la misma.

Los conceptos, los tipos y los patrones TLA que se han extraído se conocen colectivamente con el nombre de **resultados de extracción**, y actúan como los descriptores, o los cimientos, de las categorías. También puede utilizar conceptos, tipos y patrones en las reglas de categoría. Además, las técnicas automáticas utilizan conceptos y tipos para generar categorías.

La minería de textos es un proceso repetitivo en el que los resultados de la extracción se revisan de acuerdo con el contexto de los datos de texto, se ajustan para generar resultados nuevos y después se reevalúan. Después de la extracción debe revisar los resultados y realizar los cambios que considere necesarios modificando los recursos lingüísticos. Puede ajustar los recursos, en parte, directamente desde el panel Resultados de la extracción, panel Datos, recuadro de diálogo Definiciones de categoría, o recuadro de diálogo Definiciones de clúster. Consulte el tema [“Refinar resultados de extracción”](#) en la página 89 para obtener más información. También puede hacerlo directamente en la vista Editor de recursos. Consulte el tema [“La vista Editor de recursos”](#) en la página 73 para obtener más información.

Después del ajuste, puede repetir la extracción para ver los nuevos resultados. El ajuste preciso desde el principio de los resultados de la extracción permite garantizar que la próxima vez que vuelva a realizar la extracción, obtendrá resultados idénticos en las definiciones de su categoría, perfectamente adaptados al contexto de los datos. De esta forma, los documentos/registros se asignarán a las definiciones de categoría de una manera más precisa y repetible.

Conceptos

Durante el proceso de extracción, los datos de texto se exploran y analizan para identificar palabras únicas que resulten de interés o relevancia (como *election* o *peace*) y frases de palabras (como *presidential election*, *election of the president* o *peace treaties*) en el texto. Estas palabras y frases se conocen colectivamente con el nombre de *términos*. Mediante los recursos lingüísticos, los términos relevantes se extraen y, a continuación, los términos similares se agrupan bajo un término principal llamado **concepto**.

Puede ver el conjunto de términos subyacentes de un concepto pasando el ratón por encima del nombre del concepto. Al hacerlo aparecerá una etiqueta con información que muestra el nombre del concepto y varias líneas de términos agrupados bajo dicho concepto. Estos términos subyacentes incluyen los sinónimos definidos en los recursos lingüísticos (independientemente de si se encontraron en el texto o no), así como los términos en plural/singular extraídos, términos permutados, términos de agrupación difusa, etc. Puede copiar estos términos o ver el conjunto completo de términos subyacentes pulsando con el botón derecho del ratón en el nombre del concepto y seleccionando la opción del menú contextual.

De forma predeterminada, los conceptos se muestran en minúsculas y se clasifican en orden descendente de acuerdo con el número de documentos (columna Doc.) . Cuando se extraen conceptos, se les asigna un tipo para contribuir a agrupar conceptos similares. Están codificados por colores según este tipo. Los colores están definidos en las propiedades de tipo en el Editor de recursos. Para obtener más información, consulte el tema [“Diccionarios de tipo”](#) en la [página 187](#).

Siempre que se esté utilizando un concepto, tipo, o patrón en una definición de categoría, aparece un icono en la tabla de ordenación **En** columna .

types

Tipos son agrupaciones semánticas de conceptos. Cuando se extraen conceptos, se les asigna un tipo para contribuir a agrupar conceptos similares. Se entregan varios tipos incorporados con IBM SPSS Modeler Text Analytics , tales como <Location>, <Organization>, <Person>, <Positive>, <Negative> and so on. Por ejemplo, el tipo <Location> agrupa palabras clave geográficas y lugares. Este tipo se asignaría a conceptos como *chicago*, *paris* y *tokyo*. Para la mayoría de idiomas, los conceptos que no se encuentran en ningún diccionario de tipo, pero que se extraen del texto, se escriben automáticamente como <Unknown> Consulte el tema [“Tipos incorporados”](#) en la [página 188](#) para obtener más información.

Cuando selecciona la vista Tipo, los tipos extraídos aparecen de forma predeterminada en orden descendente por frecuencia global . Comprobará también que estos tipos están codificados por colores para que sea más fácil distinguirlos. Los colores forman parte de las propiedades de tipo. Para obtener más información, consulte el tema [“Crear tipos”](#) en la [página 189](#). También puede crear sus propios tipos.

Patrones

Los patrones también pueden extraerse a partir de los datos de texto. Sin embargo, necesita disponer de una biblioteca que contenga algunas reglas de patrones TLA (Análisis de enlace de texto) en el Editor de recursos. También deberá optar por extraer estos patrones en IBM SPSS Modeler Text Analytics la configuración del nodo en el cuadro de diálogo Extraer utilizando la opción **Activar extracción de patrones de análisis de enlace de texto**. Para obtener más información, consulte el tema [Capítulo 11, “Exploración del Análisis de enlaces de texto”](#), en la [página 149](#).

Extraer datos

Cuando se necesita una extracción, el panel Resultados extraídos aparece de color amarillo y se muestra el mensaje **Pulse el botón Extraer para extraer conceptos** debajo de la barra de herramientas de este panel.

Puede que necesite extraer si aún no tiene resultados de extracción, si ha realizado cambios en los recursos lingüísticos y desea actualizar los resultados de extracción, o si ha vuelto a abrir un sesión en el que no ha guardado los resultados de extracción (**Herramientas > Opciones**).

Nota: Si cambia el nodo de origen para la ruta después de que los resultados de la extracción se hayan almacenado en la memoria caché con la opción **Utilizar trabajo de sesión...**, deberá ejecutar una nueva extracción una vez que se inicie la sesión interactiva del entorno de trabajo si desea obtener resultados de extracción actualizados.

Cuando ejecuta una extracción, aparece un indicador de progreso que ofrece información sobre el estado de la extracción. Durante este tiempo, el motor de extracción lee todos los datos de texto e identifica los términos y patrones relevantes, los extrae y les asigna un tipo. A continuación, el motor intenta agrupar los términos sinónimos bajo un término principal, llamado concepto. Cuando finaliza el proceso, los conceptos, tipos y patrones resultantes aparecen en el panel Resultados extraídos.

El proceso de extracción genera un conjunto de conceptos y tipos, así como de patrones TLA (Análisis de enlace de texto), si está activado. Puede ver y trabajar con estos conceptos y tipos en el panel Resultados extraídos en las vistas Conceptos y Categorías. Si ha extraído patrones TLA, puede verlos en la vista Análisis de enlace de texto.

Nota: Hay relación entre el tamaño de conjuntos de datos y el tiempo que lleva completar el proceso de extracción. Siempre puede considerar insertar un nodo de ejemplo en sentido ascendente u optimizar la configuración de su máquina.

Extraer datos

1. En los menús elija **Herramientas > Extraer**. Como alternativa, pulse en el botón de la barra de herramientas **Extraer**.
2. Si opta por que siempre se muestre el diálogo Configuración de extracción, aparecerá para que pueda realizar cambios. Más adelante en este tema encontrará los descriptores de cada valor de configuración.
3. Pulse **Extraer** para empezar el proceso de extracción. Una vez se inicia la extracción, se abre el cuadro de diálogo de progreso. Después de la extracción, los resultados aparecen en el panel Resultados extraídos. De forma predeterminada, los conceptos se muestran en minúsculas y se clasifican en orden descendente de acuerdo con el número de documentos (columna Doc.) .

Puede repasar los resultados utilizando las opciones de la barra de herramientas y establecer los resultados con un orden diferente, o cambiar de vista (conceptos o tipos). También puede refinar los resultados de la extracción trabajando con los recursos lingüísticos. Para obtener más información, consulte el tema [“Refinar resultados de extracción”](#) en la página 89.

Posibles problemas de extracción

Varias sesiones de área de trabajo interactivas pueden provocar un comportamiento muy lento. SPSS Modeler Text Analytics y SPSS Modeler comparten un motor de tiempo de ejecución Java cuando se inicia una sesión de área de trabajo. En función del número de sesiones de área de trabajo interactivas que invoque durante una sesión de SPSS Modeler, aunque abra y cierre la misma sesión, es posible que la memoria del sistema provoque que la aplicación funcione muy lentamente. Este efecto puede ser especialmente apreciable si está trabajando con una gran cantidad de datos o tiene una máquina que tenga menos RAM que el valor recomendado de 4 GB. Si observa que la máquina responde muy lentamente, es recomendable que guarde todos sus trabajos, cierre SPSS Modeler, y vuelva a iniciar la aplicación. Ejecutar SPSS Modeler Text Analytics en una máquina con menos memoria de la recomendada, especialmente cuando se trabaje con grandes conjuntos de datos o durante prolongados períodos de tiempo, puede provocar que Java que quede sin memoria, y se cierre. Se recomienda encarecidamente actualizar al valor al recomendado, o más, de memoria (o utilice el servidor de SPSS Modeler Text Analytics) si trabaja con cantidades de datos grandes.

Para textos en neerlandés, inglés, francés, alemán, italiano, portugués y español

El cuadro de diálogo Configuración de extracción contiene algunas opciones básicas de extracción.

Habilitar extracción de patrones de análisis de enlaces de textos. Especifica que desea extraer patrones TLA a partir de los datos de texto. También presupone que dispone de reglas de patrones TLA en una de sus bibliotecas en el Editor de recursos. Esta opción puede prolongar significativamente el proceso de extracción. Para obtener más información, consulte el tema [Capítulo 11, “Exploración del Análisis de enlaces de texto”](#), en la página 149.

Arreglar errores de puntuación. Esta opción normaliza temporalmente el texto que contiene errores de puntuación (por ejemplo, el uso inapropiado) durante la extracción para mejorar la capacidad de extracción de los conceptos. Esta opción es extremadamente útil cuando el texto es breve y de calidad mediocre (como, por ejemplo, en respuestas de encuestas con final abierto, correo electrónico y datos CRM) o cuando el texto contiene muchas abreviaturas.

Acomodar la ortografía a una longitud mínima de caracteres de palabra de [n] Esta opción aplica una técnica de agrupación difusa que ayuda a agrupar bajo un concepto las palabras que suelen escribirse

mal o que tienen una ortografía parecida. El algoritmo de agrupación difusa elimina temporalmente todas las vocales (excepto la primera) y elimina las consonantes dobles/triples de las palabras extraídas y luego las compara para ver si son iguales, de forma que `modeling` y `modelling` se agruparían. Sin embargo, si a cada término se le asigna un tipo diferente, excluyendo el tipo `<Unknown>`, la técnica de agrupación difusa no se aplicará.

También puede definir el número mínimo de caracteres *raíz* necesarios para poder utilizar la agrupación difusa. El número de caracteres raíz de un término se calcula sumando todos los caracteres y restando los que forman los sufijos de las declinaciones, y en el caso de términos de palabras compuestas, también los determinantes y las preposiciones. Por ejemplo, el término `exercises` se contaría como 8 caracteres raíz en la forma `"exercise,"` ya que la letra `s` al final de la palabra es una inflexión (forma plural. De forma similar, `apple sauce` se cuenta como 10 caracteres raíz (`"apple sauce"`) y `manufacturing of cars` se cuenta como 16 caracteres raíz (`"manufacturing car"`). Este método de recuento de caracteres solo se utiliza para comprobar si debe aplicarse la agrupación difusa, pero no influye en la forma de coincidencia de las palabras.

Nota: Si encuentra que ciertas palabras posteriormente se agrupan incorrectamente, puede excluir pares de palabras de esta técnica declarándolos explícitamente en la sección **Agrupación difusa: Excepciones** en la pestaña Recursos avanzados. Para obtener más información, consulte el tema ["Agrupación difusa"](#) en la página 201.

Extraer unitérminos Esta opción extrae palabras simples (unitérminos) siempre que la palabra no forme parte de una palabra compuesta, y si es un sustantivo o una categoría léxica no reconocida.

Extraer entidades no lingüísticas Esta opción extrae entidades no lingüísticas, como números de teléfono, números de la seguridad social, horas, fechas, monedas, dígitos, porcentajes, direcciones de correo electrónico y direcciones de HTTP. Puede incluir o excluir ciertos tipos de entidades no lingüísticas en la sección **Entidades no lingüísticas: Configuración** de la pestaña Recursos avanzados. Si se desactivan las entidades innecesarias, el motor de extracción no malgastará tiempo de proceso. Para obtener más información, consulte el tema ["Configuración"](#) en la página 205.

Algoritmo de mayúsculas Esta opción extrae términos simples y compuestos que no están en los diccionarios incorporados, siempre que la primera letra del término esté en mayúscula. Esta opción supone un buen método para extraer la mayoría de los nombres propios.

Agrupar nombres de persona parciales y completos juntos cuando sea posible Esta opción agrupa nombres que aparecen de diferente manera juntos en el texto. Esta característica es útil porque a menudo se hace referencia a los nombres completos al principio del texto, y más adelante se utiliza la versión abreviada. Esta opción intenta hacer coincidir cualquier unitérmino que tenga el tipo `<Unknown>` con la última palabra de cualquier término compuesto que se haya tipificado como `<Person>`. Por ejemplo, si se encuentra *garcía*, que inicialmente se tipificó como `<Unknown>`, el motor de extracción comprobará si hay algún término compuesto en el tipo `<Person>` con el término *garcía* como la última palabra, como en *juan garcía*. Esta opción no se aplica a los nombres propios, porque muchos de ellos no se extraen nunca como unitérminos.

Permutación máxima de palabra no funcional Esta opción especifica el número máximo de palabras no funcionales que debe haber para poder aplicar la técnica de permutación. Esta técnica de permutación agrupa frases similares que difieren entre sí solo en las palabras no funcionales (por ejemplo, `de y el`), independientemente de la flexión. Por ejemplo, digamos que ha establecido este valor en dos palabras como máximo y tanto `company officials` como `officials of the company` se han extraído. En este caso, ambos términos extraídos se agruparían en la lista de conceptos final, ya que ambos términos se consideran iguales cuando se omite `of the`.

Utilizar la derivación al agrupar multitérminos Cuando procese grandes volúmenes de datos, seleccione esta opción para agrupar multitérminos mediante las reglas de derivación.

Opción de índice para el mapa de conceptos especifica que desea generar el índice del mapa durante la extracción para que puedan dibujarse los mapas de conceptos rápidamente en otro momento. Pulse **Configuración** para editar la configuración del índice. Para obtener más información, consulte el tema ["Creación de índices de mapas de conceptos"](#) en la página 88.

Mostrar siempre este diálogo antes de comenzar una extracción Especifique si desea ver el diálogo Configuración de extracción cada vez que realice una extracción, si no desea verlo nunca a menos que

vaya al menú Herramientas, o si desea que cada vez que realice una extracción se le pregunte si desea editar algún valor de la configuración de extracción.

Filtración de resultados de extracción

Cuando está trabajando con conjuntos de datos muy grandes, el proceso de extracción puede producir millones de resultados. Para varios usuarios, puede hacer que sea más difícil revisar los resultados de manera efectiva. Por lo tanto, para acercarse a los que son más interesantes, puede filtrar estos resultados mediante el diálogo Filtrar, disponible en el panel Resultados extraídos.

Tenga en cuenta que todos los valores de este diálogo Filtro se utilizan conjuntamente para filtrar los resultados de extracción que están disponibles para las categorías.

Filtrar por frecuencia Puede filtrar para visualizar sólo esos resultados con un cierto valor global o valor de frecuencia de documento .

- **Frecuencia global** es el número total de veces que un concepto aparece en el conjunto completo de documentos o registros y se muestra en la columna **Global**.
- **Frecuencia de documentos** es el número total de documentos o registros en los que un concepto aparece y se muestra en la columna **Documentos**.

Por ejemplo, si el concepto nato apareció 800 veces en 500 registros, podríamos decir que este concepto tiene una frecuencia global de 800 y una frecuencia de documento de 500.

Y por tipo Puede filtrar para visualizar únicamente los resultados que pertenecen a determinados tipos. Puede elegir todos los tipos o sólo tipos específicos.

Y por coincidencia de texto También puede filtrar para visualizar sólo los resultados que coincidan con la regla que defina aquí. Especifique el conjunto de caracteres que van a coincidir en el campo **Coincidir texto** y después seleccione la condición en cuál aplicar la coincidencia.

Tabla 15. Coincidir condiciones de texto

Condición	Descripción
Contiene	El texto coincide si la cadena ocurre en cualquier lado. (opción predeterminada)
Comienza con	El texto coincide sólo si el concepto o tipo comienza con el texto especificado.
Termina con	El texto coincide sólo si el concepto o tipo termina con el texto especificado.
Coincidencia exacta	La serie entera debe coincidir con el nombre de tipo o concepto.

Resultados mostrados en el panel Resultados de extracción

A continuación se muestran algunos ejemplos de cómo los resultados pueden visualizarse, en inglés, en la barra de herramientas del panel Resultados de extracción basados en los filtros.

Tabla 16. Ejemplos de comentarios sobre el filtro

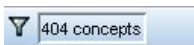
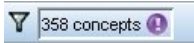
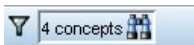
Comentarios sobre el filtro	Descripción
	La barra de herramientas muestra el número de resultados. Puesto que no había ningún texto coincidente de filtro y el máximo no se cumplió, no se muestran iconos adicionales.
	La barra de herramientas muestra que los resultados fueron limitados al máximo especificado en el filtro, que en este caso fue de 300. Si aparece un icono púrpura, esto significa que el número máximo de conceptos se ha cumplido. Pase el ratón por encima del icono para obtener más información.

Tabla 16. Ejemplos de comentarios sobre el filtro (continuación)

Comentarios sobre el filtro	Descripción
	La barra de herramientas muestra que los resultados fueron limitados utilizando un filtro de texto coincidente. A esto lo muestra el icono de lupa.

Para filtrar los resultados

1. En los menús elija **Herramientas > Filtrar**. Se abre el cuadro de diálogo Filtrar.
2. Seleccione y perfeccione los filtros que desea utilizar.
3. Pulse **Aceptar** para aplicar los filtros y ver los nuevos resultados en el panel Resultados de extracción.

Exploración de mapas de conceptos

Puede crear un mapa de conceptos para explorar cómo los conceptos están interrelacionados. Al seleccionar un solo concepto y pulsando **Mapa**, una ventana de mapa conceptos se abre para que pueda explorar el conjunto de conceptos que estén relacionados con el concepto seleccionado. Puede filtrar los conceptos que se visualizan editando los valores como qué tipo puede incluir, qué tipo de relación puede buscar y así sucesivamente.

Importante: Antes de que se pueda crear un mapa, se debe generar un índice. Este proceso puede durar unos minutos. Sin embargo, una vez que haya generado el índice, no tiene que generarlo de nuevo hasta volver a extraer. Si desea que el índice se genere automáticamente cada vez que se extrae, seleccione dicha opción en la configuración de extracción. Para obtener más información, consulte el tema [“Extraer datos”](#) en la página 82.

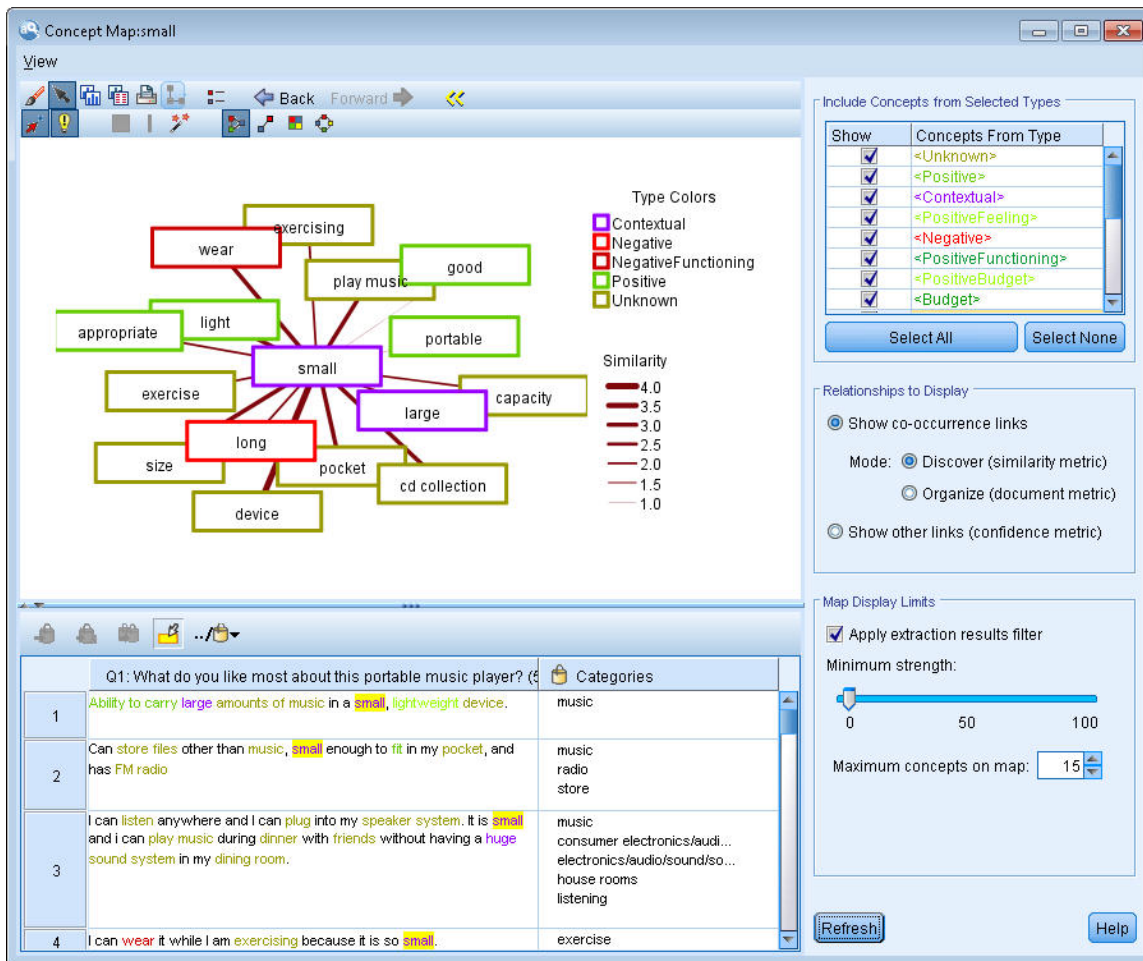


Figura 27. Un mapa de concepto para el concepto seleccionado

Para ver un mapa de conceptos

1. En el panel de Resultados de extracción, seleccione una concepto único.
2. En la barra de herramientas de este panel, pulse el botón **Mapa**. Si el índice de mapa ya se ha generado el concepto de mapa se abre en un diálogo separado. Si el índice de correlación no se ha generado o estaba fuera de la fecha, el índice debe ser reconstruido. Este proceso puede durar varios minutos.
3. Pulse alrededor del mapa para explorar. Si pulsa dos veces sobre un concepto enlazado, el mapa se volverá a diseñar y le mostrará los conceptos enlazados para el concepto al que acaba de pulsar dos veces.
4. La barra superior ofrece algunas herramientas de mapa básicas, como volver a un mapa anterior, filtrar enlaces de acuerdo con los puntos fuertes de la relación, y abrir el diálogo de filtro para controlar los tipos de conceptos que aparecen, así como el tipo de relaciones que representan. Una segunda línea de barra de herramientas contiene herramientas de edición de gráficos. Para obtener más información, consulte el tema [“Uso de barras de herramientas y paletas de gráficos”](#) en la página 163.
5. Si no está satisfecho con los tipos de enlaces que se han encontrado, revise los valores para esta presentación de correlación en la parte derecha de la correlación.

Valores de correlación: Incluye conceptos de tipos seleccionados

Sólo los conceptos que pertenecen a los tipos seleccionados en la tabla se muestran en la correlación. Para ocultar los conceptos de un tipo determinado, desmarque ese tipo en la tabla.

Valores de correlación: Relaciones para visualizar

Mostrar enlaces de co-ocurrencia Si desea mostrar los enlaces de co-ocurrencia, elija la modalidad. La modalidad afecta cómo el nivel de enlace se ha calculado.

- *Descubrir (métrica de similitud)*. Con esta medida, la potencia del enlace se calcula utilizando un cálculo más complejo que tiene en cuenta con qué frecuencia aparecen dos conceptos aparte, así como con qué frecuencia aparecen juntos. Un valor alto de potencia significa que un par de conceptos tienden a aparecer más frecuentemente juntos que lo que aparecen separados. Con la fórmula siguiente, todos los valores de coma flotante se convierten en enteros.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Figura 28. Fórmula de coeficiente de similitud

En esta fórmula, C_I es el número de documentos o registros en los que el concepto I ocurre.

C_J es el número de documentos o registros en el que aparece el concepto J.

C_{IJ} es el número de documentos o registros en que el par de conceptos I y J vuelven a ocurrir en el conjunto de documentos.

- *Organización (métrica de documentos)*. La fuerza de los vínculos con esta medida está determinado por el recuento en bruto de re-ocurrencias. En general, cuanto más frecuentes sea los dos conceptos, más probable es, más probable es que vayan a aparecer juntos. Un valor de potencia alta significa que un par de conceptos aparecen juntos con frecuencia.

Mostrar otros enlaces (métrica de confianza). Puede elegir otros enlaces para visualizar; estos pueden ser semánticos, de derivación (morfológico) o inclusión (sintácticos) y están relacionados con cuántos pasos eliminados de un concepto están del concepto al que está enlazado. Estos le pueden ayudar a ajustar los recursos, particularmente la sinonimia o a desambiguar. Para descripciones cortas de cada uno de estos grupos de técnicas, consulte [“Configuración lingüística avanzada”](#) en la página 108

Nota: Recuerde que si estos no se han seleccionado cuando se generó el índice o si no se ha encontrado ninguna relación, no se visualizará ninguno. Para obtener más información, consulte el tema [“Creación de índices de mapas de conceptos”](#) en la página 88.

Valores de correlación: Límites de la visualización de la correlación

Aplique filtros de resultados de extracción. Si no desea utilizar todos los conceptos, puede utilizar el filtro en el panel de resultados de extracción para limitar lo que se muestra. Después seleccione esta opción y IBM SPSS Modeler Text Analytics buscará conceptos relacionados utilizando este conjunto filtrado. Para obtener más información, consulte el tema [“Filtración de resultados de extracción”](#) en la página 85.

Potencia mínima. Establezca la potencia de enlace mínimo aquí. Los conceptos relacionados con una potencia de relación inferior a este límite se ocultarán del mapa.

Número máximo de conceptos de correlación. Especifique el número máximo de relaciones para mostrar en el mapa.

Creación de índices de mapas de conceptos

Antes de crear un mapa, se debe generar un índice de relaciones de concepto. Siempre que crea un mapa de concepto, IBM SPSS Modeler Text Analytics hace referencia a este índice. Puede elegir cuáles relaciones indexar seleccionando las técnicas en este diálogo.

Agrupación de técnicas. Elija una o más técnicas. Para obtener descripciones de cada una de estas técnicas, consulte [“Acerca de técnicas lingüísticas”](#) en la página 110 No todas las técnicas están disponibles para todos los idiomas de texto.

Impedir el emparejamiento de conceptos específicos. Seleccione esta casilla de verificación para detener el proceso de agrupación o emparejamiento de dos conceptos en la salida. Para crear o

gestionar pares de conceptos, pulse **Gestionar pares**. Para obtener más información, consulte el tema [“Administración de pares de excepciones de enlace”](#) en la página 110.

Construir el índice puede tardar varios minutos. Sin embargo, una vez que haya generado el índice, no tiene que generarlo de nuevo hasta que lo vuelve a extraer o a menos que quiera cambiar los valores para incluir más relaciones. Si desea generar un índice cada vez que extraiga, puede seleccionar esa opción en los valores de extracción. Para obtener más información, consulte el tema [“Extraer datos”](#) en la página 82.

Refinar resultados de extracción

La extracción es un proceso iterativo mediante el cual puede extraer, revisar los resultados, realizar cambios en los mismos y, después, volver a extraer para actualizar los resultados. Puesto que la precisión y la continuidad son básicas para una categorización y minería de textos satisfactorias, ajustar de forma precisa los resultados de extracción desde el principio garantiza que, cada vez que vuelva a extraer, obtendrá de forma precisa los mismos resultados en las definiciones de categoría. De esta forma, los registros y los documentos se asignarán a las categorías de una manera más precisa y repetible.

Los resultados de la extracción constituyen los cimientos de las categorías. Cuando crea categorías utilizando estos resultados de extracción, los registros y documentos se asignan automáticamente a categorías si contienen texto que coincida con uno o más descriptores de categoría. Aunque empiece a categorizar antes de realizar ajustes en los recursos lingüísticos, conviene revisar los resultados de la extracción al menos una vez antes de comenzar.

Mientras revisa los resultados puede encontrar elementos que desearía que el motor de extracción gestionara de otra forma. Observe los ejemplos siguientes:

- **Sinónimos no reconocidos.** Suponga que encuentra varios conceptos que considera sinónimos, como *smart*, *intelligent*, *bright* y *knowledgeable*, y todos aparecen como conceptos individuales en los resultados de la extracción. Puede crear una definición de sinónimo en la que *intelligent*, *bright* y *knowledgeable* estén todos agrupados bajo el concepto objetivo *smart*. Hacerlo los agruparía todos con *smart*, y el recuento de frecuencia global sería más alto también. Para obtener más información, consulte el tema [“Añadir sinónimos”](#) en la página 90.
- **Conceptos mal escritos.** Supongamos que los conceptos de los resultados extraídos aparecen en un tipo y desea que se asignen a otro distinto. En otro ejemplo, imagine que encuentra 15 conceptos de verduras en sus resultados de extracción y desea que todos se añadan a un nuevo tipo llamado <Vegetable>. Para la mayoría de idiomas, los conceptos que no se encuentran en ningún diccionario de tipo, pero que se extraen del texto, se escriben automáticamente como <Unknown> Puede añadir conceptos a tipos. Para obtener más información, consulte el tema [“Añadir conceptos a tipos”](#) en la página 91.
- **Conceptos insignificantes.** Supongamos que encuentra un concepto que se ha extraído y que tiene un recuento de frecuencia muy alta, es decir, que se encuentra en muchos registros o documentos. Sin embargo, considera que es un concepto sin trascendencia para su análisis. Puede excluirlo de la extracción. Para obtener más información, consulte el tema [“Excluir conceptos de una extracción”](#) en la página 92.
- **Coincidencias incorrectas.** Suponga que al revisar los registros o documentos que contienen un determinado concepto, descubre que dos palabras se han agrupado incorrectamente, por ejemplo *faculty* y *facility*. Esta coincidencia puede deberse a un algoritmo interno, al que se conoce como agrupación difusa, que pasa por alto provisionalmente las vocales y las consonantes duplicadas o triplicadas con el objeto de agrupar errores ortográficos comunes. Puede añadir estas palabras a una lista de parejas de palabras que no deben agruparse. Consulte el tema [“Agrupación difusa”](#) en la página 201 para obtener más información.
- **Conceptos no extraídos.** Supongamos que espera encontrar determinados conceptos extraídos, pero al revisar el texto del registro o documento detecta que algunas palabras o frases no se han extraído. A menudo, estas palabras son verbos o adjetivos en los que no está interesado. Sin embargo, algunas veces sí desea utilizar una palabra o frase que no se haya extraído como parte de una definición de categoría. Para extraer estos conceptos, puede forzar un término en un diccionario de tipo. Para obtener más información, consulte el tema [“Forzado de palabras en la extracción”](#) en la página 93.

Muchos de estos cambios pueden realizarse directamente desde el panel de resultados de extracción , panel de datos, recuadro de diálogo de definiciones de categoría, o recuadro de diálogo de definiciones de clúster seleccionando uno o más elementos y efectuando una pulsación con el botón derecho del ratón para acceder a los menús contextuales.

Después de realizar los cambios, el color de fondo del panel cambia para mostrar que debe volver a extraer para ver los cambios. Para obtener más información, consulte el tema [“Extraer datos”](#) en la [página 82](#). si está trabajando con conjuntos de datos más grandes, podría ser más eficaz volver a extraer tras realizar varios cambios, en lugar de después de cada cambio.

Nota: Puede ver todo el conjunto de recursos lingüísticos editables utilizados para producir los resultados de extracción en la vista Editor de recursos (Ver > Editor de recursos). Estos recursos aparecen en forma de bibliotecas y diccionarios en esta vista. Puede personalizar los conceptos y los tipos directamente en las bibliotecas y en los diccionarios. Para obtener más información, consulte el tema [Capítulo 15, “Trabajo con bibliotecas”](#), en la [página 179](#).

Añadir sinónimos

Los *Sinónimos* asocian dos o más palabras con el mismo significado. Los sinónimos se utilizan a menudo para agrupar términos con sus abreviaturas, o para agrupar palabras que suelen escribirse mal con la ortografía correcta. Con el uso de los sinónimos, la frecuencia del concepto de destino es mayor, lo cual facilita mucho la detección de información similar que se presenta de distintas formas en los datos de texto.

Las bibliotecas y las plantillas de los recursos lingüísticos que se proporcionan con el producto contienen muchos sinónimos predefinidos. Sin embargo, si detecta sinónimos que no se reconocen, puede definirlos para que la próxima vez que realice una extracción puedan detectarse.

El primer paso es decidir cuál será el concepto objetivo. El *concepto objetivo* es aquel bajo el que desea agrupar todos los términos sinónimos en los resultados finales. Durante la extracción, los sinónimos se agrupan bajo este concepto objetivo. El segundo paso es identificar todos los sinónimos de este concepto. El concepto objetivo se sustituye por todos sus sinónimos en la extracción final. Para que un término sea considerado sinónimo, debe estar extraído. Sin embargo, no es necesario extraer el concepto objetivo para que se produzca la sustitución. Por ejemplo, si desea que *intelligent* se sustituya por *smart*, *intelligent* es el sinónimo y *smart* es el concepto objetivo.

Si crea una nueva definición de sinónimo, se añade un nuevo concepto de destino al diccionario. A continuación debe añadir sinónimos al concepto objetivo. Siempre que cree o edite sinónimos, estos cambios quedarán registrados en los diccionarios de sinónimos del Editor de recursos. Si desea ver el contenido completo de estos diccionarios de sinónimos o si desea realizar un número importante de cambios, puede trabajar directamente en Editor de recursos. Para obtener más información, consulte el tema [“Diccionarios de sustitución/sinónimos”](#) en la [página 194](#).

Los sinónimos nuevos se almacenarán automáticamente en la primera biblioteca listada en el árbol de bibliotecas en la vista Editor de recursos, de forma predeterminada, esta es la **Biblioteca local**.

Nota: Si busca una definición de sinónimo y no puede encontrarla a través de los menús contextuales o directamente en Editor de recursos, puede que se haya dado una coincidencia desde una técnica de agrupación difusa interna. Para obtener más información, consulte el tema [“Agrupación difusa”](#) en la [página 201](#).

Crear un sinónimo nuevo

1. En el panel Resultados de extracción , panel Datos, recuadro de diálogo Definiciones de categoría, o recuadro de diálogo Definiciones de clúster, seleccione los conceptos para los que desee crear un sinónimo nuevo.
2. Desde los menús, seleccione **Edición > Añadir a sinónimo > Nuevo**. Aparecerá el cuadro de diálogo Crear sinónimo.
3. Escriba un concepto objetivo en el cuadro de texto Objetivo. Se trata del concepto bajo el que se agruparán todos los sinónimos.

4. Si desea añadir más sinónimos, escríbalos en el cuadro de lista Sinónimos. Utilice el separador global para separar cada término de sinónimo. Para obtener más información, consulte el tema [“Opciones: separador Sesión”](#) en la página 75.
5. Pulse **Aceptar** para aplicar los cambios. El cuadro de diálogo se cierra y el color de fondo del panel Resultados extraídos cambia para indicar que debe repetir la extracción para ver los cambios. Si tiene previstos varios cambios, realícelos antes de repetir la extracción.

Añadir un sinónimo

1. En el panel Resultados de extracción , panel Datos, recuadro de diálogo Definiciones de categoría, o recuadro de diálogo Definiciones de clúster, seleccione los conceptos que desee añadir a una definición de sinónimo existente.
2. Desde los menús, seleccione **Edición > Añadir a sinónimo** . El menú muestra un conjunto de sinónimos, y el creado más recientemente figura al principio de la lista. Seleccione el nombre del sinónimo al que desea añadir los conceptos seleccionados. Si encuentra el sinónimo que está buscando, selecciónelo; acto seguido los conceptos seleccionados se añaden a dicha definición de sinónimo. Si no lo encuentra, seleccione **Más** para mostrar el cuadro de diálogo Todos los sinónimos.
3. En el cuadro de diálogo Todos los sinónimos, puede ordenar la lista por orden de clasificación natural (orden de creación) o en sentido ascendente o descendente. Seleccione el nombre del sinónimo al que desea añadir los conceptos seleccionados y pulse en **Aceptar**. El cuadro de diálogo se cierra y los conceptos se añaden a la definición de sinónimos.

Añadir conceptos a tipos

Siempre que se ejecuta una extracción, los conceptos extraídos se asignan a tipos en un esfuerzo de agrupar términos que tienen algo en común. IBM SPSS Modeler Text Analytics se entrega con muchos tipos incorporados. Consulte el tema [“Tipos incorporados”](#) en la [página 188](#) para obtener más información. Para la mayoría de idiomas, los conceptos que no se encuentran en ningún diccionario de tipo, pero que se extraen del texto, se escriben automáticamente como <Unknown>

Cuando revise los resultados, es posible que encuentre algunos conceptos en un tipo que en realidad desea que se asigne a otro concepto, o un grupo de palabras que en realidad pertenece a un nuevo tipo por sí mismo. En estos casos, puede reasignar estos conceptos a otro tipo o crear un tipo nuevo.

Por ejemplo, supongamos que está trabajando con los datos de una encuesta relacionada con el mundo automovilístico, y le interesa que la categorización se centre en diferentes ámbitos de los vehículos. Puede crear un tipo llamado <Dashboard> para agrupar todos los conceptos relacionados con los indicadores y mandos que se encuentran en el salpicadero de los vehículos. A continuación, puede asignar conceptos como `gas gauge`, `heater`, `radio` y `odometer` a ese nuevo tipo.

En otro ejemplo, supongamos que está trabajando con datos de encuesta relacionados a universidades y colegios, y la extracción ha tomado el tipo de Johns Hopkins (la universidad) como un tipo de <Person> en lugar de como un tipo de <Organization>. En este caso, puede añadir este concepto al tipo <Organization>.

Siempre que se crea un tipo o se añaden conceptos a un tipo en una lista de términos del tipo, estos cambios quedan registrados en los diccionarios de tipo en las bibliotecas de recursos lingüísticos de Editor de recursos. Si desea ver el contenido de estas bibliotecas o si desea realizar un número importante de cambios, puede trabajar directamente en Editor de recursos. Para obtener más información, consulte el tema [“Añadir términos”](#) en la [página 190](#).

Añadir un concepto a un tipo

1. En el panel Resultados de extracción , panel Datos, recuadro de diálogo Definiciones de categoría, o recuadro de diálogo Definiciones de clúster, seleccione los conceptos que desee añadir al tipo existente.
2. Pulse el botón derecho del ratón para abrir el menú contextual.

3. Desde los menús, seleccione **Edición > Añadir a tipo** . El menú muestra un conjunto de tipos, y el creado más recientemente figura al principio de la lista. Seleccione el nombre del tipo al que desea añadir los conceptos seleccionados. Si encuentra el nombre del tipo que está buscando, selecciónelo; acto seguido los conceptos seleccionados se añaden a dicho tipo. Si no lo encuentra, seleccione **Más** para mostrar el cuadro de diálogo Todos los tipos.
4. En el cuadro de diálogo Todos los tipos, puede ordenar la lista por orden de clasificación natural (orden de creación) o en sentido ascendente o descendente. Seleccione el nombre del tipo al que desea añadir los conceptos seleccionados y pulse en **Aceptar**. El cuadro de diálogo se cierra y se añadirán como términos al tipo.

Crear un tipo nuevo

1. En el panel Resultados de extracción , panel Datos, recuadro de diálogo Definiciones de categoría, o recuadro de diálogo Definiciones de clúster, seleccione los conceptos para los que desee crear un tipo nuevo.
2. Desde los menús, seleccione **Edición > Añadir a tipo > Nuevo**. Aparecerá el cuadro de diálogo Propiedades de tipo.
3. Escriba un nombre nuevo para este tipo en el cuadro de texto Nombre y realice los cambios necesarios en el resto de los campos. Para obtener más información, consulte el tema [“Crear tipos” en la página 189](#).
4. Pulse **Aceptar** para aplicar los cambios. El recuadro de diálogo se cierra y el color de fondo del panel Resultados de extracción cambia, lo que indica que debe volver a extraer para ver los cambios. Si tiene varios cambios, realícelos antes de volver a extraer.

Excluir conceptos de una extracción

Al revisar los resultados puede encontrar, ocasionalmente, conceptos que no deseaba que se extrajeran ni que se utilizaran en ninguna técnica automática de generación de categorías. En algunos casos, estos conceptos tienen un recuento de frecuencia muy alto y son totalmente intrascendentes en su análisis. En ese caso, puede marcar un concepto para que se excluya de la extracción final. Por lo general, los conceptos que añada a esta lista serán palabras o frases de relleno que se utilizan en el texto para conferir continuidad, pero que no aportan información relevante y que pueden cargar innecesariamente los resultados de la extracción. Si añade conceptos al diccionario de exclusión, tendrá la seguridad de que no se extraerán nunca.

Al excluir estos conceptos, la próxima vez que realice la extracción desaparecerán de los resultados de la misma todas las variaciones de los conceptos excluidos. Si este concepto ya aparece como descriptor en una categoría, permanecerá en la categoría con un recuento de cero después de volver a extraer.

Cuando realiza la exclusión, estos cambios quedan registrados en un diccionario de exclusión en el Editor de recursos. Si desea ver todas las definiciones excluidas y editarlas, puede trabajar directamente en Editor de recursos. Para obtener más información, consulte el tema [“Diccionarios de exclusión” en la página 197](#).

Excluir conceptos

1. En el panel Resultados de extracción , panel Datos, recuadro de diálogo Definiciones de categoría, o recuadro de diálogo Definiciones de clúster, seleccione los conceptos que desee excluir de la extracción.
2. Pulse el botón derecho del ratón para abrir el menú contextual.
3. Seleccione **Excluir de extracción**. El concepto se añade al diccionario de exclusión en Editor de recursos y el color de fondo del panel Resultados de extracción cambia, lo que indica que se debe volver a extraer para ver los cambios. Si tiene varios cambios, realícelos antes de volver a extraer.

Nota: Las palabras que excluya se almacenarán automáticamente en la primera biblioteca listada en el árbol de bibliotecas en Editor de recursos; de forma predeterminada, esta es la **Biblioteca local**.

Forzado de palabras en la extracción

Al revisar los datos de texto en el panel Datos después de la extracción, puede encontrar que algunas palabras o frases no se han extraído. A menudo, estas palabras son verbos o adjetivos en los que no está interesado. Sin embargo, algunas veces sí desea utilizar una palabra o frase que no se haya extraído como parte de una definición de categoría.

Si desea que esas palabras o frases se extraigan, puede forzar un término en una biblioteca de tipo. Para obtener más información, consulte el tema [“Forzar términos”](#) en la página 192.

Importante: Marcar un término en el diccionario como forzado no es infalible. Esto significa que aunque haya añadido un término explícitamente en un diccionario, es posible que no siempre esté presente en el panel Resultados extraídos después de haber repetido la extracción, o puede que aparezca pero no exactamente como lo especificó. Aunque esta circunstancia es rara, puede ocurrir cuando una palabra o frase ya se había extraído como parte de una frase más larga. Para evitarlo, aplique la opción **Completo (sin compuestos)** a este término en el diccionario de tipo. Para obtener más información, consulte el tema [“Añadir términos”](#) en la página 190.

Capítulo 9. Categorización de datos de texto

En la vista Categorías y conceptos , puede crear *categorías* que representen, en esencia, conceptos de nivel superior, o temas, que capturarán las ideas clave, el conocimiento y las actitudes expresadas en el texto.

Hasta la versión IBM SPSS Modeler Text Analytics 14, las categorías pueden tener una estructura jerárquica, lo que significa que pueden contener subcategorías y estas subcategorías pueden también tener sus propias subcategorías y así sucesivamente. Puede importar estructuras de categorías predefinidas, antes denominadas marcos de código, con categorías jerárquicas así como crear estas categorías jerárquicas en el producto.

De hecho, las categorías jerárquicas le permiten crear una estructura de árbol con una o más subcategorías para agrupar elementos como áreas de temas o conceptos diferentes de manera más precisa. Un ejemplo simple puede estar relacionado a actividades de tiempo libre; respondiendo una pregunta como *¿Qué actividad desearía hacer si tuviese más tiempo?* podría tener categorías principales como *deportes, arte y artesanía, pesca*, etc.; en un nivel más bajo, debajo de *deportes*, podría tener subcategorías para ver si esto es *juegos de pelota, acuáticos*, etc.

Las Categorías están formadas por un grupo de descriptores, como *conceptos, tipos, patrones y reglas de categorías*. Todos estos descriptores en conjunción se utilizan para identificar si un documento o registro pertenece a una categoría determinada. El texto de un documento o registro puede ser explorado para comprobar si existe texto que coincida con un descriptor. Si se halla una coincidencia, el documento o registro se asigna a dicha categoría. Este proceso se denomina *categorización*.

Puede trabajar con las categorías, generarlas y explorarlas visualmente mediante los datos que se presentan en los cuatro paneles de la vista Categorías y Conceptos, cada una de los cuales puede ocultarse o mostrarse seleccionando su nombre en el menú Ver.

- **Panel de categorías.** Genere y administre las categorías en este panel. Para obtener más información, consulte el tema [“El panel de categorías”](#) en la página 96.
- **Panel de resultados de extracción.** Explore y trabaje con los conceptos y tipos extraídos en este panel. Consulte el tema [“Resultados de extracción: Conceptos y tipos”](#) en la página 81 para obtener más información.
- **Panel de visualización.** Explore visualmente las categorías y compruebe su interacción en este panel. Consulte el tema [“Gráficos de categoría”](#) en la página 159 para obtener más información.
- **Panel de datos.** Explore y revise el texto que contienen los documentos y registros que corresponden a las selecciones en este panel. Consulte el tema [“El panel Datos”](#) en la página 103 para obtener más información.

Nota: Si hay más categorías que puedan caber en el panel visible, puede utilizar los controles de la parte inferior del panel para desplazarse hacia delante y hacia atrás por las categorías, o escriba un número de página para ir a la misma.

Tabla de árbol de categorías

La tabla de árbol de este panel presenta el conjunto de categorías, subcategorías y descriptores. El árbol dispone también de algunas columnas que presentan información para cada elemento de árbol. Las siguientes columnas pueden estar disponibles para su visualización:

- **Código** Muestra el valor del código para cada categoría. Esta columna está oculta de forma predeterminada. Puede visualizar esta columna a través de los menús: **Vista > Panel Categorías**.
- **Categoría**. Contiene el árbol de categorías con el nombre de la categoría y subcategorías. Además, si se pulsa el icono de la barra de herramientas de los descriptores, se mostrará el conjunto de descriptores.
- **Descriptores**. Proporciona el número de descriptores que componen su definición. Este recuento no incluye el número de descriptores en las subcategorías. No se proporciona un recuento cuando el nombre de un descriptor se muestra en la columna **Categorías**. Puede visualizar u ocultar los descriptores en el árbol a través de los menús: **Vista > Panel Categorías > Todos los descriptores**.
- **Docs** Después de la puntuación, esta columna proporciona el número de documentos o registros categorizados en una categoría y todas sus subcategorías. Por lo tanto, si 5 registros coinciden con su categoría principal en función de sus descriptores, y 7 registros distintos coinciden con una subcategoría en función de sus descriptores, el recuento de documentos total para la categoría principal es una suma de los dos; en este caso, sería 12. Sin embargo, si el mismo registro coincidía con la categoría principal y su subcategoría, el recuento sería 11.

Aunque no exista ninguna categoría, la tabla sigue conteniendo dos filas. La fila principal, denominada **Todos los documentos**, es el número total de documentos o registros. Una segunda fila, llamada **Sin categorizar**, muestra el número de documentos/registros que todavía deben categorizarse.

Para cada categoría del panel, hay un pequeño icono amarillo cuadrado delante del nombre de categoría. Si efectúa una doble pulsación en una categoría, o selecciona **Ver > Definiciones de categoría** en los menús, se abre el cuadro de diálogo Definiciones de categoría y presenta todos los elementos, denominados *descriptores*, que componen su definición, como conceptos, tipos, patrones y reglas de categoría. Para obtener más información, consulte el tema [“Acerca de las categorías”](#) en la página 102. De forma predeterminada, la tabla de árbol de categorías no muestra los descriptores en las categorías. Si desea ver los descriptores directamente en el árbol en lugar de en el cuadro de diálogo Definiciones de categorías, pulse en el botón de conmutación con el icono lápiz en la barra de herramientas. Cuando se selecciona el botón de conmutación, puede expandir su árbol para ver los descriptores.

Recuento de categorías

La columna **Docs** en la tabla de árbol de categorías muestra el número de documentos o registros que se clasifican en esa categoría específica. Si los números están fuera de la fecha de caducidad o no están calculados, aparece un icono en esa columna. Puede pulsar **Puntuar** en la barra de herramientas del panel para volver a calcular el número de documentos. Tenga en cuenta que el proceso de recuento puede tardar si trabaja con conjuntos de datos grandes.

Selección de categorías en el árbol

Cuando realice selecciones en el árbol, sólo puede seleccionar categorías hermanas; o sea, si selecciona las categorías de nivel superior, no puede seleccionar una subcategoría. O si selecciona 2 subcategorías de una categoría dada, no puede seleccionar simultáneamente una subcategoría de otra categoría. Seleccionar una categoría que no sea contigua generará la pérdida de la selección anterior.

Mostrar en los paneles de datos y visualización

Cuando selecciona una fila de la tabla, puede pulsar el botón **Mostrar** para actualizar los paneles de Visualización y de Datos con la información correspondiente a su selección. Si alguno de los paneles no se muestra, al pulsar **Mostrar** el panel aparecerá.

Refinamiento de las categorías

Es posible que la categorización no arroje un resultado perfecto para sus datos en el primer intento, y bien puede haber categorías que desee eliminar o combinar con otras categorías. También es posible, examinando los resultados de la extracción, que algunas categorías que no haya creado le resulten útiles. En tal caso, puede realizar cambios manuales en los resultados para ajustarlos a su contexto particular. Para obtener más información, consulte el tema [“Edición y refinamiento de categorías”](#) en la página 138.

Métodos y estrategias para crear categorías

Si todavía no ha realizado la extracción o sus resultados están desactualizados, el uso de una de estas técnicas de ampliación o generación de categorías le indicará la realización de una extracción automática. Una vez aplicada la técnica, los conceptos y tipos que estaban agrupados en una categoría siguen estando disponibles para la generación de categorías mediante otras técnicas. Esto significa que puede ver un concepto en varias categorías a menos que decida no reutilizarlos.

Para ayudarle a crear las mejores categorías, revise lo siguiente:

- **Métodos de creación de categorías**
- **Estrategias de creación de categorías**
- **Sugerencias de creación de categorías**

Métodos para crear categorías

Puesto que cada conjunto de datos es exclusivo, el número de métodos para crear categorías y el orden en el que los aplique puede cambiar con el tiempo. Además, puesto que los objetivos de la minería de textos puede diferir de un conjunto de datos a otro, puede que deba experimentar con diferentes métodos para comprobar con cuál de ellos se obtiene el mejor resultado para los datos de texto determinados. Ninguna de las técnicas automáticas categorizará perfectamente sus datos; por lo tanto, recomendamos buscar y aplicar una o más técnicas automáticas que funcionen correctamente con sus datos.

Aparte de utilizar paquetes de análisis de texto (TAP, *.tap) con conjuntos de categorías pregeneradas, también puede categorizar sus respuestas utilizando una combinación de los métodos siguientes:

- **Técnicas de creación automáticas.** Hay disponibles varias opciones de categorías basadas en frecuencia y basadas en lingüística para generar categorías automáticamente. Para obtener más información, consulte el tema [“Crear categorías”](#) en la página 106.
- **Técnicas de ampliación automáticas.** Hay disponibles varias técnicas lingüísticas para ampliar las categorías existentes añadiendo y mejorando los descriptores para que capturen más registros. Para obtener más información, consulte el tema [“Ampliación de categorías”](#) en la página 115.
- **Técnicas manuales.** Hay varios métodos manuales, como la función de arrastrar y soltar. Para obtener más información, consulte el tema [“Creación manual de categorías”](#) en la página 118.

Estrategias para crear categorías

La siguiente lista de estrategias no es exhaustiva en absoluto, pero puede proporcionarle algunas ideas sobre cómo afrontar la generación de las categorías.

- Cuando defina el nodo de minería de textos, seleccione un conjunto de categorías desde un paquete de análisis de texto (TAP) para comenzar su análisis con algunas categorías creadas previamente. Estas categorías pueden categorizar suficientemente el texto desde el primer momento. Sin embargo, si desea añadir más categorías, puede editar los valores de Generar categorías (**Categorías > Valores**

de compilación). Abra el cuadro de diálogo **Configuración avanzada: Lingüística**, seleccione la opción Entrada de categorías **Resultados de extracción no usados** y cree las categorías adicionales.

- Cuando defina el nodo, seleccione un conjunto de categorías desde un TAP en la vista Categorías y Conceptos en el área de trabajo interactiva. A continuación, arrastre los conceptos o patrones no usados y suéltelos en las categorías que considere apropiadas. A continuación, amplíe las categorías existentes que acaba de editar (**Categorías > Ampliar categorías**) para obtener más descriptores relacionados con los descriptores de categorías existentes.
- Crear categorías automáticamente utilizando los valores lingüísticos avanzados (**Categorías > Generar categorías**). A continuación, refine las categorías manualmente eliminando descriptores, eliminando categorías o fusionando categorías similares hasta que esté conforme con las categorías resultantes. Además, si generó categorías originalmente **sin** utilizar la opción **Generalizar con comodines cuando sea posible**, también puede intentar simplificar las categorías automáticamente utilizando Ampliar categorías con la opción **Generalizar**.
- Importe un archivo de categoría predefinida con anotaciones y/o nombres de categorías muy descriptivos. Además, si ha importado originalmente **sin** elegir la opción de importar o generar descriptores a partir de nombres de categoría, más adelante puede utilizar el diálogo Ampliar categorías y elegir **Ampliar categorías vacías con descriptores generados a partir del nombre de categoría**. A continuación, amplíe dichas categorías una segunda vez pero esta vez utilice las técnicas de agrupación.
- Cree manualmente un primer conjunto de categorías clasificando conceptos o patrones de conceptos por frecuencia, y luego arrastrando y soltando los más interesantes en el panel Categorías. Una vez que tenga ese conjunto inicial de categorías, utilice la función Ampliar (**Categorías > Ampliar categorías**) para expandir y refinar todas las categorías seleccionadas, de modo que incluyan otros descriptores relacionados y, por lo tanto, coincidan con más registros.

Después de aplicar estas técnicas, recomendamos que revise las categorías resultantes y emplee técnicas manuales para realizar pequeños ajustes, eliminar clasificaciones incorrectas o añadir registros o palabras que pueden haberse dejado de lado. Además, puesto que el uso de diferentes técnicas también puede generar categorías repetidas, puede fusionar o eliminar categorías si es necesario. Para obtener más información, consulte el tema [“Edición y refinamiento de categorías”](#) en la página 138.

Sugerencias para crear categorías

Para facilitar la creación de categorías mejores, puede revisar algunos conceptos que le ayudarán a tomar decisiones.

Consejos sobre la relación de "categoría a documento"

Las categorías a las que se asignan los documentos y los registros a menudo no son mutuamente excluyentes durante el análisis de texto cualitativo al menos por dos razones:

- En primer lugar, una regla empírica dice que cuanto más largo sea el documento o el registro de texto, más diversas son las ideas y las opiniones que se expresan. Así, las oportunidades de que puedan asignarse numerosas categorías a un documento o a un registro aumentan considerablemente.
- En segundo lugar, suele haber varias formas de agrupar e interpretar los documentos o registros de texto que no están separados de manera lógica. En el caso de una encuesta con preguntas abiertas sobre las tendencias políticas del encuestado, se pueden crear categorías, como *liberal* y *conservador* o *republicano* y *demócrata*, así como otras categorías más específicas, como *socio-liberal*, *económicamente conservador*, etc. Estas categorías no tienen por qué ser mutuamente excluyentes y exhaustivas.

Sugerencias sobre el número de categorías que conviene crear

La creación de categorías debe fluir directamente de los datos, a medida que vea algo interesante con respecto a sus datos, puede crear una categoría para representar esa información. En general, no existe un límite máximo recomendado sobre el número de categorías que pueden crearse. Sin embargo, es muy posible que se creen demasiadas categorías y que resulten inmanejables. Se aplican dos principios:

- **Frecuencia de categorías.** Para que una categoría sea útil, debe contener un número mínimo de documentos o registros. Uno o dos documentos pueden incluir algo muy interesante, pero si es uno o dos de 1000 documentos, la información que contienen puede no tener la frecuencia suficiente en la población para ser útil en la práctica.
- **Complejidad.** Cuantas más categorías cree, más información deberá leer y resumir después de completar el análisis. Sin embargo, el hecho de que haya demasiadas categorías, además de añadir complejidad, no aporta necesariamente detalles útiles.

Desgraciadamente, no existe ninguna regla que determine cuántas categorías se consideran demasiadas, ni para determinar el número mínimo de registros por categoría. Deberá tomar ese tipo de decisiones según las demandas de su situación particular.

Sin embargo, podemos ofrecerle consejos acerca de dónde comenzar. Aunque el número de categorías no debería ser excesiva, en los primeros estadios del análisis se recomienda pecar por exceso que por defecto. Agrupar categorías que sean relativamente similares resulta más fácil que dividir los casos en nuevas categorías, de manera que una estrategia de trabajo de pasar de más a menos categorías suele ser una buena técnica. Teniendo en cuenta la naturaleza repetitiva de la minería de textos y la facilidad con la que puede realizarse con un programa de software, se considera aceptable generar más categorías al comienzo.

Selección de los mejores descriptores

La siguiente información contiene varias directrices sobre cómo seleccionar o generar los mejores descriptores (conceptos, tipos, patrones TLA y reglas de categoría) para sus categorías. Los descriptores son los cimientos de las categorías. Cuando todo el texto de un documento o registro o parte del mismo coincide con un descriptor, el documento o registro se vincula con la categoría.

A menos que un descriptor contenga o se corresponda con un concepto o patrón extraído, no se vinculará con ningún documento o registro. Por lo tanto, utilice conceptos, tipos, patrones y reglas de categoría como se describe en las siguientes secciones.

Como los conceptos no sólo se representan a sí mismos sino que también representan un conjunto de términos subyacentes que pueden ir desde formas en singular/plural hasta sinónimos, pasando por variaciones ortográficas, sólo se debe utilizar el propio concepto como descriptor o como parte de un descriptor. Para obtener más información sobre los términos subyacentes de cualquier concepto específico, pulse en el nombre del concepto en el panel Resultados extraídos de la vista Conceptos y Categorías. Cuando pase el ratón por encima del nombre del concepto, aparecerá una etiqueta con información y se mostrarán los términos subyacentes encontrados en su texto durante la última extracción. No todos los conceptos tienen términos subyacentes. Por ejemplo, si `car` y `vehicle` eran sinónimos pero `car` se ha extraído como el concepto con `vehicle` como término subyacente, solo desea utilizar `car` en un descriptor ya que coincidirá automáticamente con documentos o registros con `vehicle`.

Conceptos y tipos como descriptores

Utilice un concepto como descriptor cuando quiera encontrar todos los documentos o registros que contengan dicho concepto (o cualquiera de sus términos subyacentes). En este caso, el uso de una regla de categoría más compleja no es necesario debido a que el nombre de concepto exacto es suficiente. Recuerde que cuando utiliza recursos que extraen opiniones, a veces los conceptos pueden variar durante la extracción de patrón TLA para capturar el verdadero sentido de la frase (consulte el ejemplo de la sección siguiente sobre TLA).

Por ejemplo, una respuesta de encuesta que indique las frutas favoritas de cada persona como, por ejemplo, *“La manzana y la piña son las mejores”* puede dar como resultado la extracción de `apple` (manzana) y `pineapple` (piña). Al añadir el concepto `apple` como descriptor a su categoría, todas las respuestas que contengan el concepto `apple` (o cualquiera de sus términos subyacentes) coinciden con esa categoría.

Sin embargo, si está interesado en saber simplemente qué respuestas mencionan *manzana* de alguna manera, puede escribir una regla de categoría como `* apple *` y también capturará respuestas que contengan conceptos como `apple`, `apple sauce` o `french apple tart`.

También puede capturar todos los documentos o registros que contienen conceptos que se han escrito de la misma manera utilizando un tipo como descriptor directamente como <Fruit>. Recuerde que no puede utilizar * con tipos.

Consulte el tema [“Resultados de extracción: Conceptos y tipos”](#) en la página 81 para obtener más información.

Descriptores y patrones TLA (Análisis de enlace de texto)

Utilice un resultado de patrón TLA como descriptor cuando quiera capturar ideas más finas con más matices. Cuando el texto se analiza durante la extracción de TLA, el texto se procesa frase a frase o cláusula a cláusula en lugar de tomar el texto completo (el documento o registro). Al considerar todas las partes de una frase en conjunto, TLA puede identificar opiniones, relaciones entre dos elementos o, por ejemplo, una negación y comprender el sentido verdadero. Puede utilizar patrones de conceptos o patrones de tipo como descriptores. Para obtener más información, consulte el tema [“Patrones de tipo y concepto”](#) en la página 151.

Por ejemplo, si tuviéramos el texto *"la sala no estaba tan limpia"*, podrían extraerse los siguientes conceptos: room (sala) y clean (limpia). Sin embargo, si la extracción de TLA se ha habilitado en el valor de extracción, TLA podría detectar que clean se ha utilizado de forma negativa y corresponde realmente a not clean, que es un sinónimo del concepto dirty. Aquí, puede ver que el uso del concepto clean como un descriptor por sí mismo coincidiría con este texto, pero también podría capturar otros documentos o registros que mencionan la limpieza. Por lo tanto, podría ser mejor utilizar el patrón de concepto de TLA con dirty como concepto de salida ya que coincidiría con este texto y probablemente sea un descriptor más apropiado.

Reglas empresariales de categoría como descriptores

Las reglas de categoría son sentencias que clasifican automáticamente documentos o registros en una categoría que se basa en una expresión lógica utilizando los conceptos, tipos y patrones extraídos, así como operadores booleanos. Por ejemplo, podría escribir una expresión que significara *incluir todos los registros que contienen el concepto extraído embassy pero no argentina en esta categoría*.

Puede escribir y utilizar reglas de categoría como descriptores en sus categorías para expresar varias ideas distintas utilizando &, | y ! () booleanos. Para obtener información detallada sobre la sintaxis de estas reglas y sobre cómo escribirlas y editarlas, consulte [“Uso de reglas de categoría”](#) en la página 119.

- Utilice una regla de categoría con el operador booleano & (AND) para buscar documentos o registros en los que se produzcan 2 o más conceptos. Los 2 o más conceptos conectados por operadores & no tienen que aparecer en la misma frase u oración; pueden hacerlo en cualquier parte del mismo documento o registro para que se les considere una coincidencia con la categoría. Por ejemplo, si crea la regla de categoría food & cheap como descriptor, coincidiría con un registro que contiene el texto, *"la comida era bastante cara, pero las habitaciones eran baratas"*, a pesar del hecho de que food no era el nombre que se llamaba cheap ya que el texto contenía tanto food como cheap.
- Utilice una regla de categoría con el operador booleano ! () (NOT) como descriptor para buscar documentos o registros en los que se produzcan algunas cosas y otras no. Esto puede evitar la agrupación de información que puede parecer que está relacionada basándose en las palabras pero no en el contexto. Por ejemplo, si crea la regla de categoría <Organization> & !(ibm) como descriptor, coincidiría con el texto siguiente *SPSS Inc. fue una empresa fundada en 1967* y no coincidiría con el texto siguiente *la compañía de software fue adquirida por IBM*.
- Utilice una regla de categoría con el operador booleano | (OR) como descriptor para buscar documentos o registros que contengan uno de varios conceptos o tipos. Por ejemplo, si crea la regla de categoría (personnel|staff|team|coworkers) & bad como descriptor, coincidiría con cualquier documento o registro en el que se encuentre alguno de esos nombres con el concepto bad.
- Utilice tipos en reglas de categoría para hacer que sean más genéricas y puedan implementarse en más casos. Por ejemplo, si está trabajando con datos de hotel, puede que esté muy interesado en saber lo que piensan los clientes sobre el personal del hotel. Los términos relacionados podrían incluir palabras como recepcionista, camarero, camarera, recepción, mostrador de recepción, etc. En este caso, puede crear un tipo nuevo llamado <HotelStaff> y añadir todos los términos anteriores a ese tipo. Aunque es posible crear una regla de categoría para todo tipo de personal como [* waitress * & nice], [* desk * & friendly], [* receptionist * & accommodating], puede crear una regla de

categoría única y más genérica utilizando el tipo <HotelStaff> para capturar todas las respuestas que tengan opiniones favorables del personal del hotel en forma de [<HotelStaff> & <Positive>].

Nota: Puede utilizar + y & en reglas de categoría al incluir patrones de TLA en esas reglas. Para obtener más información, consulte el tema [“Uso de patrones TLA en las reglas de categoría”](#) en la página 122.

Ejemplo de cómo los conceptos, TLA o reglas de categoría como descriptores coinciden de manera diferente

El siguiente ejemplo muestra el modo en que el uso de un concepto como descriptor, regla de categoría como descriptor o patrón TLA como descriptor afecta a cómo se categorizan los documentos o registros. Supongamos que tenía los siguientes 5 registros.

- A: *"excelente personal del restaurante, excelente comida y habitaciones cómodas y limpias."*
- B: *"el personal del restaurante fue horrible, pero las habitaciones estaban limpias."*
- C: *"Habitaciones cómodas y limpias."*
- D: *"Mi habitación no estaba tan limpia."*
- E: *"Limpia."*

Como los registros incluyen la palabra *limpia* y desea capturar esta información, podría crear uno de los descriptores que aparecen en la tabla siguiente. Basándose en lo esencial de lo que esté intentando capturar, podrá ver cómo el uso de una clase de descriptor antes que otra puede producir resultados diferentes.

Tabla 17. Modo de coincidencia entre descriptores y registros del ejemplo						
Descriptor	A	B	C	D	E	Descripción
clean	<i>coincide</i>	<i>coincide</i>	<i>coincide</i>	<i>coincide</i>	<i>coincide</i>	El descriptor es un concepto extraído. Cada registro contenía el concepto clean, incluso el registro D, ya que sin TLA, las reglas de TLA no saben automáticamente que <i>"not clean"</i> significa dirty.
clean + .	-	-	-	-	<i>coincide</i>	El descriptor es un patrón de TLA que representa clean por sí mismo. Solo se ha emparejado el registro donde clean se ha extraído sin ningún concepto asociado durante la extracción de TLA.
[clean]	<i>coincide</i>	<i>coincide</i>	<i>coincide</i>	-	<i>coincide</i>	El descriptor es una regla de categoría que busca una regla TLA que contiene clean por sí mismo o con algo más. Coincidió con todos los registros en los que se ha encontrado una salida de TLA que contiene clean, independientemente de si clean estaba enlazado a otro concepto como room y en cualquier posición de ranura.

Acerca de las categorías

Categorías hace referencia a un grupo de conceptos, opiniones o actitudes estrechamente relacionados. Para que resulte útil, una categoría debe describirse fácilmente mediante una breve frase o etiqueta que capte su significado esencial.

Por ejemplo, si está analizando las respuestas de los usuarios en una encuesta sobre un nuevo detergente, puede crear una categoría etiquetada como *olor* que contenga todas las respuestas que describan el aroma del producto. Sin embargo, una categoría así no diferenciaría entre aquellos que consideraron que el detergente tenía un aroma agradable de los que lo encontraron molesto. Puesto que IBM SPSS Modeler Text Analytics es capaz de extraer opiniones cuando se utilizan los recursos apropiados, puede crear otras dos categorías para identificar encuestados que *disfrutaron del olor* y encuestados a los que *no les gustó el olor*.

Puede crear y trabajar con sus categorías en el panel Categorías en el panel superior izquierdo de la ventana vista Categorías y conceptos. Cada categoría está definida por uno o más descriptores. Los **Descriptores** son conceptos, tipos y patrones, así como reglas de categoría, que se han utilizado para definir una categoría.

Si quiere ver los descriptores que conforman una categoría específica, puede pulsar en el icono de lápiz de la barra de herramientas del panel Categorías y, a continuación, expandir el árbol para ver los descriptores. De forma alternativa, seleccione la categoría y abra el recuadro de diálogo Definiciones de categoría (**Ver > Definiciones de categoría**).

Cuando crea categorías automáticamente utilizando técnicas de creación de categorías como la inclusión de conceptos, las técnicas utilizarán conceptos y tipos como descriptores para crear sus categorías. Si extrae patrones de TLA, también puede añadir patrones o partes de estos patrones como descriptores de categorías. Para obtener más información, consulte el tema [Capítulo 11, “Exploración del Análisis de enlaces de texto”](#), en la página 149. Y si crea clústeres, puede añadir los conceptos de un clúster a categorías nuevas o existentes. Por último, puede crear reglas de categoría manualmente para utilizarlas como descriptores en las categorías. Para obtener más información, consulte el tema [“Uso de reglas de categoría”](#) en la página 119.

Propiedades de categoría

Además de los descriptores, las categorías también tienen propiedades que pueden editarse para cambiar el nombre de las categorías, añadir una etiqueta o anotación.

Existen las propiedades siguientes:

- **Nombre.** Este nombre aparece en el árbol de forma predeterminado. Cuando se crea una categoría utilizando una técnica automática, se le asigna un nombre automáticamente.
- **Etiqueta.** El uso de etiquetas resulta útil para crear descripciones de categoría con un significado más detallado que se utilizan en otros productos o en otras tablas o gráficos. Si elige la opción de visualizar la etiqueta, ésta se utilizará en la interfaz para identificar la categoría.
- **Código.** El número de código se corresponde con el valor de código de esta categoría.
- **Anotación.** Puede añadir una descripción breve a cada categoría de este campo. Cuando el diálogo Generar categorías genera una categoría, automáticamente se añade una nota a esta anotación. También puede añadir texto de ejemplo a una anotación directamente desde el panel Datos seleccionando el texto y eligiendo **Categorías > Añadir a anotación** en los menús.

El panel Datos

A medida que crea las categorías, en ocasiones puede necesitar revisar parte de los datos del texto con los que está trabajando. Por ejemplo, si crea una categoría en la que hay 640 documentos categorizados, es posible que desee examinar algunos o todos esos documentos para ver qué texto fue escrito en realidad. Puede revisar registros o documentos en el panel Datos, que se encuentra en la parte inferior derecha. Si no está visible de forma predeterminada, seleccione **Ver > Paneles > Datos** en los menús.

El panel Datos representa una fila por documento o registro correspondiente a la selección en el panel de Categorías, el panel de Resultados de extracción, o el cuadro de diálogo Definiciones de categoría hasta un cierto límite de visualización. De forma predeterminada, el número de documentos o registros que se muestra en el panel Datos está limitado para que pueda ver los datos más rápidamente. Sin embargo, puede ajustar esto en el cuadro de diálogo Opciones. Si trata con conjuntos de datos muy grandes,

la velocidad de visualización puede mejorarse, desactivando la opción de mostrar las categorías. Para obtener más información, consulte el tema [“Opciones: separador Sesión”](#) en la página 75.

Nota: Si hay más registros que puedan caber en el panel visible, puede utilizar los controles de la parte inferior del panel para desplazarse hacia delante y hacia atrás por los registros, o escriba un número de página para ir a la misma.

Visualización y renovación del panel Datos

El panel Datos no renueva su visualización automáticamente dado que con conjuntos de datos más grande, la renovación automática de datos podría tardar en completarse. Por lo tanto, siempre que realice una selección en otro panel en esta vista o en el cuadro de diálogo Definiciones de categoría, pulse

Desplegar para renovar el contenido del panel Datos.

Documentos o registros de texto

Si los datos de texto se encuentran en forma de registros y el texto es relativamente corto en longitud, el campo de texto en el panel Datos muestra los datos de texto en su totalidad. Sin embargo, al trabajar con registros y grandes conjuntos de datos, la columna de campo de texto muestra un breve fragmento del texto y abre un panel Vista previa de texto a la derecha para mostrar más o todo el texto del registro que ha seleccionado en la tabla. Si los datos de texto están en forma de documentos individuales, el panel Datos muestra el nombre de archivo del documento. Cuando selecciona un documento, se abre el panel de Vista previa con el texto del documento seleccionado.

Colores y resaltado

Siempre que se visualizan los datos, los conceptos y descriptores que se encuentran en esos documentos o registros se resaltan en color para ayudarlo a identificarlos fácilmente en el texto. La codificación por colores corresponde a los tipos a los que pertenecen los conceptos. También puede pasar el ratón por encima de los elementos con codificación por color para mostrar el concepto bajo el que se extrajo y el tipo al que se asignó. El texto que no se ha extraído aparece en negro. Generalmente, estas palabras no extraídas suelen ser conectores (y o *con*), pronombres (*mi* o *ellos*), y verbos (*es*, *tiene* o *tomar*).

Columnas del panel Datos

Mientras que la columna de campo de texto siempre está visible, también puede mostrar otras columnas. Para visualizar otras columnas, seleccione **Ver > Panel Datos** en los menús y, a continuación, seleccione la columna que desea visualizar en el panel Datos. Las siguientes columnas pueden estar disponibles para su visualización:

- **"Nombre de campo de texto" (#)/Documentos.** Añade una columna para los datos de texto desde los que se extrajeron los conceptos y el tipo. Si los datos están en documentos, la columna se denomina Documentos y sólo el nombre de archivo del documento o la vía de acceso completa es visible. Para ver el texto de dichos documentos debe buscar en el panel Vista previa de texto. El número de filas en el panel Datos se muestra entre paréntesis después de este nombre de columna. Puede haber ocasiones en que no todos los documentos o registros se muestran debido a un límite en el diálogo Opciones utilizado para aumentar la velocidad de carga. Si se alcanza el máximo, el número será seguido de **-Max**. Si desea más información, consulte [“Opciones: separador Sesión”](#) en la página 75.
- **Categorías.** Enumera cada una de las categorías a las que pertenece un registro. Cuando se muestra esta columna, la actualización del panel Datos puede tardar un poco más hasta mostrar la información más actualizada.
- **Forzar dentro.** Lista las categorías en las cuales ha forzado. Puede forzarse la inclusión de documentos en una categoría mediante la selección de menús **Editar > Forzar dentro**. Si desea más información, consulte [“Forzar los documentos de las categorías”](#) en la página 140.
- **Forzar fuera.** Lista las categorías de las cuales ha eliminado un documento. Puede forzarse la exclusión de documentos de una categoría mediante la selección de menús **Editar > Forzar fuera**. Por ejemplo, esto podría utilizarse cuando el sarcasmo de un encuestado hace que una respuesta caiga en la categoría incorrecta. Si desea más información, consulte [“Forzar los documentos de las categorías”](#) en la página 140.
- **Recuentos de categorías.** Lista el número de las categorías a las que pertenece un registro.

- **Rangos de relevancia.** Proporciona un orden para cada registro de una categoría. Este orden muestra el grado de adecuación del registro en la categoría si se compara con otros registros de la misma categoría. Para ver el rango, seleccione una categoría del panel Categorías (panel superior izquierdo). Si desea más información, consulte [“Relevancia de categoría”](#) en la página 105.
- **Marcas de respuesta.** Añada una columna que muestra las marcas que puedan estar utilizándose. Pulse dentro de la columna para cambiar el tipo de marca que asigna a los documentos. Podría marcar los documentos con la marca "completo" o la marca "importante", o bien eliminar las marcas. Esto es útil para revisar la precisión de un modelo de categoría. Si desea más información, consulte [“Marcar respuestas”](#) en la página 105.

Relevancia de categoría

Para mejorar la generación de categorías, puede revisar la relevancia de los documentos o registros de cada categoría, así como la relevancia de todas las categorías a las que pertenece un documento o registro.

Relevancia de una categoría en un registro

Cuando aparece un documento o registro en el panel Datos, todas las categorías a las que pertenece se muestran en una lista en la columna Categorías. Cuando un documento o registro pertenece a varias categorías, las categorías de esta columna aparecen en orden de mayor a menor relevancia. Se considera que la categoría que se muestra en primer lugar es la que mejor se corresponde con este documento o registro. Consulte el tema [“El panel Datos”](#) en la página 103 para obtener más información.

Relevancia de un registro en un categoría

Cuando selecciona una categoría, puede revisar la relevancia de cada uno de sus registros en la columna Rango de relevancia en el panel Datos. Este rango de relevancia indica el grado de idoneidad con el que el documento o registro se adapta a la categoría seleccionada si se compara con el resto de los registros de dicha categoría. Para ver la ordenación de los registros de una sola categoría, seleccione la categoría en el panel Categorías (situado arriba a la izquierda) y la ordenación del documento o registro aparece en la columna. Esta columna no puede verse de manera predeterminada, pero puede seleccionar la opción para mostrarla. Consulte el tema [“El panel Datos”](#) en la página 103 para obtener más información.

Cuanto más bajo es el número de la ordenación de relevancia del registro, mejor es el ajuste o más relevancia tiene este registro para la categoría seleccionada, de modo que 1 es el mejor. Si hay más de un registro con la misma relevancia, cada uno aparece con el mismo grado de ordenación seguido de un signo igual (=) para indicar que tienen la misma relevancia. Por ejemplo, puede tener los valores de ordenación siguientes 1=, 1=, 3, 4, etc.; significa que hay dos registros que se consideran como mejor opción de relación para esta categoría.

Sugerencia: Puede añadir el texto del registro más relevante a la anotación de categoría para ayudar a proporcionar una mejor descripción de la categoría. Añada el texto directamente desde el panel Datos seleccionando el texto y eligiendo **Categorías > Añadir a anotación** en los menús.

Marcar respuestas

Para ayudarle en su trabajo, puede marcar los documentos con marcas (distintivos) en el panel Datos. esta característica solo está disponible si el documento de origen contiene un ID exclusivo. Si el documento de origen no contiene un ID exclusivo, puede añadir un nodo Derivar entre el documento de origen y el nodo Text Mining.

Existen muchas razones por las cuales podría desear marcar un documento, entre ellas:

- Para marcar documentos que ha revisado manualmente, de forma que sepa donde reanudar más adelante.
- Para marcar un documento que no sabe muy bien cómo manejar

Una vez ha señalado un documento con una marca, puede continuar trabajando con el mismo. Las marcas solo tienen fines de registro. Puede elegir entre las siguientes marcas:

Tabla 18. Descripciones de las marcas

Marca	Descripción
	La marca de Completo señala los documentos que considera acabados.
	La marca de Importante señala los documentos que considera importantes.

Para marcar un documento con una marca:

1. Desde el panel Datos, pulse con el botón derecho del ratón sobre el documento que desea marcar.
2. Desde el menú de contexto, seleccione **Vista > Panel Datos > Marcas de respuesta** y seleccione a continuación el tipo de marca que desea utilizar (marca de Importante o Completo). Se asignará la marca seleccionada. Si la columna Marca del panel Datos no estaba visible, ahora aparecerá.

Para borrar las marcas:

1. Desde el panel Datos, pulse con el botón derecho del ratón sobre los documentos de los que desea eliminar la marca.
2. Desde el menú de contexto, seleccione **Marcar respuestas con > Borrar marcas**. Las marcas seleccionadas se eliminan.

Crear categorías

Por un lado puede tener categorías procedentes de un paquete de análisis de texto, y también puede generar categorías automáticamente utilizando una serie de técnicas lingüísticas y de frecuencia. A través del recuadro de diálogo Valores de creación de categorías, puede aplicar las técnicas automáticas lingüísticas y de frecuencia para producir categorías a partir de conceptos o de patrones de conceptos.

En general, las categorías pueden estar formadas por diferentes tipos de descriptores (tipos, conceptos, patrones TLA, reglas de categoría). Cuando se generan categorías utilizando las técnicas automáticas de generación de categorías, las categorías resultantes se indican después de un concepto o de un patrón de conceptos (según la entrada que haya seleccionado) y contiene un conjunto de descriptores. Estos descriptores pueden presentarse en forma de reglas de categoría o conceptos e incluyen todos los conceptos relacionados descubiertos por las técnicas.

Después de generar las categorías, puede aprender mucho sobre ellas revisándolas en el panel Categorías o explorándolas en gráficos y diagramas. Luego puede utilizar técnicas manuales para realizar pequeños ajustes, eliminar clasificaciones incorrectas o añadir registros o palabras que pueden haberse dejado de lado. Una vez aplicada la técnica, los conceptos, tipos y patrones que estaban agrupados en una categoría siguen estando disponibles para otras técnicas. Y puesto que el uso de diferentes técnicas también puede generar categorías repetidas o inapropiadas, puede fusionar o eliminar categorías. Para obtener más información, consulte el tema [“Edición y refinamiento de categorías”](#) en la página 138.

Importante: En versiones anteriores, las reglas de coocurrencia y de sinónimos se colocaban entre corchetes. En esta versión, los corchetes ahora indican un resultado de patrón de análisis de enlace de texto. En su lugar, las reglas de coocurrencia y de sinónimos se encapsularán entre paréntesis, como (speaker systems|speakers).

Para crear categorías

1. En los menús, seleccione **Categorías > Crear categorías**. A menos que haya elegido que no se le solicite nunca, se muestra un recuadro de mensaje.
2. Elija si desea generar ahora o editar primero la configuración.

- Pulse **Generar ahora** para empezar a generar categorías utilizando la configuración actual. La configuración seleccionada de manera predeterminada suele ser suficiente para comenzar el proceso de categorización. El proceso de generación de categorías comienza y aparece un diálogo de progreso.
- Pulse **Editar** para revisar y modificar la configuración de la generación.

Nota: El número máximo de categorías que se pueden mostrar es 10.000. Se muestra un aviso si se alcanza o supera este número. Si esto sucede, debe cambiar las opciones de creación o ampliación de categorías para reducir el número de categorías creadas.

Entradas

Las categorías se generan a partir de descriptores derivados de cada tipo o patrón de tipos. En esta tabla, puede seleccionar los tipos individuales o los patrones que se incluirán en el proceso de generación de categorías.

Patrones de tipo. Si selecciona patrones de tipo, las categorías se generarán a partir de patrones en lugar de tipos y conceptos individuales. De esta forma se categorizarán los registros o documentos que contengan un patrón de conceptos que pertenezca al patrón de tipo seleccionado. Por lo tanto, si selecciona el patrón de tipo <Budget> y <Positive> en la tabla, podrían generarse categorías como *cost & <Positive>* o *rates & excellent*.

Al utilizar patrones de tipo como entrada para la generación de categorías automática, a veces las técnicas identifican varias maneras de formar la estructura de la categoría. Técnicamente, no hay una única manera correcta de producir las categorías; sin embargo, puede que encuentre una estructura más adecuada para su análisis que otra. Para ayudar a personalizar la salida en este caso, puede designar un tipo como el foco preferido. Todas las categorías de nivel superior producidas provendrán de un concepto del tipo que seleccione aquí (y de ningún otro tipo). Cada subcategoría contendrá un patrón de enlace de texto de este tipo. Seleccione este tipo en el campo **Estructurar categorías por tipo de patrón:** y la tabla se actualizará para mostrar sólo los patrones aplicables que contengan el tipo seleccionado. Lo más frecuente es que <Unknown> ya esté preseleccionado para usted. El resultado es que se seleccionan todos los patrones que contienen el tipo <Unknown> . La tabla muestra los tipos en orden descendente empezando por el que tiene el mayor número de registros o documentos (recuento de **Doc.**).

Tipos. Si selecciona tipos, las categorías se generarán a partir de los conceptos que pertenecen a los tipos seleccionados. Por lo tanto, si selecciona el tipo <Budget> en la tabla, pueden generarse categorías como *cost* o *price*, ya que *cost* y *price* son conceptos asignados al tipo <Budget>.

De manera predeterminada, sólo se seleccionan los tipos que capturan el número máximo de registros o de documentos. Esta preselección permite centrarse rápidamente en los tipos más interesantes y evitar la generación de categorías irrelevantes. En la tabla se muestran los tipos en orden descendente, comenzando por el que tiene el mayor número de registros o documentos (recuento de **Doc.**). Los tipos de la biblioteca *Opinions* no están seleccionados de manera predeterminada en la tabla de tipos.

La entrada que ha seleccionado afecta a las categorías que obtiene. Cuando decida utilizar Tipos como entrada, podrá ver los conceptos relacionados claramente con más facilidad. Por ejemplo, si crea categorías utilizando Tipos como entrada, puede obtener una categoría *Fruit* con conceptos como *apple*, *pear*, *citrus fruits*, *orange*, etc. Si elige Patrones de tipo como entrada en su lugar y selecciona el patrón <Unknown> + <Positive>, por ejemplo, puede obtener una categoría *fruit + <Positive>* con uno o dos tipos de fruta, como *fruit + tasty* y *apple + good*. Este segundo resultado sólo muestra 2 patrones de conceptos porque las otras apariciones de frutas no tienen por qué haberse calificado positivamente. Y aunque esto puede ser suficientemente bueno para los datos de texto actuales, en los estudios longitudinales en los que se utilizan distintos conjuntos de documentos, es posible que desee añadir manualmente otros descriptores, como *citrus fruit + positive* o tipos de uso. Si sólo utiliza tipos como entrada, esto le ayudará a encontrar toda la fruta posible.

Técnicas

Puesto que cada conjunto de datos es exclusivo, el número de métodos y el orden en el que los aplique puede cambiar con el tiempo. Puesto que los objetivos de la minería de textos puede diferir de un

conjunto de datos a otro, puede que deba experimentar con diferentes técnicas para comprobar con cuál de ellas se obtiene el mejor resultado para los datos de texto determinados.

No es necesario que sea un experto en este tipo de configuración para poder utilizarla. De forma predeterminada, los valores más comunes ya aparecen seleccionados. Por lo tanto, puede pasar por alto los diálogos de configuración avanzada y empezar a generar categorías directamente. Igualmente, si realiza cambios aquí, no es necesario que vuelva al diálogo de configuración cada vez, puesto que siempre se conservan los últimos valores.

Seleccione las técnicas lingüísticas o de frecuencia y pulse en el botón Configuración avanzada para que se muestren los valores de las técnicas seleccionadas. Ninguna de las técnicas automáticas categorizará perfectamente sus datos; por lo tanto, recomendamos buscar y aplicar una o más técnicas automáticas que funcionen correctamente con sus datos. No se pueden utilizar técnicas lingüísticas y de frecuencia simultáneamente para el proceso de creación.

- **Técnicas lingüísticas avanzadas.** Para obtener más información, consulte [“Configuración lingüística avanzada”](#) en la página 108.
- **Técnicas de frecuencia avanzadas.** Para obtener más información, consulte [“Configuración avanzada de frecuencia”](#) en la página 114.

Configuración lingüística avanzada

Al crear categorías, puede seleccionar entre una serie de técnicas de creación de categorías lingüísticas avanzadas como, por ejemplo, *inclusión de concepto* y *redes semánticas* (solo texto en inglés). Estas técnicas pueden utilizarse solas o combinadas entre sí para crear categorías.

Tenga en cuenta que, puesto que cada conjunto de datos es exclusivo, el número de métodos y el orden en el que los aplique puede cambiar con el tiempo. Puesto que los objetivos de la minería de textos puede diferir de un conjunto de datos a otro, puede que deba experimentar con diferentes técnicas para comprobar con cuál de ellas se obtiene el mejor resultado para los datos de texto determinados. Ninguna de las técnicas automáticas categorizará perfectamente sus datos; por lo tanto, recomendamos buscar y aplicar una o más técnicas automáticas que funcionen correctamente con sus datos.

Los campos y áreas siguientes están disponibles dentro del recuadro de diálogo Valores avanzados: Lingüística:

Entrada y salida

Entrada de categoría Seleccione a partir de qué se generarán las categorías:

- **Resultados de extracción no usados.** Esta opción permite que se generen categorías a partir de los resultados de extracción que no se utilizan en ninguna categoría existente. De esta manera se minimiza la tendencia que tienen los registros de hacer coincidir varias categorías y limita el número de categorías que se generan.
- **Todos los resultados de la extracción.** Esta opción permite generar categorías utilizando los resultados de extracción. Esto resulta especialmente útil cuando no existen categorías, o existen muy pocas.

Salida de categoría Seleccione la estructura general de las categorías que se generarán:

- **Jerárquico con subcategorías.** Esta opción permite la creación de subcategorías y sub-subcategorías. Puede establecer la profundidad de sus categorías seleccionando el número máximo de niveles (campo **Niveles máximos creados**) que se puede crear. Si selecciona 3, las categorías podrían contener subcategorías y dichas subcategorías también podrían tener subcategorías.
- **Categorías planas (nivel único).** Esta opción sólo permite la creación de un nivel de categorías, lo que significa que no se generará ninguna subcategoría.

Agrupación de técnicas

Cada una de las técnicas disponibles resulta idónea para determinados tipos de datos y situaciones, pero a menudo conviene combinar técnicas en el mismo análisis para capturar el rango completo de documentos o registros. Puede ver un concepto en diversas categorías o detectar categorías redundantes.

Inclusión de conceptos. Esta técnica genera categorías agrupando los conceptos multitérmino (palabras compuestas) basándose en si contienen palabras que son subconjuntos o superconjuntos de una palabra en la otra. Por ejemplo, el concepto `seat` se agruparía con `safety seat`, `seat belt` y `seat belt buckle`. Para obtener más información, consulte el tema [“Inclusión de conceptos”](#) en la página 112.

Red semántica. Esta técnica comienza identificando los posibles sentidos de cada concepto a partir de un amplio índice de relaciones de palabras, y luego crea categorías agrupando los conceptos relacionados. Esta técnica resulta idónea cuando los conceptos son conocidos en la red semántica y no son muy ambiguos. Es menos idónea si el texto contiene terminología específica o jerga desconocida en la red. En un ejemplo, el concepto `granny smith apple` se podría agrupar con `gala apple` y `winesap apple` ya que son hermanos de Granny Smith. En otro ejemplo, el concepto `animal` podría agruparse con `cat` y `kangaroo` ya que son hipónimos de `animal`. En esta versión esta técnica está disponible solo para texto en inglés. Para obtener más información, consulte el tema [“Redes semánticas”](#) en la página 112.

Nota: La opción **Distancia máxima de búsqueda** sólo está disponible si selecciona la opción **Red semántica**.

Distancia máxima de búsqueda Seleccione el alcance de búsqueda de las técnicas antes de generar las categorías. Cuanto más bajo sea el valor, menor será la cantidad de resultados obtenidos; sin embargo, estos resultados serán menos ruidosos y más propensos a ser enlazados o asociados significativamente uno con el otro. Cuando más alto sea el valor, mayor será la cantidad de resultados obtenidos; sin embargo, estos resultados pueden ser menos fiables o relevantes. Esta opción se aplica globalmente en todas las técnicas, pero sus resultados son mejores en las co-ocurrencias y en las redes semánticas.

Impedir el emparejamiento de conceptos específicos. Seleccione esta casilla de verificación para detener el proceso de agrupación o emparejamiento de dos conceptos en la salida. Para crear o gestionar pares de conceptos, pulse **Gestionar pares**. Para obtener más información, consulte el tema [“Administración de pares de excepciones de enlace”](#) en la página 110.

Generalizar con comodines cuando sea posible Seleccione esta opción para que el producto genere reglas genéricas en las categorías utilizando el comodín de asterisco. Por ejemplo, en lugar de producir varios descriptores como `[apple tart + .]` y `[apple sauce + .]`, utilizar comodines puede producir `[apple * + .]`. Si generaliza con comodines, obtendrá a menudo exactamente el mismo número de registros o documentos que antes. Sin embargo, esta opción tiene la ventaja de reducir el número y simplificar los descriptores de categoría. Además, esta opción aumenta la posibilidad de categorizar más registros o documentos utilizando estas categorías en datos de texto nuevos (por ejemplo, en estudios por fases o longitudinales).

Otras opciones para generar categorías

Además de seleccionar las técnicas de agrupación que se aplicarán, puede editar otras muchas opciones de generación, como se indica a continuación:

Número máximo de categorías de nivel superior creadas. Utilice esta opción para limitar el número de categorías que pueden generarse cuando pulsa en el botón Generar categorías. En algunos casos, puede obtener mejores resultados si establece un valor alto y luego suprime cualquiera de las categorías irrelevantes.

Número mínimo de descriptores y/o subcategorías por categoría. Utilice esta opción para definir el número mínimo de descriptores y subcategorías que debe contener una categoría para que pueda crearse. Esta opción ayuda a limitar la creación de categorías que no capturen un número significativo de registros o documentos.

Permitir a los descriptores aparecer en más de una categoría Cuando está seleccionada, esta opción permite el uso de descriptores en más de una de las categorías que se generarán a continuación. Esta opción suele generarse porque los elementos se clasifican de forma natural o habitual en dos o más categorías, y dejar que esto ocurra conlleva la creación de categorías de mayor calidad. Si no selecciona esta opción, se reduce el solapamiento de registros en varias categorías, y en función del tipo de datos que tenga, puede ser una situación deseable. Sin embargo, con la mayoría de los tipos de datos, el hecho de restringir los descriptores a una sola categoría suele resultar en una pérdida de calidad o de cobertura de la categoría. Por ejemplo, digamos que usted tenía el concepto `car seat manufacturer`. Con esta opción, este concepto podría aparecer en una categoría basada en el texto `car seat` y en otra basada

en `manufacturer`. Pero si esta opción no está seleccionada, aunque todavía puede obtener ambas categorías, el concepto `car seat manufacturer` solo aparecerá como un descriptor en la categoría que mejor coincide en función de varios factores, incluido el número de registros en los que `car seat` y `manufacturer` aparecen por separado.

Resolver nombres de categoría duplicados mediante Seleccione cómo tratar las categorías o subcategorías nuevas cuyos nombres sean iguales que los de categorías existentes. Puede fusionar las nuevas (y sus descriptores) con las categorías existentes que tengan el mismo nombre. Asimismo, puede omitir la creación de categorías si se encuentra un nombre duplicado en las categorías existentes.

Administración de pares de excepciones de enlace

Durante el proceso de generación de categorías, clúster y asignación de conceptos, los algoritmos internos agrupan las palabras en asociaciones conocidas. Para impedir que se emparejen dos conceptos, o que se enlacen, puede activar esta función en el diálogo **Valores avanzados de creación de categorías**, **Crear clústeres** diálogo y el diálogo **Valores de indexación de correlación de conceptos** y pulsar el botón **Gestionar pares**.

En el cuadro de diálogo resultante **Administrar excepciones de enlace**, puede añadir, editar o eliminar parejas de conceptos. Introduzca un par por cada línea. La introducción de pares aquí evitará que se produzca el emparejamiento al generar o ampliar categorías, clúster y asignación de conceptos. Introduzca las palabras exactas que quiera, por ejemplo, una palabra con tilde no es igual que una palabra sin tilde.

Por ejemplo, si desea asegurarse de que `hot dog` y `dog` no están agrupados, puede añadir el par como una línea separada en la tabla.

Acerca de técnicas lingüísticas

Al crear o ampliar categorías, puede seleccionar entre una serie de técnicas de creación de categorías lingüísticas avanzadas como, por ejemplo, *inclusión de concepto* y *redes semánticas* (solo texto en inglés). Estas técnicas pueden utilizarse solas o combinadas entre sí para crear categorías.

No es necesario que sea un experto en este tipo de configuración para poder utilizarla. De forma predeterminada, los valores más comunes ya aparecen seleccionados. Puede pasar por alto los diálogos de configuración avanzada y empezar a generar o ampliar categorías directamente. Igualmente, si realiza cambios aquí, no es necesario que vuelva al diálogo de configuración cada vez, puesto que siempre se conservan los últimos valores.

Tenga en cuenta que, puesto que cada conjunto de datos es exclusivo, el número de métodos y el orden en el que los aplique puede cambiar con el tiempo. Puesto que los objetivos de la minería de textos puede diferir de un conjunto de datos a otro, puede que deba experimentar con diferentes técnicas para comprobar con cuál de ellas se obtiene el mejor resultado para los datos de texto determinados. Ninguna de las técnicas automáticas categorizará perfectamente sus datos; por lo tanto, recomendamos buscar y aplicar una o más técnicas automáticas que funcionen correctamente con sus datos.

Las principales técnicas lingüísticas automáticas para la generación de categorías son:

- **Inclusión de conceptos.** Esta técnica crea categorías tomando un concepto y buscando otros conceptos que lo incluyan. Para obtener más información, consulte el tema [“Inclusión de conceptos”](#) en la [página 112](#).
- **Red semántica.** Esta técnica comienza identificando los posibles sentidos de cada concepto a partir de un amplio índice de relaciones de palabras, y luego crea categorías agrupando los conceptos relacionados. Para obtener más información, consulte el tema [“Redes semánticas”](#) en la [página 112](#). Esta opción sólo está disponible para textos en inglés.

Derivación de raíz de conceptos

La técnica de derivación de raíz de conceptos crea categorías tomando un concepto y buscando otros conceptos que estén relacionados con el primero analizando si alguno de los componentes de los conceptos está morfológicamente relacionado. Un componente es una palabra. Esta técnica intenta

agrupar conceptos observando la terminación (el sufijo) de cada componente de un concepto y buscando otros conceptos que puedan derivar de los primeros. La idea radica en que cuando las palabras derivan unas de otras, probablemente tienen el mismo significado o parecido. Para poder identificar las terminaciones, se utilizan reglas específicas de idioma. Por ejemplo, el concepto `opportunities to advance` se agruparía con los conceptos `opportunity for advancement` y `advancement opportunity`.

Puede utilizar la derivación de raíz de conceptos en cualquier tipo de texto. Por sí mismo genera pocas categorías, y cada una de ellas suele contener pocos conceptos. Los conceptos de cada categoría son sinónimos o están relacionados por posición. Puede resultarle útil emplear este algoritmo aunque esté generando categorías manualmente; los sinónimos que encuentre pueden ser sinónimos de los conceptos en los que está particularmente interesado.

Nota: Puede impedir que se agrupen conceptos juntos especificándolos explícitamente. Para obtener más información, consulte el tema [“Administración de pares de excepciones de enlace”](#) en la página 110.

Desglose de términos en componentes y desarticulación

Cuando se aplican las técnicas de derivación de raíz o inclusión de conceptos, en primer lugar los términos se desglosan en componentes (palabras) y luego los componentes se desarticulan. Cuando se aplica una técnica, los conceptos y sus términos asociados se cargan y se dividen en componentes basándose en separadores, como espacios, guiones y apóstrofes. Por ejemplo, el término `system administrator` se divide en componentes, como `{administrator, system}`.

Sin embargo, es posible que algunas partes del término original no se utilicen, las cuales se consideran palabras vacías. En inglés, algunos de estos componentes que se pueden omitir pueden incluir `a`, `and`, `as`, `by`, `for`, `from`, `in`, `of`, `on`, `or`, `the`, `to` y `with`.

Por ejemplo, el término `examination of the data` tiene el conjunto de componentes `{data, examination}`, y se considera que tanto `of` como `the` se puede omitir. Además, el orden de los componentes no se refleja en un conjunto de componentes. De este modo, los tres términos siguientes podrían ser equivalentes: `cough relief for child`, `child relief from a cough` y `relief of child cough` ya que todos tienen el mismo conjunto de componentes `{child, cough, relief}`. Cada vez que una pareja de términos se identifica como equivalente, los conceptos correspondientes se fusionan para formar un concepto nuevo que haga referencia a todos los términos.

Además, puesto que los componentes de un término pueden estar declinados, internamente se aplican reglas específicas del idioma para identificar los términos equivalentes independientemente de la variación de la declinación, por ejemplo, las formas plurales. De este modo, los términos `level of support` y `support levels` se pueden identificar como equivalentes, ya que la forma singular sin conjugar sería `level`.

Cómo funciona la derivación de raíz de conceptos

Cuando se ha aplicado la estructuración en componentes y la desarticulación de los términos (consulte la sección anterior), el algoritmo de derivación de raíz de conceptos analiza las terminaciones o sufijos de los componentes con el fin de encontrar la raíz del componente y luego agrupar los conceptos con otros que tengan raíces iguales o similares. Las terminaciones se identifican a partir de un conjunto de reglas de derivación lingüística específicas del idioma del texto. Por ejemplo, en inglés existe una regla de derivación por la que la terminación del componente de un término con el sufijo `ical` puede derivar de un término que tenga la misma raíz y terminación con el sufijo `ic`. Usando esta regla (y no conjugación), el algoritmo podría agrupar los conceptos `epidemiologic study` y `epidemiological studies`.

Puesto que los términos ya están componentizados y los componentes omitibles (por ejemplo, `in` y `of`) se han identificado, el algoritmo de derivación de raíz de concepto también podría agrupar el concepto `studies in epidemiology` con `epidemiological studies`.

Se ha elegido el conjunto de reglas de derivación de componentes para que la mayoría de los conceptos agrupados por este algoritmo sean sinónimos: los conceptos `epidemiologic studies`, `epidemiological studies`, `studies in epidemiology` son términos equivalentes. Para aumentar la precisión, existen algunas reglas de derivación que permiten al algoritmo agrupar conceptos que están

relacionados por posición. Por ejemplo, el algoritmo puede agrupar conceptos ingleses como `empire builder` y `empire building`.

Inclusión de conceptos

La técnica de inclusión de conceptos genera categorías tomando un concepto y, mediante los algoritmos de series léxicas, identifica los conceptos que están incluidos en otros conceptos. La idea radica en que cuando las palabras de un concepto forman un subconjunto de otro concepto, refleja una relación semántica subyacente. La inclusión es una potente técnica que puede utilizarse con cualquier tipo de texto.

Esta técnica funciona bien en combinación con las redes semánticas, pero puede utilizarse por separado. La inclusión de conceptos puede arrojar mejores resultados cuando los documentos o registros contienen una gran cantidad de jerga o terminología específica del dominio. Esto es especialmente cierto si ha ajustado los diccionarios de antemano para permitir la extracción y agrupación apropiada de términos especiales (con sinónimos).

Funcionamiento de la inclusión de conceptos

Antes de aplicar el algoritmo de inclusión de conceptos, los términos se estructuran en componentes y se desarticulan. Para obtener más información, consulte el tema [“Derivación de raíz de conceptos” en la página 110](#). A continuación, el algoritmo de inclusión de conceptos analiza los conjuntos de componentes. Para cada conjunto de componentes, el algoritmo busca otro conjunto de componentes que sea un subconjunto del primer conjunto de componentes.

Por ejemplo, si tiene el concepto `continental breakfast`, que tiene el conjunto de componentes `{breakfast, continental}`, y tiene el concepto `breakfast`, que tiene el conjunto de componentes `{breakfast}`, el algoritmo concluirá que `continental breakfast` es un tipo de `breakfast` y los agrupará.

En un ejemplo más grande, si tiene el concepto `seat` en el panel Resultados de extracción y aplica este algoritmo, conceptos como `safety seat`, `leather seat`, `seat belt`, `seat belt buckle`, `infant seat carrier` y `car seat laws` también se agruparán en esa categoría.

Puesto que los términos ya están componentizados y los componentes omitibles (por ejemplo, `in` y `of`) se han identificado, el algoritmo de inclusión de conceptos reconocerá que el concepto `advanced spanish course` incluye el concepto `course in spanish`.

Nota: Puede impedir que se agrupen conceptos especificándolos explícitamente. Para obtener más información, consulte el tema [“Administración de pares de excepciones de enlace” en la página 110](#).

Redes semánticas

En esta versión, la técnica de redes semánticas solo está disponible para texto en inglés.

Esta técnica genera categorías utilizando una red incorporada de relaciones de palabras. Por esta razón, esta técnica puede generar resultados muy buenos cuando los términos son concretos y no contienen demasiadas ambigüedades. Sin embargo, no confíe en que esta técnica encuentre muchos enlaces entre conceptos especializados y muy técnicos. Cuando trabaje con este tipo de conceptos, el empleo de las técnicas de inclusión y derivación de raíz de conceptos le resultará más útil.

Funcionamiento de la red semántica

La idea que encierra la técnica de la red semántica es aprovechar las relaciones de las palabras comunes para crear categorías de sinónimos o hipónimos. Un **hipónimo** es un concepto que constituye una especie de concepto secundario en una relación jerárquica, también conocida como relación "ISA". Por ejemplo, si `animal` es un concepto, entonces `cat` y `kangaroo` son hipónimos de `animal` ya que son tipos de animales.

Además de las relaciones de sinónimos e hipónimos, la técnica de red semántica también examina parte y enlaces completos entre los conceptos del tipo `<Location>`. Por ejemplo, la técnica agrupará los conceptos `normandy`, `provence` y `france` en una categoría porque la Normandía y la Provenza son partes de Francia.

Las redes semánticas empiezan identificando los sentidos posibles de cada concepto de la red semántica. Cuando los conceptos se identifican como sinónimos o hipónimos, se agrupan en una sola categoría. Por ejemplo, la técnica crearía una única categoría con estos tres conceptos: `eating apple`, `dessert apple` y `granny smith`, ya que la red semántica contiene la información de que: 1) `dessert apple` es un sinónimo de `eating apple` y 2) `granny smith` es una clase de `eating apple` (lo que significa que es un hipónimo de `eating apple`).

Si se consideran individualmente, muchos conceptos, sobre todo los unitérminos, son ambiguos. Por ejemplo, el concepto `buffet` puede significar un tipo de comida o un mueble. Si el conjunto de conceptos incluye `meal`, `furniture` y `buffet`, el algoritmo se ve obligado a elegir entre agrupar `buffet` con `meal` o con `furniture`. Tenga en cuenta que en algunos casos, las opciones que elige el algoritmo pueden no ser apropiadas en el contexto de un conjunto particular de registros o documentos.

La técnica de la red semántica puede generar un mejor rendimiento de la inclusión de conceptos con determinados tipos de datos. Aunque tanto la red semántica como la inclusión de conceptos reconocen que `apple pie` es una clase de `pie`, solo la red semántica reconoce que `tart` también es una clase de `pie`.

Las redes semánticas funcionarán en conjunción con el resto de las técnicas. Por ejemplo, suponga que ha seleccionado tanto la red semántica como las técnicas de inclusión y que la red semántica ha agrupado el concepto `teacher` con el concepto `tutor` (porque un tutor es una clase de profesor). El algoritmo de inclusión puede agrupar el concepto `graduate tutor` con `tutor` y, como resultado, los dos algoritmos colaboran para generar una categoría de salida que contiene los tres conceptos: `tutor`, `graduate tutor` y `teacher`.

Opciones de la red semántica

Existe una serie de valores adicionales que pueden ser de interés para esta técnica.

- Cambie la **Distancia máxima de búsqueda**. Seleccione el alcance de búsqueda de las técnicas antes de generar las categorías. Cuanto más bajo sea el valor, menor será la cantidad de resultados producidos; sin embargo, estos resultados serán menos ruidosos y más propensos a ser enlazados o asociados significativamente uno con el otro. Cuando más alto sea el valor, mayor será la cantidad de resultados obtenidos; sin embargo, estos resultados pueden ser menos fiables o relevantes.

Por ejemplo, dependiendo de la distancia, el algoritmo busca desde `Danish pastry` hasta `coffee roll` (su padre), luego `bun` (abuelo) y hacia arriba a `bread`.

Al reducir la distancia de búsqueda, esta técnica produce categorías más pequeñas con las que debería ser más fácil trabajar si cree que las categorías que se están produciendo son demasiado grandes o agrupan demasiados elementos.

Importante: Además, se recomienda no aplicar la opción **Acomodar la ortografía a un límite mínimo de caracteres raíz de** (definido en la pestaña Experto del nodo o en el cuadro de diálogo Extraer) para la agrupación difusa cuando se utiliza esta técnica, puesto que algunas agrupaciones falsas pueden tener un impacto muy negativo en los resultados.

Reglas de coocurrencia

Las reglas de co-ocurrencia le permiten descubrir y agrupar conceptos fuertemente relacionados dentro del conjunto de documentos o registros. La idea radica en que cuando en los documentos y registros a menudo se encuentran conceptos que aparecen juntos, esa co-ocurrencia refleja una relación subyacente que probablemente sea valiosa para las definiciones de categorías. Esta técnica crea reglas de co-ocurrencia que pueden utilizarse para crear una categoría nueva, para ampliar una categoría o como entrada a otra técnica de categoría. Se considera que la co-ocurrencia de dos conceptos es muy alta si estos aparecen con frecuencia juntos en un conjunto de registros y lo hacen raramente separados en el resto de los registros. Esta técnica puede generar buenos resultados con conjuntos de datos más extensos que tengan al menos varios centenares de documentos o registros.

Por ejemplo, si varios registros contienen las palabras `price` y `availability`, estos conceptos podrían agruparse en una regla de co-ocurrencia, (`price & available`). En otro ejemplo, si los conceptos `peanut butter`, `jelly` y `sandwich` aparecen en conjunto con más frecuencia que separados, se agruparían en una regla de co-ocurrencia de concepto (`peanut butter & jelly & sandwich`).

Importante: En versiones anteriores, las reglas de coocurrencia y de sinónimos se colocaban entre corchetes. En esta versión, los corchetes ahora indican un resultado de patrón de análisis de enlace de texto. En su lugar, las reglas de coocurrencia y de sinónimos se encapsularán entre paréntesis, como (speaker systems|speakers).

Funcionamiento de las reglas de coocurrencia

Esta técnica explora los documentos o registros en busca de dos o más conceptos que tiendan a aparecer juntos. Se considera que dos o más conceptos son co-ocurrentes cuando aparecen con frecuencia juntos en un conjunto de documentos o registros y si raramente aparecen separados en cualquiera de los otros documentos o registros.

Cuando se encuentran conceptos co-ocurrentes, se genera una regla de categoría. Estas reglas constan de dos o más conceptos conectados entre sí mediante un operador booleano &. Estas reglas son sentencias lógicas que clasifican automáticamente a un documento o registro en una categoría siempre que el conjunto de conceptos de la regla co-ocurran en ese documento o registro.

Opciones de las reglas de coocurrencia

Si utiliza la técnica de reglas de co-ocurrencia, puede ajustar varios de los valores de configuración que influyen en las reglas resultantes:

- Cambie la **Distancia máxima de búsqueda**. Seleccione hasta qué punto desea que la técnica busque co-ocurrencias. A medida que aumenta la distancia de búsqueda, el valor de similitud mínimo necesario para cada co-ocurrencia se reduce; como resultado, pueden producirse muchas co-ocurrencias, pero aquellas que tengan un valor de similitud bajo a menudo tendrán poca importancia. A medida que reduce la distancia de búsqueda, aumenta el valor de similitud mínimo necesario; como resultado, se producen menos reglas de co-ocurrencia, pero estas tenderán a ser más significativas (más fuertes).
- **Número mínimo de documentos**. El número mínimo de registros o documentos que deben contener un determinado par de conceptos para que se considere una co-ocurrencia; cuanto más bajo establezca esta opción, más fácil será encontrar co-ocurrencias. El aumento del valor da como resultado menos co-ocurrencias, pero más significativas. Como ejemplo, supongamos que se encuentran los conceptos "manzana" y "pera" juntos en 2 registros (y que ninguno de los dos conceptos aparece en otros registros). Con **Número mínimo de documentos** establecido en 2 (el valor predeterminado), la técnica de co-ocurrencia creará una regla de categoría (manzana y pera). Si el valor se aumenta a 3, la regla ya no se creará.

Nota: con pequeños conjuntos de datos (< 1000 respuestas) es posible que no encuentre ninguna co-aparición con los valores predeterminados. Si es así, intente aumentar el valor de distancia de búsqueda.

Nota: Puede impedir que se agrupen conceptos especificándolos explícitamente. Para obtener más información, consulte el tema [“Administración de pares de excepciones de enlace”](#) en la página 110.

Configuración avanzada de frecuencia

Puede generar categorías basándose en una técnica de frecuencia mecánica o directa. Con esta técnica, puede generar una categoría para cada elemento (tipo, concepto o patrón) que se haya encontrado por encima del recuento de un registro o documento determinado. También puede generar una sola categoría para todos los elementos que se produzcan con menos frecuencia. Por recuento se entiende el número de registros o documentos que contienen el concepto extraído (y cualquiera de sus sinónimos) o el tipo o patrón en cuestión, en contraposición al número total de apariciones en todo el texto.

La agrupación de elementos que aparecen con frecuencia puede arrojar resultados interesantes, ya que puede indicar una respuesta común o significativa. Esta técnica es muy útil si se ejecuta sobre los resultados de extracción sin utilizar después de haber aplicado otras técnicas. Otra aplicación es ejecutar esta técnica inmediatamente después de la extracción si no existe ninguna otra categoría, editar los resultados para suprimir las categorías que no interesen, y luego ampliar esas categorías para que coincidan con más registros o documentos. Para obtener más información, consulte el tema [“Ampliación de categorías”](#) en la página 115.

En lugar de utilizar esta técnica, puede clasificar los conceptos o los patrones de conceptos disminuyendo el número de registros o de documentos en el panel Resultados extraídos y luego arrastrar los principales y soltarlos en el panel Categorías para crear las categorías correspondientes.

Los campos siguientes están disponibles dentro del recuadro de diálogo Valores avanzados: Frecuencias:

Generar descriptores de categoría en. Seleccione este tipo de entrada para los descriptores. Para obtener más información, consulte el tema [“Crear categorías”](#) en la página 106.

- **Nivel de conceptos.** Si selecciona esta opción significa que se utilizarán las frecuencias de los conceptos o de los patrones de conceptos. Se utilizarán conceptos si se han seleccionado tipos como entrada para la generación de categorías, y patrones de conceptos si se han seleccionado patrones de tipo. En general, la aplicación de esta técnica al nivel de conceptos genera resultados más específicos, ya que los conceptos y los patrones de conceptos representan un nivel inferior de medición.
- **Nivel de tipos.** Si selecciona esta opción significa que se utilizarán las frecuencias de tipos o de patrones de tipo. Se utilizarán tipos si se han seleccionado tipos como entrada para la generación de categorías, y patrones de tipos si se han seleccionado patrones de tipo. La aplicación de esta técnica al nivel de tipo le permite obtener una vista rápida con respecto al tipo de información presente determinado.

Registro doc. número para los elementos que tienen su propia categoría. Esta opción permite generar categorías a partir de elementos que aparecen con frecuencia. Esta opción restringe el resultado sólo a las categorías que contengan un descriptor que haya aparecido como mínimo en X registros o documentos, donde X es el valor que hay que introducir para esta opción.

Agrupar todos los elementos restantes en una categoría denominada. Esta opción permite agrupar en una única categoría con el nombre de su elección todos los conceptos o tipos que aparecen rara vez. De forma predeterminada, esta categoría se llama *Otros*.

Entrada de categoría. Seleccione el grupo al que aplicar las técnicas:

- **Resultados de extracción no usados.** Esta opción permite que se generen categorías a partir de los resultados de extracción que no se utilizan en ninguna categoría existente. De esta manera se minimiza la tendencia que tienen los registros de hacer coincidir varias categorías y limita el número de categorías que se generan.
- **Todos los resultados de la extracción.** Esta opción permite generar categorías utilizando los resultados de extracción. Esto resulta especialmente útil cuando no existen categorías, o existen muy pocas.

Resolver nombres de categoría duplicados mediante. Seleccione cómo tratar las categorías o subcategorías nuevas cuyos nombres sean iguales a categorías existentes. Puede fusionar las nuevas (y sus descriptores) con las categorías existentes que tengan el mismo nombre. Asimismo, puede omitir la creación de categorías si se encuentra un nombre duplicado en las categorías existentes.

Ampliación de categorías

La ampliación es un proceso a través del cual se añaden descriptores o se mejoran automáticamente para "aumentar" las categorías existentes. El objetivo es generar una categoría mejor que capture los registros o documentos relacionados que no se asignaron originalmente a dicha categoría.

Las técnicas automáticas de agrupación que seleccione intentarán identificar conceptos, patrones TLA y reglas de categoría relacionadas con los descriptores de categorías existentes. Estos nuevos conceptos, patrones y reglas de categoría se añadirán como nuevos descriptores, o lo harán a los descriptores existentes. Las técnicas de agrupación para ampliar incluyen la *Derivación de raíz de concepto*, la *inclusión conceptual*, las *redes semánticas* (solo en inglés) y las *reglas de coaparición*. El método **Ampliar categorías vacías con descriptores generados desde el nombre de categorías** genera descriptores utilizando las palabras de los nombres de categoría, por lo tanto, cuanto más descriptivos sean los nombres de las categorías, mejores serán los resultados.

Nota: Las técnicas de frecuencia no están disponibles al ampliar categorías.

La ampliación es una excelente manera de mejorar interactivamente las categorías. He aquí algunos ejemplos de cuándo ampliar una categoría:

- Después de arrastrar y soltar patrones de conceptos para crear categorías en el panel Categorías
- Después de crear categorías manualmente y añadir reglas de categoría y descriptores simples
- Después de importar un archivo de categorías predefinidas en el que las categorías tenían nombres muy descriptivos
- Después de refinar las categorías que procedían del TAP que eligió

Puede ampliar una categoría varias veces. Por ejemplo, si ha importado un archivo de categoría predefinida con nombres muy descriptivos, puede realizar la ampliación utilizando la opción **Ampliar categorías vacías con descriptores generados desde el nombre de categorías** para obtener un primer conjunto de descriptores, y luego volver a ampliar estas categorías. Sin embargo, en otros casos, realizar la ampliación en distintas ocasiones puede dar como resultado una categoría demasiado genérica si los descriptores se amplían cada vez más. Puesto que las técnicas de agrupación de generación y de ampliación utilizan algoritmos subyacentes similares, es improbable que la ampliación directa después de generar las categorías genere resultados más interesantes.

Consejo:

- Si intenta ampliar y no desea utilizar los resultados, siempre puede deshacer la operación (**Editar > Deshacer**) inmediatamente después de haber ampliado.
- La ampliación puede generar dos o más reglas de categoría en una categoría que coincidan exactamente con el mismo conjunto de documentos, puesto que las reglas se generan de manera independiente durante el proceso. Si lo desea, puede revisar las categorías y eliminar redundancias editando manualmente la descripción de la categoría. Para obtener más información, consulte el tema [“Edición de descriptores de categoría” en la página 139](#).

Ampliar categorías

1. En el panel Categorías, seleccione las categorías que desea ampliar.
 2. En los menús, seleccione **Categorías > Ampliar categorías**. A menos que haya elegido que no se le pregunte nunca, aparecerá un cuadro de mensaje.
 3. Elija si desea generar ahora o editar primero la configuración.
- Pulse **Ampliar ahora** para empezar a ampliar categorías utilizando la configuración actual. El proceso comienza y aparece un diálogo de progreso.
 - Pulse **Editar** para revisar y modificar la configuración.

Después de intentar la ampliación, las categorías para las que se encuentren nuevos descriptores se marcan mediante la palabra **Ampliado** en el panel Categorías, para que pueda identificarlas rápidamente. El texto Ampliado permanecerá hasta que amplíe de nuevo, edite la categoría de otra forma o lo borre mediante el menú contextual.

Nota: El número máximo de categorías que se pueden mostrar es 10.000. Se muestra un aviso si se alcanza o supera este número. Si esto sucede, debe cambiar las opciones de creación o ampliación de categorías para reducir el número de categorías creadas.

Cada una de las técnicas disponibles al generar o ampliar categorías resulta idónea para determinados tipos de datos y situaciones, pero a menudo conviene combinar técnicas en el mismo análisis para capturar el rango completo de documentos o registros. En el área de trabajo interactiva, los conceptos y tipos agrupados en una categoría siguen estando disponibles la próxima vez que cree categorías. Esto significa que puede ver un concepto en diversas categorías o detectar categorías redundantes.

Las siguientes áreas y campos están disponibles dentro del recuadro de diálogo Ampliar categorías: Valores:

Ampliar con. Seleccione qué entrada se utilizará para ampliar las categorías:

- **Resultados de extracción no usados.** Esta opción permite que se generen categorías a partir de los resultados de extracción que no se utilizan en ninguna categoría existente. De esta manera se minimiza la tendencia que tienen los registros de hacer coincidir varias categorías y limita el número de categorías que se generan.

- **Todos los resultados de la extracción.** Esta opción permite generar categorías utilizando los resultados de extracción. Esto resulta especialmente útil cuando no existen categorías, o existen muy pocas.

Técnicas de agrupación

Para obtener descripciones cortas de cada una de estas técnicas, consulte [“Configuración lingüística avanzada”](#) en la página 108. Estas técnicas incluyen:

- **Derivación de raíz de concepto**
- **Red semántica** (solo texto en inglés, y no se utiliza si está seleccionada la opción Solo generalizar.)
- **Inclusión de conceptos**
- **Co-ocurrencia** y la subopción **Número mínimo de documentos**.

Hay un número de tipos que están permanentemente excluidos de la técnica de redes semánticas, porque dichos tipos no generan resultados relevantes. Incluyen <Positive>, <Negative>, <IP>, otros tipos no lingüísticos, etc.

Distancia máxima de búsqueda Seleccione el alcance de búsqueda de las técnicas antes de generar las categorías. Cuanto más bajo sea el valor, menor será la cantidad de resultados obtenidos; sin embargo, estos resultados serán menos ruidosos y más propensos a ser enlazados o asociados significativamente uno con el otro. Cuando más alto sea el valor, mayor será la cantidad de resultados obtenidos; sin embargo, estos resultados pueden ser menos fiables o relevantes. Esta opción se aplica globalmente en todas las técnicas, pero sus resultados son mejores en las co-ocurrencias y en las redes semánticas.

Impedir el emparejamiento de conceptos específicos. Seleccione esta casilla de verificación para detener el proceso de agrupación o emparejamiento de dos conceptos en la salida. Para crear o gestionar pares de conceptos, pulse **Gestionar pares**. Para obtener más información, consulte el tema [“Administración de pares de excepciones de enlace”](#) en la página 110.

Donde sea posible: Elija si desea simplemente ampliar, generalizar los descriptores utilizando comodines, o ambos.

- **Ampliar y generalizar.** Esta opción ampliará las categorías seleccionadas y, a continuación, generalizará los descriptores. Si selecciona generalizar, el producto creará reglas de categoría genéricas en categorías mediante el comodín de asterisco. Por ejemplo, en lugar de generar varios descriptores como [apple tart + .] y [apple sauce + .], utilizar comodines puede generar [apple * + .]. Si generaliza con comodines, obtendrá a menudo exactamente el mismo número de registros o documentos que antes. Sin embargo, esta opción tiene la ventaja de reducir el número y simplificar los descriptores de categoría. Además, esta opción aumenta la posibilidad de categorizar más registros o documentos utilizando estas categorías en datos de texto nuevos (por ejemplo, en estudios por fases o longitudinales).
- **Sólo ampliar.** Esta opción ampliará sus categorías sin generalizar. Puede ser de utilidad seleccionar primero la opción **Sólo ampliar** para las categorías creadas manualmente y, a continuación, ampliar las mismas categorías de nuevo mediante la opción **Ampliar y generalizar**.
- **Sólo generalizar.** Esta opción generalizará los descriptores sin ampliar sus categorías de ningún otro modo.

Nota: La selección de esta opción inhabilita la opción **Red semántica**; esto se debe a que la opción **Red semántica** solo está disponible cuando se va a ampliar una descripción.

Otras opciones para ampliar categorías

Además de seleccionar las técnicas que se aplicarán, puede editar cualquiera de las opciones siguientes:

Número máximo de elementos por los que ampliar un descriptor. Cuando se amplía un descriptor con elementos (conceptos, tipos y otras expresiones), define el número máximo de elementos que pueden añadirse a un solo descriptor. Si establece este límite en 10, no podrá añadir más de 10 elementos adicionales a un descriptor existente. Si hay más de 10 elementos para añadir, las técnicas dejan de añadir elementos nuevos cuando se alcanza el número diez. Con ello se puede reducir la lista de un descriptor, pero no se garantiza que se utilicen en primer lugar los elementos más interesantes. Puede

resultar preferible reducir el tamaño de la ampliación sin comprometer la calidad; para ello utilice la opción **Generalizar con comodines cuando sea posible**. Esta opción solo se aplica a los descriptores que contienen los operadores booleanos & (AND) o ! (NOT).

Ampliar también subcategorías. Esta opción también ampliará cualquier subcategoría por debajo de las categorías seleccionadas.

Ampliar categorías vacías con descriptores generados a partir del nombre de categoría. Este método solo se aplica a categorías vacías, que tienen 0 descriptores. Si una categoría ya contiene descriptores, no se ampliará de esta forma. Esta opción intenta generar descriptores automáticamente para cada categoría basándose en las palabras que conforman el nombre de la categoría. El nombre de la categoría se explora para comprobar si las palabras que conforman el nombre coinciden con alguno de los conceptos extraídos. Si se reconoce un concepto, se utilizará para buscar patrones de conceptos coincidentes, y ambos se utilizarán para generar descriptores para la categoría. Esta opción genera los mejores resultados cuando los nombres de categoría son largos y descriptivos. Se trata de un método rápido para generar descriptores de categorías, que a su vez permiten a la categoría capturar registros que contienen dichos descriptores. Esta opción es muy útil cuando se importan categorías desde otro punto o cuando crea categorías manualmente con nombres descriptivos largos.

Generar descriptores como. Esta opción sólo se aplica si la opción anterior está seleccionada.

- **Conceptos.** Seleccione esta opción para producir los descriptores resultantes en forma de conceptos, independientemente de si se han extraído del texto de origen.
- **Patrones.** Seleccione esta opción para producir los descriptores resultantes en forma de patrones, independientemente de si se han extraído los patrones resultantes o cualquier patrón.

Creación manual de categorías

Además de crear categorías utilizando las técnicas de generación automática de categorías y el editor de reglas, también puede crear categorías manualmente. Existen los métodos manuales siguientes:

- Crear una categoría vacía en la que se añadirán elementos uno a uno. Para obtener más información, consulte el tema [“Creación de categorías nuevas o cambio de nombre de categorías”](#) en la página 118.
- Arrastrar términos, tipos y patrones al panel de categorías. Para obtener más información, consulte el tema [“Creación de categorías mediante el método de arrastrar y soltar”](#) en la página 119.

Creación de categorías nuevas o cambio de nombre de categorías

Puede crear categorías vacías en las que añadir conceptos y tipos. También puede cambiar el nombre de las categorías.

Para crear una categoría vacía nueva

1. Vaya al panel Categorías.
2. En los menús, seleccione **Categorías > Crear categoría vacía**. Aparecerá el cuadro de diálogo Propiedades de categoría.
3. Escriba un nombre para esta categoría en el campo Nombre.
4. Pulse **Aceptar** para aceptar el nombre y cerrar el cuadro de diálogo. El cuadro de diálogo se cierra y el nombre de la nueva categoría aparece en el panel.

Ahora ya puede empezar a añadir elementos a esta categoría. Para obtener más información, consulte el tema [“Añadir descriptores a las categorías”](#) en la página 138.

Para cambiar el nombre de una categoría

1. Seleccione una categoría y elija **Categorías > Cambiar nombre de categoría**. Aparecerá el cuadro de diálogo Propiedades de categoría.
2. Escriba un nuevo nombre para esta categoría en el campo Nombre.
3. Pulse **Aceptar** para aceptar el nombre y cerrar el cuadro de diálogo. El cuadro de diálogo se cierra y el nombre de la nueva categoría aparece en el panel.

Creación de categorías mediante el método de arrastrar y soltar

La técnica de arrastrar y soltar es manual y no se basa en algoritmos. Puede crear categorías en el panel Categorías arrastrando a éste los elementos siguientes:

- Conceptos, tipos o patrones extraídos desde el panel Resultados extraídos al panel Categorías.
- Conceptos extraídos desde el panel Datos del panel Categorías.
- Filas completas desde el panel Datos del panel Categorías. Esto creará una categoría compuesta de todos los conceptos y patrones extraídos contenidos en esa fila.

Nota: El panel Resultados de extracción da soporte a varias selecciones para facilitar el arrastrado y soltado de varios elementos.

Importante: No pueden arrastrarse y soltarse conceptos desde el panel Datos que no se hayan extraído del texto. Si desea forzar la extracción de un concepto que ha encontrado en los datos, debe añadir este concepto a un tipo. Luego vuelva a ejecutar la extracción. Los nuevos resultados de la extracción contendrán el concepto que acaba de añadir. Luego puede utilizarlo en la categoría. Para obtener más información, consulte el tema [“Añadir conceptos a tipos”](#) en la página 91.

Para crear categorías utilizando la función arrastrar y soltar:

1. En el panel Resultados extraídos o en el panel Datos, seleccione uno o más conceptos, patrones, tipos, registros o registros parciales.
2. Mientras mantiene pulsado el botón del ratón, arrastre el elemento hasta una categoría existente o a un área del panel para crear una categoría nueva.
3. Cuando haya alcanzado el área donde desea soltar el elemento, suelte el botón del ratón. El elemento se añade al panel Categorías. Las categorías que se han modificado aparecen con un color de fondo especial. Este color se llama **fondo de comentario de categoría**. Para obtener más información, consulte el tema [“Opciones de configuración”](#) en la página 74.

Nota: La categoría resultante se ha denominado automáticamente. Si desea cambiar el nombre, puede hacerlo. Para obtener más información, consulte el tema [“Creación de categorías nuevas o cambio de nombre de categorías”](#) en la página 118.

Si desea ver qué registros están asignados a una categoría, seleccione la categoría en el panel Categorías. El panel de datos se actualiza automáticamente y muestra todos los registros de dicha categoría.

Uso de reglas de categoría

Puede crear categorías de muchas formas. Una de estas formas es definir reglas de categoría para expresar ideas. Las reglas de categoría son sentencias que clasifican automáticamente documentos o registros en una categoría que se basa en una expresión lógica utilizando los conceptos, tipos y patrones extraídos, así como operadores booleanos. Por ejemplo, podría escribir una expresión que significara *incluir todos los registros que contienen el concepto extraído embassy pero no argentina en esta categoría*.

Aunque algunas reglas de categoría se producen automáticamente al crear categorías utilizando técnicas de agrupación como *coocurrencia* y *Derivación de raíz de concepto* (**Categorías > Valores de compilación > Valores avanzados: Lingüística**), también puede crear reglas de categoría manualmente en el editor de reglas aprovechando su comprensión de categorías de los datos y el contexto. Cada regla se adjunta a una única categoría, de manera que cada documento o registro que coincida con la regla se registre en dicha categoría.

Las reglas de categoría ayudan a mejorar la calidad y la productividad de los resultados de la minería de textos y del análisis cuantitativo al permitirle categorizar las respuestas con mayor grado de especificidad. Tanto su experiencia como los conocimientos empresariales pueden proporcionarle mayor profundización en los datos y en el contexto. Puede aprovechar esta profundización para trasladar los conocimientos a reglas de categoría y categorizar así los documentos o los registros con mayor eficacia y precisión, ya que podrá combinar los elementos extraídos con la lógica booleana.

La capacidad de crear estas reglas mejora la precisión de la codificación, la eficacia y la productividad, ya que le permite desglosar en capas sus conocimientos empresariales en la tecnología de extracción del producto.

Nota: Para ver ejemplos sobre cómo coinciden las reglas con el texto, consulte [“Ejemplos de reglas de categoría”](#) en la página 127

Sintaxis de regla de categoría

Aunque algunas reglas de categoría se generan automáticamente al crear categorías utilizando técnicas de agrupación como *coocurrencia* y *Derivación de raíz de conceptos* (**Categorías > Valores de compilación > Valores avanzados: Lingüística**), también puede crear reglas de categoría manualmente en el editor de reglas. Cada regla se adjunta a una única categoría, de manera que cada documento o registro que coincida con la regla es anotado automáticamente en dicha categoría.

Nota: Para ver ejemplos sobre cómo coinciden las reglas con el texto, consulte [“Ejemplos de reglas de categoría”](#) en la página 127

Cuando cree o edite una regla, deberá tenerla abierta en un editor de reglas. Puede añadir conceptos, tipos o patrones y utilizar comodines para ampliar las coincidencias. Cuando se utilizan conceptos, tipos y patrones extraídos, puede beneficiarse de encontrar todos los conceptos relacionados.

Importante: para evitar errores comunes, se recomienda arrastrar los conceptos directamente desde el panel Resultados extraídos, desde el panel de análisis de enlace de texto o desde el panel Datos y soltarlos en el editor de reglas o añadirlos mediante los menús contextuales siempre que sea posible.

Cuando se reconocen los conceptos, tipos y patrones, aparece un icono junto al texto.

Tabla 19. Iconos de extracción	
Icono	Descripción
	Concepto extraído
	Tipo extraído
	Patrón extraído

Sintaxis y operadores de reglas

En la tabla siguiente encontrará los caracteres con los que puede definir la sintaxis de las reglas. Utilice estos caracteres junto con los conceptos, tipos y patrones para crear la regla.

Tabla 20. Sintaxis soportada	
Carácter	Descripción
&	El booleano "and". Por ejemplo, a & b contiene tanto a <i>como</i> b como: - invasion & united states - 2016 & olympics - good & apple

Tabla 20. Sintaxis soportada (continuación)

Carácter	Descripción
	El booleano "or" es inclusivo, lo que significa que si se encuentran algunos o todos los elementos, existe coincidencia. Por ejemplo, <code>a b</code> contiene tanto <code>a</code> como <code>b</code> como: - <code>attack france</code> - <code>condominium apartment</code>
!()	El booleano "not". Por ejemplo, <code>!(a)</code> no contiene <code>a</code> . tales como, <code>!(good & hotel), assassination & !(austria)</code> o <code>!(gold) & !(copper)</code>
*	Comodín que representa cualquier cosa, desde un solo carácter a una palabra completa dependiendo de cómo se utilice. Para obtener más información, consulte el tema “Uso de comodines en reglas de categoría” en la página 124.
()	Un delimitador de expresión. Las expresiones que están entre paréntesis se evalúan en primer lugar.
+	Conector de patrones que se utiliza para formar un patrón específico de orden. Cuando está presente, deben utilizarse corchetes. Para obtener más información, consulte el tema “Uso de patrones TLA en las reglas de categoría” en la página 122.
[]	Se requiere el delimitador de patrón si está buscando coincidir basándose en un patrón TLA extraído de una regla de categoría. El contenido de los corchetes representa los patrones TLA y no coincidirá nunca con los conceptos o tipos basados en una simple co-ocurrencia. Si no ha extraído este patrón TLA, ninguna coincidencia será posible. Para obtener más información, consulte el tema “Uso de patrones TLA en las reglas de categoría” en la página 122. No utilice corchetes si desea hacer coincidir conceptos y tipos en lugar de patrones. <i>Nota:</i> En versiones anteriores, la co-ocurrencia y las reglas de sinónimo generadas por las técnicas de creación de categorías solían estar entre corchetes. En todas las versiones nuevas, los corchetes ahora indican la presencia de un patrón TLA. En su lugar, las reglas generadas por la técnica de coocurrencia y los sinónimos se encapsularán entre paréntesis, como por ejemplo <code>(speaker systems speakers)</code> .

Los operadores `&` y `|` son conmutativos, de forma que `a & b = b & a` y `a | b = b | a`.

Salto de caracteres con la barra invertida

Si tiene un concepto que contiene cualquier carácter que sea también un carácter de sintaxis, deberá colocar una barra inclinada invertida delante de dicho carácter para que la regla se interprete correctamente. El carácter de barra inclinada invertida se utiliza para caracteres de escape que pueden tener un significado especial. Cuando realiza la acción de arrastrar y soltar en el editor, las barras inclinadas invertidas se colocan automáticamente.

Los caracteres de sintaxis de reglas deben ir precedidos por una barra inclinada invertida si quiere que se consideren como tal en vez de como sintaxis de regla:

```
& ! | + < > ( ) [ ] *
```

Por ejemplo, puesto que el concepto `r&d` contiene el operador "and" (`&`), se requiere la barra inclinada invertida cuando se escribe en el editor de reglas, como por ejemplo: `r\d`.

Uso de patrones TLA en las reglas de categoría

Los patrones de análisis de enlaces de texto se pueden especificar explícitamente en reglas de categoría para permitirle obtener resultados aún más específicos y contextuales. Cuando define un patrón en una regla de categoría, está omitiendo los resultados más sencillos de la extracción de conceptos, y sólo compara los documentos y registros basados en los resultados de patrones de análisis de enlaces de texto extraídos.

Importante: Para poder hacer coincidir documentos utilizando patrones TLA en sus reglas de categoría, debe haber ejecutado una extracción con el análisis de enlaces de texto activado. La regla de categoría buscará las coincidencias encontradas durante dicho proceso. Si no ha seleccionado explorar los resultados de TLA en la pestaña Modelo de su nodo de Text Mining, puede seleccionar activar la extracción de TLA en la configuración de extracción en la sesión interactiva y luego volver a extraer. Para obtener más información, consulte el tema [“Extraer datos”](#) en la página 82.

Delimitación con corchetes. Un patrón TLA debe colocarse entre corchetes [] si lo está utilizando dentro de una regla de categoría. Se requiere el delimitador de patrón si está buscando coincidir basándose en un patrón TLA extraído. Puesto que las reglas de categoría pueden contener tipos, conceptos o patrones; los corchetes sirven para aclarar a la regla que su contenido representa el patrón TLA extraído. Si no ha extraído este patrón TLA, ninguna coincidencia será posible. Si ve un patrón sin corchetes como `apple + good` en el panel Categorías, esto probablemente significa que el patrón se ha añadido directamente a la categoría fuera del editor de reglas de categoría. Por ejemplo, si añade un patrón de concepto directamente a la categoría de la vista análisis de enlaces de texto, no aparecerá con corchetes. Sin embargo, cuando se utiliza un patrón dentro de una regla de categoría, debe encapsular el patrón dentro de los corchetes dentro de la regla de categoría, como por ejemplo `[banana + ! (good)]`.

Utilización del signo + en patrones. En IBM SPSS Modeler Text Analytics, puede tener patrones de hasta seis partes (espacios). Para indicar que el orden es importante, utilice el signo + para conectar cada elemento, como por ejemplo `[company1 + acquired + company2]`. Aquí el orden es importante porque de lo contrario cambiaría el significado de qué empresa adquiere a cuál. El orden no se determina por la estructura de la frase, sino por cómo está estructurada la salida del patrón TLA. Por ejemplo, si tiene el texto *"I love Paris"* y desea extraer esta idea, es probable que el patrón de TLA sea `[paris + like]` o `[<Location> + <Positive>]` en lugar de `[<Positive> + <Location>]`, ya que los recursos de opinión predeterminados generalmente colocan las opiniones en la segunda posición en patrones de 2 partes. Por ello, utilizar el patrón directamente como descriptor en la categoría puede resultar útil para evitar problemas. Sin embargo, si necesita utilizar un patrón como parte de una frase más compleja, preste especial atención al orden de los elementos dentro de los patrones presentados en la vista de análisis de enlaces de texto, ya que el orden desempeña un papel importante en la búsqueda de una coincidencia.

Por ejemplo, supongamos que tenía en los dos textos de ejemplo siguientes la expresión: *"Me gusta la piña"* y *"Odio la piña"*. Sin embargo, me *gustan* las fresas". La expresión `like & pineapple` coincidiría con ambos textos, ya que es una expresión de concepto y no una regla de enlace de texto (no está entre corchetes). La expresión `pineapple + like` solo coincide con *"I like pineapple"* ya que en el segundo texto, la palabra *like* está asociada a *strawberries*.

Agrupación con patrones. Puede simplificar las reglas con sus propios patrones. Digamos que desea capturar las siguientes tres expresiones, `cayenne peppers + like`, `chili peppers + like` y `peppers + like`. Puede agruparlos en una regla de categoría única como, por ejemplo, `[* peppers & like]`. Si tiene otra expresión `hot peppers + good`, puede agrupar las cuatro con una regla como, por ejemplo, `[* peppers + <Positive>]`.

Ordenación en patrones. Para ordenar mejor la salida, las reglas de análisis de enlace de texto proporcionadas en las plantillas que ha instalado con su producto intentan sacar patrones básicos en el mismo orden independientemente del orden de las palabras en la frase. Por ejemplo, si tuviese un registro que contenga el texto *"Buenas presentaciones."* y otro registro que contiene *"las presentaciones fueron buenas"*, ambos textos coinciden con la misma regla y salida en el mismo orden que `presentation + good` en los resultados del patrón de concepto en lugar de `presentation + good` y también `good + presentation`. Y en un patrón de dos ranuras como los del ejemplo, los

conceptos asignados a los tipos en la biblioteca de Opiniones se presentarán en último lugar en la salida, de forma predeterminada, como por ejemplo `apple + bad`.

Tabla 21. Sintaxis de patrones y uso de booleanos	
Expresión	Coincide con un documento o registro que
[]	Contiene cualquier patrón TLA. Se requiere el delimitador de patrón <i>en las reglas de categoría</i> si está buscando coincidir basándose en un patrón TLA extraído. El contenido entre los corchetes se refiere a los patrones TLA y no a conceptos y tipos sencillos. Si no ha extraído este patrón TLA, ninguna coincidencia será posible. Si deseara crear una regla que no incluyese patrones, podría utilizar ! ([]).
[a]	Contiene un patrón del que al menos un elemento es a independientemente de su posición en el patrón. Por ejemplo, [deal] puede coincidir con [deal + good] o simplemente con [deal + .]
[a + b]	Contiene un patrón de concepto. Por ejemplo, [deal + good]. <i>Nota:</i> Si sólo desea capturar este patrón sin añadir otros elementos, recomendamos añadir el patrón directamente en la categoría en lugar de hacer una regla con el mismo.
[a + b + c]	Contiene un patrón de concepto. El signo + denota que el orden de los elementos coincidentes es importante. Por ejemplo, [company1 + acquired + company2].
[<A> +]	Contiene cualquier patrón con el tipo <A> en el primer espacio y el tipo en el segundo espacio, y hay exactamente dos espacios. El signo + denota que el orden de los elementos coincidentes es importante. Por ejemplo, [<Budget> + <Negative>]. <i>Nota:</i> Si sólo desea capturar este patrón sin añadir otros elementos, recomendamos añadir el patrón directamente en la categoría en lugar de hacer una regla con el mismo.
[<A> &]	Contiene cualquier patrón de tipo con el tipo <A> y . Por ejemplo, [<Budget> & <Negative>]. Este patrón de TLA nunca se extraerá; sin embargo, cuando se escribe como tal, es realmente igual a [<Budget> + <Negative>] [<Negative> + <Budget>]. El orden de los elementos coincidentes no es importante. Además, otros elementos pueden estar en el patrón, pero debe tener al menos a <Budget> y <Negative>.
[a + .]	Contiene un patrón donde a es el único concepto y no existe nada en ningún otro espacio de este patrón. Por ejemplo: [deal + .] coincide con el patrón de concepto donde la única salida es el concepto deal. Si ha añadido el concepto deal como un descriptor de categoría, obtendrá todos los registros del acuerdo como un concepto, incluidas las sentencias positivas sobre un acuerdo. Sin embargo, el uso de [deal + .] solo coincidirá con los resultados del patrón de registros que representen deal y no con otras relaciones u opiniones y no coincidiría con deal + fantastic. <i>Nota:</i> Si sólo desea capturar este patrón sin añadir otros elementos, recomendamos añadir el patrón directamente en la categoría en lugar de hacer una regla con el mismo.

Tabla 21. Sintaxis de patrones y uso de booleanos (continuación)

Expresión	Coincide con un documento o registro que
[<A> + <>]	Contiene un patrón donde <A> es el único tipo. Por ejemplo, [<Budget> + <>] coincide con el patrón en el que el único resultado es un concepto del tipo <Budget>. <i>Nota:</i> Puede utilizar <> para indicar un tipo vacío solo al colocarlo después del símbolo + de patrón en el patrón de tipo, como [<Budget> + <>] pero no [price + <>]. <i>Nota:</i> Si sólo desea capturar este patrón sin añadir otros elementos, recomendamos añadir el patrón directamente en la categoría en lugar de hacer una regla con el mismo.
[a + !(b)]	Contiene al menos un patrón que incluye el concepto a pero no incluye el concepto b. Debe incluir al menos un patrón. Por ejemplo, [price + !(high)] o para tipos, [!(<Fruit> <Vegetable>) + <Positive>]
!([<A> &])	No contiene un patrón específico. Por ejemplo, !([<Budget> & <Negative>]).

Nota: Para ver ejemplos sobre cómo coinciden las reglas con el texto, consulte [“Ejemplos de reglas de categoría”](#) en la página 127

Uso de comodines en reglas de categoría

Pueden añadirse comodines a los conceptos de las reglas para ampliar las posibilidades de coincidencia. El comodín asterisco (*) puede colocarse delante y/o detrás de una palabra para indicar cómo deben coincidir los conceptos. Existen dos tipos de usos de los comodines:

- **Comodines afijos.** Estos comodines se colocan inmediatamente antes o después de una cadena sin dejar espacios en blanco entre ésta y el asterisco. Por ejemplo, *operat** podría coincidir con *operat*, *operate*, *operates*, *operations*, *operational*, etc.
- **Comodines de palabras.** Estos comodines se colocan delante o detrás de un concepto con un espacio entre éste y el asterisco. Por ejemplo, ** operation* podría coincidir con *operation*, *surgical operation*, *post operation*, etc. Además, se puede utilizar un comodín de palabras a lo largo de un comodín affix como, por ejemplo, ** operat* **, que podría coincidir con *operation*, *surgical operation*, *telephone operator*, *operatic aria*, etc. Como puede comprobar en este último ejemplo, se recomienda utilizar los comodines con precaución para que el rango no sea excesivamente amplio y no se capturen coincidencias no deseadas.

¡Excepciones!

- Un comodín nunca puede ser un elemento individual. Por ejemplo, (apple | *) no se aceptará.
- Un comodín no se puede utilizar nunca para hacer coincidir nombres de tipo. <Negative>* no coincidirá con ningún nombre de tipo.
- No puede filtrar algunos tipos para evitar que coincidan con conceptos encontrados a través de los comodines. El tipo al que está asignado el concepto se utiliza automáticamente.
- Un comodín no puede nunca estar situado en el centro de una secuencia de palabras, ni al final o el comienzo de una palabra (abrir* cuenta) ni como componente independiente (abrir * cuenta). Tampoco puede utilizar comodines en nombres de tipo. Por ejemplo, word* word, como apple* recipe, no coincidirá con "applesauce recipe" ni con ninguna otra cosa. Sin embargo, apple* * coincidiría con *applesauce recipe*, *apple pie*, *apple*, etc. En otro ejemplo, word * word, como apple * toast, no

coincidirá con *apple cinnamon toast* ni ninguna otra cosa, ya que el asterisco aparece entre otras dos palabras. Sin embargo, *apple ** coincidiría con *apple cinnamon toast*, *apple*, *apple pie*, etc.

Tabla 22. Uso de los comodines	
Expresión	Coincide con un documento o registro que
*apple	Contiene un concepto que termina con las letras especificadas pero puede tener un número indefinido de letras como prefijo. Por ejemplo: *apple finaliza con las letras <i>apple</i> pero puede tomar un prefijo como: - apple - pineapple - crabapple
apple*	Contiene un concepto que empieza con las letras especificadas pero puede tener un número indefinido de letras como sufijo. Por ejemplo: apple* comienza con las letras <i>apple</i> pero puede tomar un sufijo (o ninguno) como: - apple - applesauce - applejack Por ejemplo, apple* & !(pear* quince), que contiene un concepto que empieza con las letras <i>apple</i> pero no un concepto que empieza por las letras <i>pear</i> o el concepto <i>quince</i>, no coincidiría con: apple & quince pero podría coincidir con: - applesauce - apple & orange
product	Contiene un concepto que contiene las letras escritas <i>product</i>, pero puede tener cualquier número de letras como prefijo o sufijo, o ambas cosas. Por ejemplo: *product* podría coincidir con: - product - byproduct - unproductive

Tabla 22. Uso de los comodines (continuación)

Expresión	Coincide con un documento o registro que
* loan	Contiene un concepto que contiene la palabra loan pero puede ser un término compuesto con otra palabra colocada antes que ella. Por ejemplo, * loan podría coincidir con: - loan - car loan - home equity loan Por ejemplo, [* delivery + <Negative>], que contiene un concepto que termina en la palabra delivery en la primera posición y contiene un tipo <Negative> en la segunda posición, podría coincidir con los siguientes patrones de concepto: - package delivery + slow - overnight delivery + late
event *	Contiene un concepto que contiene la palabra event pero puede ser un término compuesto seguido de otra palabra. Por ejemplo, event * podría coincidir con: - event - event location - event planning committee
* apple *	Contiene un concepto que podría comenzar con cualquier palabra seguida de la palabra apple posiblemente seguida de otra palabra. * significa 0 o n, por lo que también coincide con apple. Por ejemplo, * apple * podría coincidir con: - gala applesauce - granny smith apple crumble - famous apple pie - apple Por ejemplo, [* reservation* * + <Positive>], que contiene un concepto con la palabra reservation (independientemente de dónde está en el concepto) en la primera posición y contiene un tipo <Positive> en la segunda posición, podría coincidir con los patrones de concepto: - reservation system + good - online reservation + good

Nota: Para ver ejemplos sobre cómo coinciden las reglas con el texto, consulte “Ejemplos de reglas de categoría” en la página 127

Ejemplos de reglas de categoría

Para ayudar a demostrar cómo coinciden las reglas con los registros basados de manera diferente en la sintaxis utilizada para expresarlos, observe el siguiente ejemplo.

Ejemplo de registros

Imagine que tiene dos registros:

- **Registro A:** “cuando revisé mi billetera, noté que me faltaban 5 dólares.”
- **Registro B:** “se encontraron \$5 en el área de picnic, pero faltaba la manta.”

Las dos tablas siguientes muestran qué se puede extraer de los conceptos y tipos así como los patrones de concepto y patrones de tipo.

Conceptos y tipos extraídos de un ejemplo

Tabla 23. Ejemplo de conceptos y tipos extraídos	
Concepto extraído	Conceptos escritos como
wallet	<Unknown>
missing	<Negative>
USD5	<Currency>
blanket	<Unknown>
picnic area	<Unknown>

Patrones TLA extraídos de un ejemplo

Tabla 24. Ejemplo de resultados extraídos de patrones TLA		
Patrones de concepto extraídos	Patrones de tipo extraídos	En el registro
picnic area + .	<Unknown> + <>	Registro B
wallet + .	<Unknown> + <>	Registro A
blanket + missing	<Unknown> + <Negative>	Registro B
USD5 + .	<Currency> + <>	Registro B
USD5 + missing	<Currency> + <Negative>	Registro A

Qué reglas de categoría pueden coincidir

La siguiente tabla contiene algunas sintaxis que se pueden introducir en el editor de reglas de categoría. No todas las reglas funcionan aquí y no todas coinciden con los mismos registros. Vea cómo afectan las diferentes sintaxis a los registros coincidentes.

Tabla 25. Reglas de muestra	
Sintaxis de la regla	Resultado
USD5 & missing	Coincide con los registros A y B, ya que ambos contienen el concepto extraído missing y el concepto extraído USD5. Esto equivale a: (USD5 & missing)

Tabla 25. Reglas de muestra (continuación)

Sintaxis de la regla	Resultado
missing & USD5	Coincide con los registros A y B, ya que ambos contienen el concepto extraído missing y el concepto extraído USD5. Esto equivale a: (missing & USD5)
missing & <Currency>	Coincide con los registros A y B ya que ambos contienen el concepto extraído missing y un concepto que coincide con el tipo <Currency>. Esto equivale a: (missing & <Currency>)
<Currency> & missing	Coincide con los registros A y B ya que ambos contienen el concepto extraído missing y un concepto que coincide con el tipo <Currency>. Esto equivale a: (<Currency> & missing)
[USD5 + missing]	Coincide con A pero no B ya que el registro B no produjo ninguna salida de patrón TLA que contenga USD5 + missing (consulte la tabla anterior). Esto es equivalente al resultado de patrón TLA: USD5 + missing
[missing + USD5]	No coincide con el registro A ni B ya que ningún patrón TLA extraído (consulte la tabla anterior) coincide con el orden expresado aquí con missing en la primera posición. Esto es equivalente al resultado de patrón TLA: USD5 + missing
[missing & USD5]	Coincide con A pero no B ya que no se ha extraído ningún patrón TLA del registro B. La utilización del carácter & indica que el orden no es importante al coincidir; por lo tanto, esta regla busca una coincidencia de patrón con [missing + USD5] o [USD5 + missing]. Solo [USD5 + missing] del registro A tiene una coincidencia.
[missing + <Currency>]	No coincide ni con el registro A ni con B ya que ningún patrón TLA extraído coincide con este orden. Esto no tiene equivalente, ya que una salida de TLA solo se basa en términos (USD5 + missing) o en tipos (<Currency> + <Negative>), pero no mezcla conceptos y tipos.
[<Currency> + <Negative>]	Coincide con el registro A pero no B, ya que no se ha extraído ningún patrón TLA del registro B. Esto es equivalente a la salida TLA: <Currency> + <Negative>

Tabla 25. Reglas de muestra (continuación)	
Sintaxis de la regla	Resultado
[<Negative> + <Currency>]	No coincide ni con el registro A ni con B ya que ningún patrón TLA extraído coincide con este orden. En la plantilla Opinions, de forma predeterminada, cuando se encuentra un <i>tema</i> con una <i>opinión</i> , el <i>tema</i> (<Currency>) ocupa la primera posición de la ranura y la <i>opinión</i> (<Negative>) ocupa la segunda posición de la ranura.

Creación de reglas de categoría

Cuando cree o edite una regla, deberá tenerla abierta en un editor de reglas. Puede añadir conceptos, tipos o patrones y utilizar comodines para ampliar las coincidencias. Cuando se utilizan conceptos, tipos y patrones reconocidos, tiene la ventaja de que encontrará todos los conceptos relacionados. Por ejemplo, cuando utiliza un concepto, todos sus términos asociados, formas plurales y sinónimos también se asocian con la regla. De la misma forma, cuando utiliza un tipo, la regla también captura todos sus conceptos.

Puede abrir el editor de reglas editando una regla existente o pulsando el botón derecho del ratón en el nombre de la categoría y eligiendo **Crear regla**.

Puede utilizar menús contextuales, la acción de arrastrar y soltar o especificar manualmente conceptos, tipos y patrones en el editor. A continuación, combine estos con operadores booleanos (&, !(), |) y paréntesis para formar sus expresiones de regla. Para evitar errores comunes, se recomienda arrastrar los conceptos directamente desde el panel Resultados extraídos o desde el panel Datos y soltarlos en el editor de reglas. Preste especial atención en la sintaxis de las reglas para evitar errores. Para obtener más información, consulte el tema [“Sintaxis de regla de categoría”](#) en la página 120.

Nota: Para ver ejemplos sobre cómo coinciden las reglas con el texto, consulte [“Ejemplos de reglas de categoría”](#) en la página 127.

Para crear una regla

1. Si no ha extraído aún ningún dato o su extracción está fuera de la fecha, hágalo ahora. Para obtener más información, consulte el tema [“Extraer datos”](#) en la página 82.

Nota: Si filtra una extracción de forma que ya no hayan conceptos visibles, se muestra un mensaje de error cuando intenta crear o editar una regla de categoría. Para evitar esto, modifique su filtro de extracción para que los conceptos estén disponibles.

2. En el panel Categorías, seleccione la categoría en la que desea añadir una regla.
3. En los menús, seleccione **Categorías > Crear regla**. El panel del editor de reglas de categoría se abre en la ventana.
4. En la columna Nombre de regla, escriba un nombre para la regla. Si no proporciona un nombre, se utilizará automáticamente la expresión como nombre. Puede cambiar el nombre de la regla en otro momento.
5. En el campo de texto de expresión más largo puede:
 - Escribir texto directamente en el campo o arrastrarlo desde otro panel y soltarlo aquí. Utilice solo conceptos extraídos, tipos y patrones. Por ejemplo, si especifica la palabra *cats* pero solo la forma singular, *cat*, aparece en el panel Resultados de extracción, el editor no podrá reconocer *cats*. En este último caso, el singular puede incluir automáticamente el plural, de lo contrario deberá utilizar un comodín. Para obtener más información, consulte el tema [“Sintaxis de regla de categoría”](#) en la página 120.
 - Seleccione los conceptos, tipos o patrones que desee añadir a las reglas y utilice los menús.
 - Añada operadores booleanos para enlazar los elementos de la regla. Utilice los botones de la barra de herramientas para añadir el operador booleano "y" **&**, el operador booleano "o" **|**, el operador booleano "no" **!**, paréntesis **()** y corchetes para los patrones **[]** a su regla.

6. Pulse en el botón **Probar regla** para verificar que la regla tiene un formato correcto. Para obtener más información, consulte el tema [“Sintaxis de regla de categoría”](#) en la página 120. El número de documentos o registros encontrados aparece entre paréntesis junto al texto **Resultado de prueba**. A la derecha de este texto puede ver los elementos de la regla que se han reconocido o posibles mensajes de error. Si el gráfico junto al tipo, patrón o concepto aparece con un signo de interrogación rojo, significa que el elemento no coincide con ninguna extracción conocida. Si no coincide, la regla no encontrará ningún registro.
7. Para probar una parte de la regla, seleccione dicha parte y pulse en **Probar selección**.
8. Realice los cambios necesarios y vuelva a probar la regla si encontró problemas.
9. Cuando haya finalizado, pulse **Guardar y cerrar** para volver a guardar la regla y cerrar el editor. El nuevo nombre de la regla aparece en la categoría.

Edición y eliminación de reglas

Después de crear y guardar una regla, puede editarla en cualquier momento. Para obtener más información, consulte el tema [“Sintaxis de regla de categoría”](#) en la página 120.

Si ya no desea una regla, puede eliminarla.

Para editar reglas

1. En la tabla Descriptores del cuadro de diálogo Definiciones de categoría, seleccione la regla.
2. En los menús, seleccione **Categorías > Editar regla** o efectúe una doble pulsación en el nombre de la regla. El editor se abre con la regla seleccionada.
3. Realice cambios en la regla utilizando los resultados extraídos y los botones de la barra de herramientas.
4. Vuelva a probar la regla para asegurarse de que arroja los resultados esperados.
5. Pulse **Guardar y cerrar** para volver a guardar la regla y cerrar el editor.

Para eliminar una regla

1. En la tabla Descriptores del cuadro de diálogo Definiciones de categoría, seleccione la regla.
2. En los menús, seleccione **Editar > Suprimir**. La regla se elimina de la categoría.

Importación y exportación de categorías predefinidas

Si tiene sus propias categorías almacenadas en un archivo de Microsoft Excel (*.xls, *.xlsx), puede importarlas a IBM SPSS Modeler Text Analytics .

También puede exportar las categorías que tenga en un sesión de área de trabajo interactiva abierto a un archivo de Microsoft Excel (*.xls, *.xlsx). Al exportar sus categorías, puede decidir incluir o excluir información adicional como descriptores y puntuaciones. Para obtener más información, consulte el tema [“Exportación de categorías”](#) en la página 134.

Si las categorías predefinidas no tienen códigos o desea códigos nuevos, puede generar automáticamente un nuevo conjunto de códigos para el conjunto de categorías en el panel de categorías seleccionando **Categorías > Gestionar categorías > Generar códigos automáticamente** en los menús. Esto eliminará los códigos existentes y volverá a numerarlos a todos automáticamente.

Importación de categorías predefinidas

Puede importar sus categorías predefinidas en IBM SPSS Modeler Text Analytics . Antes de la importación, asegúrese de que el archivo de categoría predefinida se encuentre en un archivo Microsoft Excel (*.xls, *.xlsx) y esté estructurado en uno de los formatos compatibles. También puede hacer que el producto detecte automáticamente el formato en su lugar. Se da soporte a los formatos siguientes:

- **Formato de lista sin formato:** Consulte el tema [“Formato de lista plana”](#) en la página 132 para obtener más información.

- **Formato compacto:** Consulte el tema [“Formato compacto”](#) en la página 132 para obtener más información.
- **Formato de sangría:** Consulte el tema [“Formato con sangrado”](#) en la página 133 para obtener más información.

Para importar categorías predefinidas

1. En los menús del entorno de trabajo interactivo, seleccione **Categorías > Gestionar categorías > Importar categorías predefinidas**. Aparecerá un asistente de importación de categorías predefinidas.
2. En la lista desplegable Buscar en, seleccione la unidad y la carpeta donde se encuentra el archivo.
3. Seleccione el archivo de la lista. El nombre del archivo aparece en el cuadro de texto Nombre de archivo.
4. Seleccione la hoja de cálculo que contenga las categorías predefinidas de la lista. El nombre de la hoja de cálculo aparece en el campo Hoja de cálculo.
5. Para comenzar a elegir el formato de datos, pulse **Siguiente**.
6. Seleccione el formato de su archivo o la opción para permitir que el producto intente detectar el formato automáticamente. La detección automática funciona mejor con los formatos más comunes.
 - **Formato de lista sin formato:** Consulte el tema [“Formato de lista plana”](#) en la página 132 para obtener más información.
 - **Formato compacto:** Consulte el tema [“Formato compacto”](#) en la página 132 para obtener más información.
 - **Formato de sangría:** Consulte el tema [“Formato con sangrado”](#) en la página 133 para obtener más información.
7. Para definir las opciones de importación adicionales, pulse **Siguiente**. Si decide que se detecte el formato automáticamente, se le llevará al paso final.
8. Si una o más filas contienen cabeceras de columna u otra información externa, seleccione el número de fila desde el que desee comenzar la importación en la opción **Comenzar importación en la fila**. Por ejemplo, si sus nombres de categoría empiezan en la fila 7, debe introducir el número 7 en esta opción para que el archivo se importe correctamente.
9. Si su archivo contiene códigos de categorías, seleccione la opción **Contiene códigos de categorías**. De este modo ayuda a que el asistente reconozca sus datos correctamente.
10. Revise las celdas codificadas por colores y la leyenda para asegurarse de que los datos se han identificado correctamente. Los errores detectados en el archivo se muestran de color rojo y se hace referencia a ellos debajo de la tabla de presentación preliminar del formato. Si se ha seleccionado un formato incorrecto, retroceda y seleccione otro formato. Si necesita realizar correcciones en su archivo, haga dichos cambios y reinicie el asistente volviendo a seleccionar el archivo. Debe corregir todos los errores antes de poder finalizar el asistente.
11. Para revisar el conjunto de categorías y subcategorías que se importarán y para definir cómo crear descriptores para estas categorías, pulse **Siguiente**.
12. Revise el conjunto de categorías que se importará a la tabla. Si no ve las palabras clave que esperaba ver como descriptores, puede que no se hayan reconocido durante la importación. Asegúrese de que tienen los prefijos adecuados y de que aparecen en la casilla correcta.
13. Elija cómo desea manejar las categorías preexistentes en su sesión.
 - **Sustituir todas las categorías existentes.** Esta opción depura todas las categorías existentes y, a continuación, las categorías recién importadas se utilizan en solitario en su lugar.
 - **Agregar a las categorías existentes.** Esta opción importará las categorías y fusionará las categorías comunes con las categorías existentes. Cuando esté añadiendo categorías a categorías existentes, debe determinar cómo desea tratar los duplicados. Una opción (opción: **Fusionar**) es fusionar las categorías que se estén importando con las categorías existentes si comparten el nombre de categoría. Otra opción (opción: **Excluir de la importación**) es prohibir la importación de categorías si existe una con el mismo nombre.

14. **Importar palabras clave como descriptores** es una opción para importar las palabras clave identificadas en sus datos como descriptores para la categoría asociada.
15. **Ampliar categorías mediante la derivación de descriptores** es una opción que generará descriptores a partir de las palabras que representan el nombre de la categoría o subcategoría y/o las palabras que componen la anotación. Si las palabras coinciden con resultados extraídos, entonces se añaden como descriptores a la categoría. Esta opción genera los mejores resultados cuando los nombres de categoría o anotaciones son largos y descriptivos. Se trata de un método rápido para generar los descriptores de categorías que permiten que la categoría capture registros que contengan dichos descriptores.
 - El campo **De** le permite seleccionar de qué texto se derivarán los descriptores, los nombres o categorías y subcategorías, las palabras de las anotaciones, o ambas.
 - El campo **Como** le permite crear estos descriptores en forma de conceptos o patrones TLA. Si la extracción de TLA no ha tenido lugar, las opciones de **patrones** están desactivadas en este asistente.
16. Para importar las categorías predefinidas en el panel Categorías, pulse **Finalizar**.

Formato de lista plana

En el formato de lista sin formato, sólo hay un nivel superior de categorías sin ninguna jerarquía, lo que significa que no hay subcategorías o subredes. Los nombres de categorías están en una única columna.

La siguiente información puede incluirse en un archivo con este formato:

- La columna opcional **códigos** contiene valores numéricos que identifican de manera exclusiva cada categoría. Si especifica que el archivo de datos contiene códigos (opción **Contiene códigos de categorías** en el paso **Configuración de contenido**), entonces debe existir una columna que contenga códigos únicos para cada categoría en la celda justo a la izquierda del nombre de categoría. Si los datos no contienen códigos, pero desea crear algunos códigos más adelante, siempre puede generar códigos más tarde (**Categorías > Gestionar categorías > Generar códigos automáticamente**).
- Una columna de **nombres de categoría** *necesaria* contiene todos los nombres de las categorías. Esta columna es obligatoria para importar mediante este formato.
- **Anotaciones** opcionales en la casilla justo a la derecha del nombre de categoría. Esta anotación consiste en texto que describe sus categorías/subcategorías.
- Se pueden importar **palabras clave** opcionales como descriptores para categorías. Para ser reconocidas, estas palabras clave deben existir en la celda directamente debajo del nombre de categoría/subcategoría asociado y la lista de palabras clave debe tener como prefijo el carácter de subrayado (_), como por ejemplo `_firearms`, `weapons` / `guns`. La celda de palabras clave puede contener una o más palabras utilizadas para describir cada categoría. Estas palabras se importarán como descriptores o se ignorarán dependiendo de lo que especifique en el último paso del asistente. Posteriormente, los descriptores se comparan con los resultados extraídos del texto. Si se encuentra una coincidencia, entonces ese registro o documento se puntúa en la categoría que contiene este descriptor.

Tabla 26. Formato de lista plana con códigos, palabras clave y anotaciones

Columna A	Columna B	Columna C
Código de categoría (opcional)	Nombre de categoría	Anotación
	Lista de _descriptores/palabras clave (opcional)	

Formato compacto

El formato compacto se estructura de forma parecida al formato de lista plana excepto en que el formato compacto se utiliza con categorías jerárquicas. Por lo tanto, se necesita una columna de nivel de código para definir el nivel jerárquico de cada categoría y subcategoría.

La siguiente información puede incluirse en un archivo con este formato:

- La columna *necesaria* **nivel de código** contiene números que indican la posición jerárquica de la información siguiente de dicha línea. Por ejemplo, si se especifican los valores 1, 2 ó 3 y tiene tanto categorías como subcategorías, entonces 1 es para categorías, 2 para subcategorías y 3 para sub-subcategorías. Si sólo tiene categorías y subcategorías, entonces 1 es para categorías y 2 para subcategorías. Así sucesivamente, hasta la profundidad de categoría deseada.
- La columna opcional **códigos** contiene valores que identifican de manera exclusiva cada categoría. Si especifica que el archivo de datos contiene códigos (opción **Contiene códigos de categorías** en el paso **Configuración de contenido**), entonces debe existir una columna que contenga códigos únicos para cada categoría en la celda justo a la izquierda del nombre de categoría. Si los datos no contienen códigos, pero desea crear algunos códigos más adelante, siempre puede generar códigos más tarde (**Categorías > Gestionar categorías > Generar códigos automáticamente**).
- Una columna de **nombres de categoría** *necesaria* contiene todos los nombres de las categorías y subcategorías. Esta columna es obligatoria para importar mediante este formato.
- **Anotaciones** opcionales en la casilla justo a la derecha del nombre de categoría. Esta anotación consiste en texto que describe sus categorías/subcategorías.
- Se pueden importar **palabras clave** opcionales como descriptores para categorías. Para ser reconocidas, estas palabras clave deben existir en la celda directamente debajo del nombre de categoría/subcategoría asociado y la lista de palabras clave debe tener como prefijo el carácter de subrayado (_), como por ejemplo `_firearms`, `weapons` / `guns`. La celda de palabras clave puede contener una o más palabras utilizadas para describir cada categoría. Estas palabras se importarán como descriptores o se ignorarán dependiendo de lo que especifique en el último paso del asistente. Posteriormente, los descriptores se comparan con los resultados extraídos del texto. Si se encuentra una coincidencia, entonces ese registro o documento se puntúa en la categoría que contiene este descriptor.

Tabla 27. Ejemplo de formato compacto con códigos

Columna A	Columna B	Columna C
Nivel de código jerárquico	Código de categoría (<i>opcional</i>)	Nombre de categoría
Nivel de código jerárquico	Código de subcategoría (<i>opcional</i>)	Nombre de subcategoría

Tabla 28. Ejemplo de formato compacto sin códigos

Columna A	Columna B
Nivel de código jerárquico	Nombre de categoría
Nivel de código jerárquico	Nombre de subcategoría

Formato con sangrado

En el formato de archivo con sangrado, el contenido es jerárquico, lo que significa que contiene categorías y uno o más niveles de subcategorías. Además, su estructura tiene sangrado para indicar esta jerarquía. Cada fila del archivo contiene una categoría o una subcategoría, pero las subcategorías tienen un sangrado con respecto a las categorías, las sub-subcategorías tienen un sangrado con respecto a las subcategorías y así sucesivamente. Puede crear manualmente esta estructura en Microsoft Excel o utilizar una que se haya exportado desde otro producto y se haya guardado en un formato Microsoft Excel.

- **Los códigos de categoría de nivel superior y nombres de categoría** ocupan las columnas A y B, respectivamente. No obstante, si no hay ningún código, el nombre de categoría está en la columna A.
- **Los códigos de subcategoría y los nombres de subcategoría** ocupan las columnas B y C, respectivamente. O, si no hay códigos presentes, el nombre de la subcategoría se encuentra en la columna B. La subcategoría es miembro de una categoría. No puede tener subcategorías si no tiene categorías de nivel superior.

Tabla 29. Estructura sangrada con códigos			
Columna A	Columna B	Columna C	Columna D
Código de categoría (opcional)	Nombre de categoría		
	Código de subcategoría (opcional)	Nombre de subcategoría	
		Código de sub-subcategoría (opcional)	Nombre de sub-subcategoría

Tabla 30. Estructura sangrada sin códigos		
Columna A	Columna B	Columna C
Nombre de categoría		
	Nombre de subcategoría	
		Nombre de sub-subcategoría

La siguiente información puede incluirse en un archivo con este formato:

- Los **códigos** opcionales deben ser valores que identifiquen de manera exclusiva cada categoría o subcategoría. Si especifica que el archivo de datos contiene códigos (opción **Contiene códigos de categorías** en el paso **Configuración de contenido**), entonces debe existir un código único para cada categoría o subcategoría en la celda justo a la izquierda del nombre de categoría/subcategoría. Si los datos no contienen códigos, pero desea crear algunos códigos más adelante, siempre puede generar códigos más tarde (**Categorías > Gestionar categorías > Generar códigos automáticamente**).
- Un **nombre necesario** para cada categoría y subcategoría. Las subcategorías deben tener un sangrado con respecto a las categorías de una casilla a la derecha en una fila separada.
- **Anotaciones** opcionales en la casilla justo a la derecha del nombre de categoría. Esta anotación consiste en texto que describe sus categorías/subcategorías.
- Se pueden importar **palabras clave** opcionales como descriptores para categorías. Para ser reconocidas, estas palabras clave deben existir en la celda directamente debajo del nombre de categoría/subcategoría asociado y la lista de palabras clave debe tener como prefijo el carácter de subrayado (_), como por ejemplo `_firearms`, `weapons` / `guns`. La celda de palabras clave puede contener una o más palabras utilizadas para describir cada categoría. Estas palabras se importarán como descriptores o se ignorarán dependiendo de lo que especifique en el último paso del asistente. Posteriormente, los descriptores se comparan con los resultados extraídos del texto. Si se encuentra una coincidencia, entonces ese registro o documento se puntúa en la categoría que contiene este descriptor.

Importante: Si utiliza un código a un nivel, debe incluir un código para cada categoría y subcategoría. De lo contrario, el proceso de importación fallará.

Exportación de categorías

También puede exportar las categorías que tenga en un sesión de área de trabajo interactiva abierto a un formato de archivo de Microsoft Excel (*.xls, *.xlsx). Los datos que se exportarán provienen en gran medida del contenido actual del panel Categorías o de las propiedades de categoría. Por lo tanto, le recomendamos que vuelva a puntuar si tiene previsto exportar también el valor de puntuación de **Docs.** .

Tabla 31. Opciones de exportación de categorías

Siempre se exporta...	Se exporta opcionalmente...
• Códigos de categorías, si los hay • Nombres de categoría (y subcategoría) • Niveles de código, si los hay (formato <i>plano/compacto</i>) • Cabeceras de columna (formato <i>plano/compacto</i>)	• Puntuaciones de Docs. :NONE. • Anotaciones de categoría • Nombres de descriptor • Recuentos de descriptores

Importante: Cuando exporta descriptores, se convierten en cadenas de texto y se les añade un guión bajo como prefijo. Si vuelve a realizar una importación en este producto, se perderá la capacidad de distinguir entre descriptores que sean patrones, los que sean reglas de categoría y los que sean conceptos planos. Si tiene la intención de volver a utilizar estas categorías en este producto, recomendamos encarecidamente que cree un archivo de paquete de análisis de texto (TAP) en su lugar, ya que el formato TAP conservará todos los descriptores del modo en que estén definidos actualmente, así como todas sus categorías, códigos y los recursos lingüísticos utilizados. Los archivos TAP pueden utilizarse tanto en IBM SPSS Modeler Text Analytics como en IBM SPSS Text Analytics for Surveys. Para obtener más información, consulte el tema [“Uso de los paquetes de análisis de texto”](#) en la página 135.

Para exportar categorías predefinidas

1. En los menús del entorno de trabajo interactivo, seleccione **Categorías > Gestionar categorías > Exportar categorías**. Aparecerá un asistente de exportación de categorías.
2. Seleccione la ubicación e introduzca el nombre del archivo que se exportará.
3. Introduzca un nombre para el archivo de salida en el cuadro de texto Nombre de archivo.
4. Para elegir el formato en el que va a exportar sus datos de categoría, pulse **Siguiente**.
5. Seleccione el formato entre las siguientes alternativas:
 - **Formato de lista sin formato o compacta:** Consulte el tema [“Formato de lista plana”](#) en la página 132 para obtener más información. La lista sin formato no contiene subcategorías. Para obtener más información, consulte el tema [“Formato compacto”](#) en la página 132. El formato de lista compacta contiene categorías jerárquicas.
 - **Formato de sangría:** Consulte el tema [“Formato con sangrado”](#) en la página 133 para obtener más información.
6. Para comenzar a elegir el contenido a exportar y para revisar los datos propuestos, pulse **Siguiente**.
7. Revise el contenido del archivo exportado.
8. Seleccione o elimine la selección de la configuración de contenido adicional que se exportará como **Anotaciones** o **Nombres de descriptor**.
9. Para exportar las categorías, pulse **Finalizar**.

Uso de los paquetes de análisis de texto

Un paquete de análisis de texto, también llamado TAP, actúa como plantilla para la categorización de las respuestas del texto. La utilización de un TAP es una forma sencilla de categorizar sus datos de texto con una intervención mínima ya que contiene el conjuntos de categorías creados previamente y los recursos lingüísticos necesarios poder para codificar un gran número de registros rápida y automáticamente. Mediante el uso de los recursos lingüísticos, los datos de texto se analizan y se realiza en ellos el proceso de minería para extraer los conceptos clave. Basándose en los conceptos clave y en los patrones que se encuentran en el texto, los registros pueden categorizarse en el conjunto de categorías que seleccionó en el TAP. Puede crear su propio TAP o actualizar uno.

Un TAP está compuesto por los elementos siguientes:

- **Conjuntos de categorías.** Un conjunto de categorías se compone fundamentalmente de categorías predefinidas, códigos de categorías, descriptores para cada categoría y, por último, un nombre para todo el conjunto de categorías. Los descriptores son elementos lingüísticos (conceptos, tipos, patrones

y reglas) como el término *barato* o el patrón *buen precio*. Los descriptores se utilizan para definir una categoría de manera que cuando el texto coincide con un descriptor de categoría, el documento o registro se coloca en la categoría.

- **Recursos lingüísticos.** Los recursos lingüísticos son un conjunto de bibliotecas y recursos avanzados que se ajustan para extraer patrones y conceptos clave. Estos conceptos y patrones de extracción, a su vez, se utilizan como los descriptores que permiten a los registros colocarse en una categoría del conjunto de categorías.

Puede crear su propio TAP, actualizar uno, o cargar paquetes de análisis de texto.

Después de seleccionar el TAP y elegir un conjunto de categorías, SPSS Modeler Text Analytics puede extraer y categorizar los registros.

Nota: Los TAP pueden crearse y utilizarse de manera intercambiable entre SPSS Text Analytics for Surveys y SPSS Modeler Text Analytics . Sin embargo, tenga en cuenta que la puntuación sobre reglas puede ser diferente en SPSS Modeler Text Analytics dependiendo de si se carga un paquete de análisis de texto (TAP) desde SPSS Modeler Text Analytics directamente, o si se carga un TAP desde IBM SPSS Text Analytics for Surveys. Se recomienda utilizar los TAP que se lleven a cabo dentro de SPSS Modeler Text Analytics ; esto es porque los TAP que se lleven a cabo en IBM SPSS Text Analytics for Surveys puede crearse mediante una versión diferente de los recursos lingüísticos.

Creación de paquetes de análisis de texto

Siempre que tenga un sesión con al menos una categoría y algunos recursos, puede crear un paquete de análisis de texto (TAP) a partir del contenido del sesión de área de trabajo interactiva abierto. El conjunto de categorías y descripciones (conceptos, tipos, reglas o resultados de patrones TLA) pueden colocarse en TAP junto con todos los recursos lingüísticos abiertos en el Editor de recursos.

Puede ver el idioma para el que se crearon los recursos. El idioma se establece en la pestaña Recursos avanzados de Editor de plantillas o Editor de recursos.

Para crear un paquete de análisis de texto

1. En los menús, seleccione **Archivo > Paquetes de análisis de texto > Realizar paquete**. Aparecerá el diálogo Realizar paquete.
2. Busque el directorio en el que se ha guardado el TAP. De forma predeterminada, los TAP se guardan en el subdirectorio \TAP del directorio de instalación del producto.
3. Introduzca un nombre para el TAP en el campo **Nombre de archivo**.
4. Introduzca una etiqueta en el campo **Etiqueta de paquete**. Cuando especifica un nombre de archivo, este nombre automáticamente aparece como la etiqueta, pero puede cambiarla si lo desea.
5. Para excluir un conjunto de categorías del TAP, quite la marca de la casilla de verificación **Incluir**. Al hacerlo se asegura de que no se añadirá al paquete. De forma predeterminada, en el TAP se incluye un conjunto de categorías por pregunta. Siempre debe haber al menos un conjunto de categorías en el TAP.
6. Cambie el nombre de los conjuntos de categorías. La columna **Conjunto(s) de categorías nuevas** contiene nombres genéricos de forma predeterminada, que se generan al añadir el prefijo Cat_ en el nombre de la variable de texto. Con solo pulsar con el ratón en la celda podrá editar el nombre. Pulse en cualquier otro lugar para que el cambio de nombre surta efecto. Si renombra un conjunto de categorías, el nombre sólo cambia en el TAP y no cambia el nombre de variable en el sesión abierto.
7. Cambie el orden de los conjuntos de categorías mediante las teclas de flecha a la derecha de la tabla de conjuntos de categorías.
8. Pulse **Guardar** para crear el paquete de análisis de texto. El cuadro de diálogo se cierra.

Carga de los paquetes de análisis de texto

Cuando se configura un nodo de modelado de Text Mining, debe especificar los recursos que se utilizarán durante la extracción. En lugar de elegir una plantilla de recursos, puede seleccionar un paquete de análisis de texto (TAP) o un proyecto de SPSS Text Analytics for Surveys (.tas) para copiar no solo sus

recursos, sino también un conjunto de categorías en el nodo. Si selecciona un archivo .tas, se convertirá en un TAP.

Los TAP resultan más interesantes cuando se crea interactivamente un modelo de categoría, ya que puede utilizar el conjunto de categorías como punto de partida para la categorización. Cuando se ejecuta la ruta, la sesión del área de trabajo interactiva se inicia y el juego de categorías aparece en el panel Categorías. De esta forma, los documentos y los registros se puntúan inmediatamente utilizando estas categorías, y luego puede seguir refinando, generando y ampliando dichas categorías hasta que cumplan sus expectativas. Si desea más información, consulte [“Métodos y estrategias para crear categorías”](#) en la página 98.

A partir de la versión 14, podrá ver además el idioma para el que se definieron los recursos en el TAP cuando pulse en **Cargar** y seleccione el TAP.

Cargar un TAP o un TAS

1. Edite el nodo de modelado Text Mining.
2. En la pestaña Modelos, seleccione *Paquete de análisis de texto* en la sección **Copiar recursos de**.
3. Pulse **Cargar**. Aparece el diálogo Cargar Paquete de análisis de texto.
4. Examine la ubicación del TAP o el proyecto de SPSS Text Analytics for Surveys (.tas) que contiene los recursos y el conjunto de categorías que desea copiar en el nodo. De forma predeterminada, se guardan en el subdirectorío \TAP del directorío de instalación del producto.
5. Introduzca un nombre para el TAP en el campo **Nombre de archivo**. La etiqueta se mostrará automáticamente.
6. Seleccione el conjunto de categorías que desee utilizar. Se trata del conjunto de categorías que aparecerá en la sesión del entorno de trabajo interactivo. Luego puede ajustar y mejorar estas categorías manualmente o bien utilizando las opciones Generar categorías o Ampliar categorías.
7. Pulse **Cargar** para copiar el contenido del paquete de análisis de texto o el proyecto de SPSS Text Analytics for Surveys en el nodo. El cuadro de diálogo se cierra. Cuando se carga el contenido, se copia en el nodo; por lo tanto, cualquier cambio que realice en los recursos y en las categorías no se verá reflejado a menos que lo actualice explícitamente y lo vuelva a cargar.

Actualización de los paquetes de análisis de texto

Si realiza mejoras en un conjunto de categorías, en los recursos lingüísticos o crea un conjunto de categorías completamente nuevo, puede actualizar un paquete de análisis de texto (TAP) para que pueda volver a utilizar estas mejoras en otro momento. Para ello, debe estar en el sesión abierto que contiene la información que desea colocar en el TAP. Al actualizar puede optar por agregar conjuntos de categorías, sustituir recursos, cambiar la etiqueta del paquete o cambiar el nombre o el orden de los conjuntos de categorías.

Actualizar un paquete de análisis de texto

1. En los menús, seleccione **Archivo > Paquetes de análisis de texto > Actualizar paquete**. Aparece el diálogo Actualizar paquete.
2. Busque el directorío que contiene el paquete de análisis de texto que desea actualizar.
3. Introduzca un nombre para el TAP en el campo **Nombre de archivo**.
4. Para sustituir los recursos lingüísticos dentro del TAP con los que se encuentran en el sesión actual, seleccione la opción **Sustituir los recursos de este paquete con los que están en la sesión abierta**. Por lo general conviene actualizar los recursos lingüísticos porque se utilizaron para extraer los conceptos clave y los patrones empleados para crear las definiciones de categorías. El hecho de tener los recursos lingüísticos más recientes, garantiza que obtenga los mejores resultados para categorizar sus registros. Si no selecciona esta opción, los recursos lingüísticos que ya estaban en el paquete se mantienen sin cambios.

5. Para actualizar únicamente los recursos lingüísticos, asegúrese de que selecciona la opción **Sustituir los recursos de este paquete por los recursos de la sesión abierta** y elige solo los conjuntos de categorías actuales que ya se encontraban en el TAP.
6. Para incluir el nuevo conjunto de categorías desde el sesión abierto en el TAP, seleccione el recuadro de selección para cada conjunto de categorías que se vaya a añadir. Puede añadir uno o varios conjuntos de categorías, o ninguno.
7. Para eliminar conjuntos de categorías del TAP, quite la marca de la casilla de verificación **Incluir**. Puede optar por eliminar un conjunto de categorías que ya estuviera en el TAP, puesto que está añadiendo uno mejorado. Para ello, quite la marca de verificación de la casilla **Incluir** del conjunto de categorías correspondiente en la columna Conjunto(s) de categorías actual(es). Siempre debe haber al menos un conjunto de categorías en el TAP.
8. Cambie el nombre de los conjuntos de categorías, si procede. Con solo pulsar con el ratón en la celda podrá editar el nombre. Pulse en cualquier otro lugar para que el cambio de nombre surta efecto. Si renombra un conjunto de categorías, el nombre sólo cambia en el TAP y no cambia el nombre de variable en el sesión abierto. Si hay dos conjuntos de categorías con el mismo nombre, los nombres aparecerán en rojo hasta que corrija la duplicación.
9. Para crear un paquete nuevo con los contenidos de la sesión fusionados con los contenidos del TAP seleccionado, pulse en **Guardar como nuevo**. Aparece el diálogo Guardar como Paquete de análisis de texto. Consulte las instrucciones siguientes.
10. Pulse **Actualizar** para guardar los cambios realizados en el TAP seleccionado.

Guardar un paquete de análisis de texto

1. Busque el directorio en el que se ha guardado el archivo TAP. De forma predeterminada, los archivos TAP se guardan en el subdirectorio \TAP del directorio de instalación.
2. Introduzca un nombre para el archivo TAP en el campo **Nombre de archivo**.
3. Introduzca una etiqueta en el campo **Etiqueta de paquete**. Cuando especifica un nombre de archivo, automáticamente se utiliza este nombre también como etiqueta. Sin embargo, puede cambiar el nombre de la etiqueta. Es necesario que tenga una etiqueta.
4. Pulse **Guardar** para crear el paquete nuevo.

Edición y refinamiento de categorías

Una vez creadas algunas categorías, querrá examinarlas y realizar en ellas algunos ajustes. Además de refinar los recursos lingüísticos, debe revisar las categorías y averiguar maneras de combinar o limpiar las definiciones, así como comprobar algunos de los documentos o registros categorizados. También puede revisar los documentos o registros de una categoría y realizar ajustes para que las categorías se definan de tal forma que puedan captarse todos los matices y distinciones.

Puede utilizar las técnicas automáticas de generación para crear las categorías; sin embargo, seguramente querrá realizar algunos ajustes en estas categorías. Después de utilizar una técnica o más, en la ventana aparecerán una serie de categorías nuevas. Luego puede revisar los datos de una categoría y realizar ajustes hasta que esté conforme con las definiciones de categoría. Para obtener más información, consulte el tema [“Acerca de las categorías”](#) en la página 102.

A continuación mostramos algunas opciones para refinar sus categorías, la mayoría de las cuales están descritas en las páginas siguientes:

Añadir descriptores a las categorías

Después de utilizar las técnicas automáticas, lo más probable es que aún tenga resultados extraídos que no se utilizaron en ninguna definición de categoría. Debe revisar esta lista en el panel Resultados extraídos. Si encuentra elementos que desearía mover a una categoría, puede añadirlos a una categoría nueva o existente.

Para añadir un concepto o un tipo a una categoría

1. Desde los paneles Datos y Resultados extraídos, seleccione los elementos que desea añadir a una categoría nueva o existente.
2. En los menús, seleccione **Categorías > Añadir a categoría**. El recuadro de diálogo Todas las categorías muestra el conjunto de categorías. Seleccione la categoría a la que desee añadir los elementos seleccionados. Si desea añadir los elementos a una categoría nueva, seleccione **Nueva categoría**. Aparece una categoría nueva en el panel Categorías con el nombre del primer elemento seleccionado.

Edición de descriptores de categoría

Una vez que haya creado algunas categorías, puede abrir cada una de ellas para ver todos los descriptores que conforman su definición. En el cuadro de diálogo Definiciones de categoría, puede realizar una serie de ediciones en los descriptores de categoría. Además, si se muestran las categorías en el árbol de categorías, puede también trabajar con ellos allí mismo.

Para editar una categoría

1. Seleccione la categoría que desea editar en el panel Categorías.
2. En los menús, seleccione **Ver > Definiciones de categoría**. Aparecerá el cuadro de diálogo Definiciones de categoría.
3. Seleccione el descriptor que desea editar y pulse en el botón correspondiente de la barra de herramientas.

La tabla siguiente describe cada botón de la barra de herramientas que puede utilizar para editar sus definiciones de categoría.

Tabla 32. Descripciones y botones de la barra de herramientas	
Iconos	Descripción
	Elimina los descriptores seleccionados de la categoría.
	Mueve los descriptores seleccionados a una categoría nueva o existente.
	Mueve los descriptores seleccionados en forma de una regla de categoría & a una categoría. Para obtener más información, consulte el tema “Uso de reglas de categoría” en la página 119 .
	Mueve cada uno de los descriptores seleccionados como su propia categoría nueva
 Visualización	Actualiza lo que se muestra en el panel Datos y en el panel Visualización en función de los descriptores seleccionados

Cómo mover categorías

Si desea colocar una categoría en otra categoría, o mover descriptores a otra categoría, puede moverla.

Para mover una categoría

1. En el panel Categorías, seleccione las categorías que desea mover a otra categoría.
 2. En los menús, seleccione **Categorías > Mover a categoría**. El menú muestra un conjunto de categorías, y la creada más recientemente figura al principio de la lista. Seleccione el nombre de la categoría a la que desea mover los conceptos seleccionados.
- Si puede ver el nombre que está buscando, selecciónelo; acto seguido los elementos seleccionados se añaden a dicha categoría.
 - Si no lo encuentra, seleccione **Más** para mostrar el cuadro de diálogo Todas las categorías, y seleccione la categoría de la lista que aparece.

Aplanamiento de categorías

Cuando dispone de una estructura jerárquica de categorías con categorías y subcategorías, puede aplanar la estructura. Cuando aplanar una categoría, todos los descriptores de las subcategorías de esa categoría se trasladan a la categoría seleccionada y todas las subcategorías ahora vacías se pierden. De esta forma, todos los documentos utilizados para coincidir con las subcategorías están ahora categorizados en la categoría seleccionada.

Para aplanar una categoría

1. En el panel Categorías, seleccione una categoría (nivel superior o subcategoría) que desea aplanar.
2. En los menús, seleccione **Categorías > Aplanar categorías**. Las subcategorías se trasladan y los descriptores se fusionan en la categoría seleccionada.

Fusión o combinación de categorías

Si desea combinar dos o más categorías en una categoría nueva, puede fusionarlas. Cuando se fusionan categorías, se crea una categoría nueva con un nombre genérico. Todos los conceptos, tipos y patrones que se utilizan en los descriptores de categorías se mueven a la nueva categoría. Más tarde podrá cambiar el nombre de la categoría editando sus propiedades.

Para fusionar una categoría o parte de una categoría

1. En el panel Categorías, seleccione los elementos que desea fusionar.
2. En los menús, seleccione **Categorías > Fusionar categorías**. Aparecerá el cuadro de diálogo Propiedades de categoría en el que debe introducir un nombre para la categoría recién creada. Las categorías seleccionadas se combinan en la nueva categoría como subcategorías.

Forzar los documentos de las categorías

Forzar la inclusión y exclusión de los documentos de las categorías le permite sustituir las definiciones de categoría creadas mediante las técnicas de creación automática de categorías sin cambiar la propia definición de la categoría. Puede darse el caso de que aunque el documento contenga términos que se utilizan para definir una categoría determinada, el propio documento no debería estar en dicha categoría. En este caso, puede forzar la exclusión del documento de dicha categoría sin tener que eliminar los términos de la definición de categoría.

Forzar es una operación que se utiliza en casos especiales en los que un documento se ajusta (o no se ajusta) a una categoría, pero por alguna razón u otra (por ejemplo, porque contiene un término determinado) está asignado (o no asignado) a esa categoría. Por ejemplo, esto puede suceder cuando un encuestado utiliza el sarcasmo en su respuesta, como *"La pizza estaba buenísima. No hay nadie a quien no le guste la pizza fría y quemada."* Supongamos que usted tenía una categoría llamada Pos: [`<Food> + <Positive>`] para capturar opiniones positivas con respecto a la comida que sirve un restaurante y esta respuesta está asignada a esa categoría. En este caso, probablemente quiera forzar la exclusión de respuesta de la categoría.

Para forzar la exclusión de las categorías

1. Desde el panel Datos, seleccione el documento que desea incluir o excluir de una categoría determinada.
2. Desde los menús, seleccione **Categorías > Forzar dentro** o bien **Categorías > Forzar fuera**. En un submenú aparecerá la lista de categorías que puede seleccionar.
3. Seleccione la categoría en la que desea incluir o de la que desea excluir este documento. Si ha creado muchas categorías, es posible que no todas aparezcan en el submenú.
 - En ese caso, elija **Más** en la parte inferior del submenú. Se abrirá el cuadro de diálogo Todas las categorías, en el que podrá seleccionar la categoría y pulsar en **Aceptar** para aplicar el cambio.
 - Si desea forzar la inclusión de un documento en una nueva categoría, seleccione **Crear categoría vacía**. Aparece una nueva categoría en el árbol de categorías con un nombre genérico.

Siempre que una categoría contiene uno o más documentos forzados, se muestra una pseudo-categoría denominada **Forzar dentro** o **Forzar fuera** debajo del nombre de la categoría en el árbol.

Borrar un estado forzado

1. Desde el panel Datos, seleccione el documento que ya no desea incluir o excluir de una categoría determinada.
2. En los menús, seleccione **Categorías > Forzar inclusión** para forzar la inclusión o seleccione **Categorías > Forzar exclusión** para forzar la exclusión. Las categorías de inclusión o exclusión del documento van precedidas por una marca de selección.
3. Seleccione la categoría en el submenú que está marcada y para la que desea eliminar la acción de forzar. La marca de selección se elimina y el documento deja de estar forzado.

Borrar todos los estados forzados

1. Desde el panel Datos, seleccione un registro que contenga una operación **Forzar dentro** o **Forzar fuera**.
2. Desde los menús, seleccione **Categorías > Borrar todo > Forzar dentro** o **Categorías > Borrar todo > Forzar fuera**. El estado forzado de los documentos se borra y se deja de forzar su inclusión o exclusión en las categorías.

Nota: esta característica solo está disponible si el texto de origen contiene un ID exclusivo. Si el texto de origen no tiene un ID exclusivo, puede añadir un nodo Derivar entre el documento de origen y el nodo Text Mining. Esta característica solo es efectiva cuando se ejecuta una sesión interactiva. Cuando se despliega el modelo de categoría para una puntuación no interactiva, esta parte de la información no se conserva ni utiliza puesto que está basada en un ID de documento.

Eliminación de categorías

Si ya no desea conservar una categoría, puede eliminarla.

Para eliminar una categoría

1. En el panel Categorías, seleccione la categoría o categorías que desea eliminar.
2. En los menús, seleccione **Editar > Suprimir**.

Capítulo 10. Análisis de clústeres

Puede construir y explorar clústeres de concepto en la vista Clústeres (**Ver > Clústeres**). Un *clúster* es una agrupación de conceptos relacionados generados por algoritmos de agrupación en clústeres basados en la frecuencia con que estos conceptos suceden en el conjunto de documentos/registros y la frecuencia con que aparecen juntos en el mismo documento, también conocido como *co-ocurrencias*. Cada concepto en un clúster co-aparece con al menos un concepto en el clúster. El objetivo de los clústeres es agrupar conceptos que se producen juntos, mientras que el objetivo de las categorías es agrupar documentos o registros basándose en la forma en que el texto que contienen coincide con los descriptores (conceptos, reglas, patrones) para cada categoría.

Un buen clúster es uno cuyos conceptos están fuertemente asociados y co-aparecen frecuentemente y con pocos enlaces a conceptos en otros clústeres. Cuando se trabaja con grandes conjuntos de datos, esta técnica puede provocar que los tiempos de espera sean significativamente más largos.

La agrupación en clústeres es un proceso que comienza por el análisis de un conjunto de conceptos y la búsqueda de conceptos que co-aparecen a menudo en los documentos. Dos conceptos que co-aparecen en un documento se consideran como un par de conceptos. A continuación, el proceso de agrupación en clústeres evalúa el *valor de similitud* de cada par de conceptos mediante la comparación del número de documentos en el que el par aparece junto, con el número de documentos en el que cada concepto aparece. Para obtener más información, consulte el tema [“Calcular valores de similitud de enlaces”](#) en la [página 145](#).

Por último, el proceso de agrupación en clústeres agrupa conceptos similares en clústeres mediante la agregación y toma en cuenta los valores de enlace y los valores definidos en el cuadro de diálogo Construir Clústeres. Por agregación, nos referimos a que los conceptos se añaden o que los clústeres pequeños se fusionan en uno más grande hasta que el clúster esté saturado. Un clúster está *saturado* cuando la fusión adicional de conceptos o clústeres más pequeños causa que el clúster supere los valores del cuadro de diálogo Construir clústeres (número de conceptos, enlaces externos o enlaces externos). Un clúster toma el nombre del concepto dentro del clúster que tenga el mayor número de enlaces que otros dentro del mismo clúster.

En conclusión, no todos los pares de conceptos terminan juntos en el mismo clúster ya que podría haber un enlace más fuerte en otro clúster o la saturación podría impedir la fusión de los clústeres en que se lleva a cabo. Por este motivo, hay enlaces externos e internos.

- *Enlaces internos* son enlaces entre pares de conceptos dentro de un clúster. No todos los conceptos en un clúster están vinculados unos a otros. Sin embargo, cada concepto está vinculado a al menos un concepto dentro del clúster.
- *Enlaces externos* son enlaces entre pares de conceptos en clústeres separados (un concepto dentro de un clúster y un concepto en otro clúster).

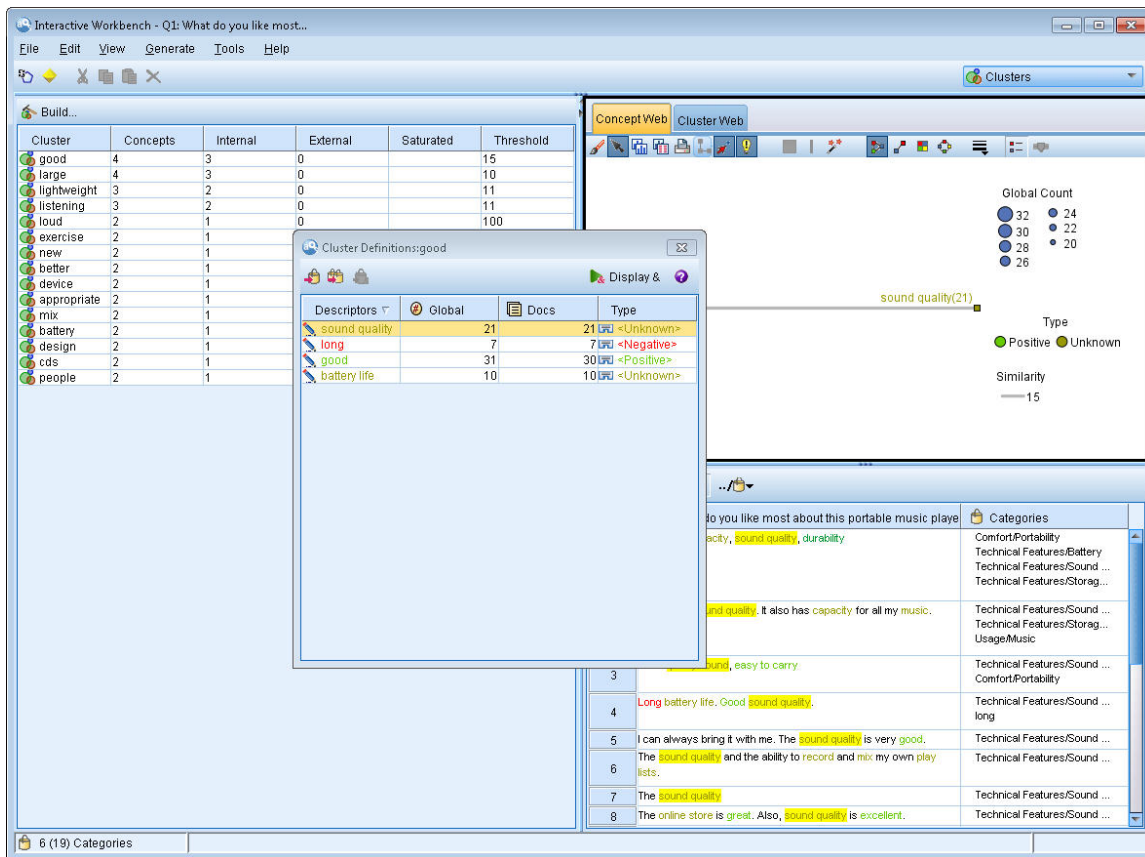


Figura 30. Vista Clústeres

La vista Clústeres se organiza en tres paneles; cada uno puede ocultarse o mostrarse mediante la selección del nombre desde la vista Menú:

- **Panel Clústeres** Puede construir y gestionar los clústeres en este panel. Para obtener más información, consulte el tema “Exploración de los clústeres” en la página 146.
- **Panel Visualización** Puede explorar visualmente los clústeres y ver cómo interactúan en el panel. Para obtener más información, consulte el tema “Gráficos de clúster” en la página 161.
- **Panel Datos** Puede explorar y revisar los textos contenidos en los documentos y los registros que corresponden a las secciones en el cuadro de diálogo Definiciones de clústeres. Para obtener más información, consulte el tema “Definiciones de clúster” en la página 147.

Creando clústeres

Cuando acceda por primera vez a la vista Clústeres, no habrá ningún clúster. Puede crear los clústeres a través de los menús (**Herramientas > Crear clústeres**) o pulsando **Crear...** en la barra de herramientas. Esta acción abre el cuadro de diálogo Construir clústeres en el que puede definir los valores y límites para la construcción de los clústeres.

Importante: Cuando los resultados de la extracción ya no coinciden con los recursos, este panel se torna amarillo como el panel Resultados de extracción. Puede repetir la extracción para obtener los resultados de extracción más recientes y el color amarillo desaparecerá. Sin embargo, cada vez que se lleve a cabo una extracción el panel Clústeres se borra, y deberá reconstruir los clústeres. Los clústeres tampoco quedan guardados de una sesión a otra.

Las siguientes áreas y campos están disponibles en el cuadro de diálogo Construir clústeres:

Entradas

Tabla de entradas Los clústeres se construyen a partir de descriptores de ciertos tipos. En la tabla, puede seleccionar los tipos para incluirlos en el proceso de construcción. Los tipos que capturan la mayoría de registros o documentos son previamente seleccionados de forma predeterminada.

Conceptos a clúster: Seleccione el método de selección de conceptos que desea utilizar para la agrupación en clústeres. Al reducir el número de conceptos, puede acelerar el proceso de agrupación en clústeres. Puede agrupar en clústeres mediante la utilización de un número o porcentaje de conceptos principales, o todos los conceptos:

- **Número basado en doc. recuento** Cuando seleccione **Número máximo de conceptos**, especifique el número de conceptos que se deben tener en cuenta para la agrupación en clúster. Los conceptos se eligen basados en el mayor valor de recuento de documentos. El recuento de documentos es el número de documentos o registros en que aparece el concepto. El valor máximo es 150.000.
- **Porcentaje basado en doc. recuento** Cuando seleccione **Porcentaje máximo de conceptos**, especifique el porcentaje de conceptos que se deben tener en cuenta para la agrupación en clúster. Los conceptos se eligen basados en el porcentaje de conceptos con el mayor valor de recuento de documentos.

Límites de la salida

Número máximo de clústeres a crear Este valor es el número máximo de clústeres que se crearán y mostrarán en el panel Clústeres. Durante el proceso de agrupación en clústeres, los clústeres saturados se presentan antes que los no saturados, y por lo tanto, muchos de los clústeres resultantes serán saturados. Para ver más clústeres no saturados, puede establecer este valor en uno mayor que el número de clústeres saturados.

Número máximo de conceptos de un clúster Este valor es el número máximo de conceptos que puede contener un clúster.

Número mínimo de conceptos de un clúster Éste es el número mínimo de conceptos que deben enlazarse para crear un clúster.

Número máximo de enlaces internos Este valor es el número máximo de enlaces internos que puede contener un clúster. Los enlaces internos son enlaces entre pares de conceptos en un clúster.

Número máximo de enlaces externos Este valor es el número máximo de enlaces a conceptos fuera del clúster. Los enlaces externos son enlaces entre pares de conceptos en clústeres separados.

Valor de enlace mínimo Éste es el valor más pequeño de enlace aceptado para que un par de conceptos se puedan agrupar en clústeres. El valor de enlace se calcula mediante una fórmula de similitud. Para obtener más información, consulte el tema [“Calcular valores de similitud de enlaces”](#) en la [página 145](#).

Impedir el emparejamiento de conceptos específicos. Seleccione esta casilla de verificación para detener el proceso de agrupación o emparejamiento de dos conceptos en la salida. Para crear o gestionar pares de conceptos, pulse **Gestionar pares**. Para obtener más información, consulte el tema [“Administración de pares de excepciones de enlace”](#) en la [página 110](#).

Calcular valores de similitud de enlaces

Saber el número de documentos en que un par de conceptos aparecen juntos no le indica cuán similares son dichos conceptos. En estos casos, el valor de similitud puede resultar útil. El valor de similitud de enlace se mide mediante el conteo de co-apariciones de documento comparado al conteo de documentos individuales para cada concepto en la relación. Cuando se calcula la similitud, la unidad de medida es el número de documentos (conteo de documentos) en el que se encuentra un concepto o par de conceptos. Un concepto o par de conceptos se "encuentra" en un documento si este se aparece *al menos* una vez en el documento. Puede elegir que el grosor de la línea en el gráfico Concepto represente el valor de similitud de enlace en las gráficas.

El algoritmo revela las relaciones que son más fuertes, lo cual significa que la tendencia de los conceptos a aparecer juntos en los datos de texto es más alta que la tendencia a aparecer de forma independiente.

Internamente, el algoritmo produce un coeficiente de similitud que varía entre 0 y 1, donde el valor 1 significa que los dos conceptos siempre aparecen juntos y nunca por separado. El resultado del coeficiente de similitud es entonces multiplicado por 100 y redondeado al número entero más cercano. El coeficiente de similitud se calcula mediante la fórmula que se muestra en la siguiente figura.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Figura 31. Fórmula de coeficiente de similitud

Donde:

- C_I es el número de documentos o registros en el que aparece el concepto I.
- C_J es el número de documentos o registros en el que aparece el concepto J.
- C_{IJ} es el número de documentos o registros en el que co-aparecen el par de conceptos I y J en el conjunto de documentos.

Por ejemplo, suponga que tiene 5,000 documentos. Deje que I y J sean conceptos extraídos y deje que IJ sea una coocurrencia de par concepto de I y J. La tabla siguiente propone dos escenarios para demostrar cómo se calcula el valor del coeficiente y del enlace.

<i>Tabla 33. Ejemplo de frecuencias de concepto</i>		
Par/concepto	Escenario A	Escenario B
Concepto: I	Aparece en 20 documentos	Aparece en 30 documentos
Concepto: J	Aparece en 20 documentos	Aparece en 60 documentos
Par de concepto: IJ	Co-aparecen en 20 documentos	Co-aparecen en 20 documentos
Coeficiente de similitud	1	0.22222
Valor de enlace de similitud	100	22

En el caso A, los conceptos I y J así como el par IJ aparecen en 20 documentos, lo cual produce un coeficiente de similitud de 1; esto significa que los conceptos siempre aparecen juntos. El valor de similitud de enlace de este par sería 100.

En el caso B, el concepto I aparece en 30 documentos y el concepto J aparece en 60 documentos, pero el par IJ aparece solo en 20 documentos. Como resultado, el coeficiente de similitud es 0.22222. El valor de similitud de enlace para este par se redondea a 22.

Exploración de los clústeres

Después de construir los clústeres, puede ver el conjunto de resultados en el panel Clústeres. Para cada clúster, está disponible la siguiente información en la tabla:

- **Clúster.** Este es el nombre del clúster. Los clústeres se denominan de acuerdo al concepto con el mayor número de enlaces internos.
- **Conceptos.** Este es el número de conceptos en el clúster. Para obtener más información, consulte el tema [“Definiciones de clúster”](#) en la página 147.
- **Interno.** Este es el número de enlaces internos en el clúster. Los enlaces internos son enlaces entre pares de conceptos en un clúster.
- **Externo.** Este es el número de enlaces externos en el clúster. Los enlaces externos son enlaces entre pares de conceptos cuando un concepto está en un clúster y el otro concepto está en otro.
- **Sat.** Si hay un símbolo presente, este indica que el clúster puede haber sido más grande pero se excedieron uno o más límites, y, por lo tanto, el proceso de agrupación en clústeres finalizó para dicho clúster y se considera que está *saturado*. Al final del proceso de agrupación en clústeres, los clústeres saturados se presentan antes que los no saturados y, por lo tanto, muchos de los clústeres resultantes

estarán saturados. Para ver más clústeres no saturados, puede modificar el valor **Número máximo de clústeres a crear** en un valor mayor que el número de clústeres saturados, o disminuir el **Valor de enlace mínimo**. Para obtener más información, consulte el tema [“Creando clústeres”](#) en la [página 144](#).

- **Umbral.** Para todos los pares de conceptos que co-aparecen en el clúster, este es el valor de enlace de similitud menor de todos en el clúster. Para obtener más información, consulte el tema [“Calcular valores de similitud de enlaces”](#) en la [página 145](#). Un clúster con un valor de umbral alto significa que los conceptos en ese clúster tienen una similitud general mayor y están más estrechamente relacionados que aquellos cuyo valor de umbral es menor.

Para obtener más información acerca de un clúster determinado, puede seleccionarlo y el panel de visualización a la derecha le mostrará dos gráficas para ayudarlo a explorar el o los clústeres. Para obtener más información, consulte el tema [“Gráficos de clúster”](#) en la [página 161](#). También puede cortar y pegar los contenidos de la tabla en otra aplicación.

Cuando el resultado de la extracción ya no coincida con los recursos, este panel se vuelve amarillo así como el panel Resultados de extracción. Puede repetir la extracción para obtener los resultados de extracción más recientes y el color amarillo desaparecerá. Sin embargo, cada vez que se realiza una extracción, el panel Clústeres se borra y deberá reconstruir los clústeres. Los clústeres tampoco se guardan de una sesión a otra.

Definiciones de clúster

Puede ver todos los conceptos que hay dentro de un clúster seleccionándolo en el panel Clústeres y abriendo el recuadro de diálogo Definiciones de clúster (**Ver > Definiciones de clúster**).

Todos los conceptos en el clúster seleccionado aparecen en el cuadro de diálogo Definiciones de clústeres. Si selecciona uno o más conceptos en el cuadro de diálogo Definiciones de clúster y pulsa **Visualizar**, el panel de datos mostrará todos los registros o documentos en los que *todos los conceptos seleccionados aparecen juntos*. Sin embargo, el panel Datos no muestra ningún registro de texto o documentos cuando selecciona un clúster en el panel Clústeres. Para obtener información general acerca del panel Datos, consulte [en](#).

Seleccionar conceptos en este cuadro de diálogo también modifica el gráfico de concepto web. Para obtener más información, consulte el tema [“Gráficos de clúster”](#) en la [página 161](#). De forma similar, cuando selecciona uno o más conceptos en el cuadro de diálogo Definiciones de clústeres, el panel Visualizaciones mostrará todos los enlaces internos y externos de dichos conceptos.

Descripciones de columna

Se muestran iconos para que pueda identificar fácilmente cada descriptor.

Tabla 34. Iconos de descriptor y columnas	
Columnas	Descripción
Descriptors	El nombre del concepto.
 GLOBAL	Muestra el número de veces que el descriptor aparece en todo el conjunto de datos, también conocido como frecuencia global.
 Documentos	Muestra el número de documentos o registros en los que aparece el descriptor, también conocido como la frecuencia del documento.
type	Muestra el tipo o tipos a los que pertenece el descriptor. Si el descriptor es una regla de categoría, no se mostrará ningún nombre de tipo en esta columna.

Acciones de la barra de herramientas

Desde este cuadro de diálogo puede seleccionar uno o más conceptos para utilizar en una categoría. Hay varias maneras de hacer esto, pero es mejor seleccionar los conceptos que aparecen en un clúster y añadirlos como una regla de categoría. Para obtener más información, consulte el tema [“Reglas de](#)

coocurrencia” en la [página 113](#). Puede utilizar los botones de la barra de herramientas para añadir los conceptos a las categorías.

<i>Tabla 35. Botones de la barra de herramientas para añadir conceptos a las categorías</i>	
Iconos	Descripción
	Añadir los conceptos seleccionados a una categoría nueva o existente
	Añada los conceptos seleccionados en la forma de una & regla de categoría a una categoría nueva o existente. Para obtener más información, consulte el tema “Uso de reglas de categoría” en la página 119 .
	Añada cada uno de los conceptos seleccionados a sus nuevas categorías propias
	Actualiza lo que se muestra en el panel Datos y en el panel Visualización en función de los descriptores seleccionados

Nota: Puede también añadir conceptos a un tipo, como los sinónimos, o excluir elementos utilizando los menús contextuales.

Capítulo 11. Exploración del Análisis de enlaces de texto

En la vista Análisis de enlaces de texto (TLA), puede explorar los resultados de patrón de análisis de enlace de texto. Los análisis de enlaces de texto es una tecnología de coincidencia de patrón que le permite definir reglas de patrones y compararlos con conceptos reales extraídos y relaciones que se encuentra en el texto.

Por ejemplo, extraer ideas sobre una organización puede que no sea lo suficientemente interesante. Al utilizar el TLA, también puede aprender acerca de los vínculos entre esta organización y otras organizaciones o las personas dentro de una organización. También puede utilizar el TLA para extraer opiniones sobre productos o, para algunos idiomas, las relaciones entre los genes.

Una vez que haya extraído algunos resultados de patrón TLA, los puede revisar en los paneles de patrones de Tipo y Concepto de la vista Análisis de enlace de texto. Para obtener más información, consulte el tema [“Patrones de tipo y concepto”](#) en la [página 151](#). También puede explorarlos en los paneles de Datos o Visualización en esta vista. Posiblemente más importante, puede añadirlos a las categorías.

Si todavía no ha elegido hacerlo, puede pulsar en **Extraer** y elegir **Habilitar extracción de patrones de análisis de enlace de texto** en el cuadro de diálogo Extraer valores. Para obtener más información, consulte el tema [“Extracción de resultados del patrón TLA”](#) en la [página 150](#).

Debe haber algunas reglas de patrones TLA definidos en la plantilla de recursos o bibliotecas que está utilizando para extraer resultados de patrón TLA. Puede utilizar los patrones TLA en algunas plantillas de recursos que se suministra con IBM SPSS Modeler Text Analytics. Estos tipos de relaciones y patrones que puede extraer dependen enteramente de las reglas TLA definidas en sus recursos. Puede definir sus propias reglas de TLA. Los patrones se componen de las macros, las listas de palabras y lagunas de palabras para formar una consulta booleano, o una regla, que se compara con el texto de entrada. Para obtener más información, consulte el tema [Capítulo 18, “Sobre las reglas de enlaces de texto”](#), en la [página 211](#).

Siempre que una regla de patrón TLA coincide con el texto, este texto puede ser extraído como un patrón y reestructurado como datos de salida. Los resultados están entonces visibles en los paneles de vista Análisis de enlace de texto. Cada panel puede estar oculto o se podrá visualizar, seleccionando su nombre en el menú Ver:

- **Paneles de patrones de Tipo y Concepto.** Puede compilar y explorar los patrones en estos dos paneles. Para obtener más información, consulte el tema [“Patrones de tipo y concepto”](#) en la [página 151](#).
- **Panel de visualización.** Puede explorar visualmente cómo los conceptos y los tipos en los patrones interactúan en este panel. Para obtener más información, consulte el tema [“Gráficos de Análisis de enlace de texto”](#) en la [página 162](#).
- **Panel de datos.** Puede explorar y revisar el texto que contienen los documentos y registros que corresponden a las selecciones en otro panel. Para obtener más información, consulte el tema [“Panel Datos”](#) en la [página 153](#).

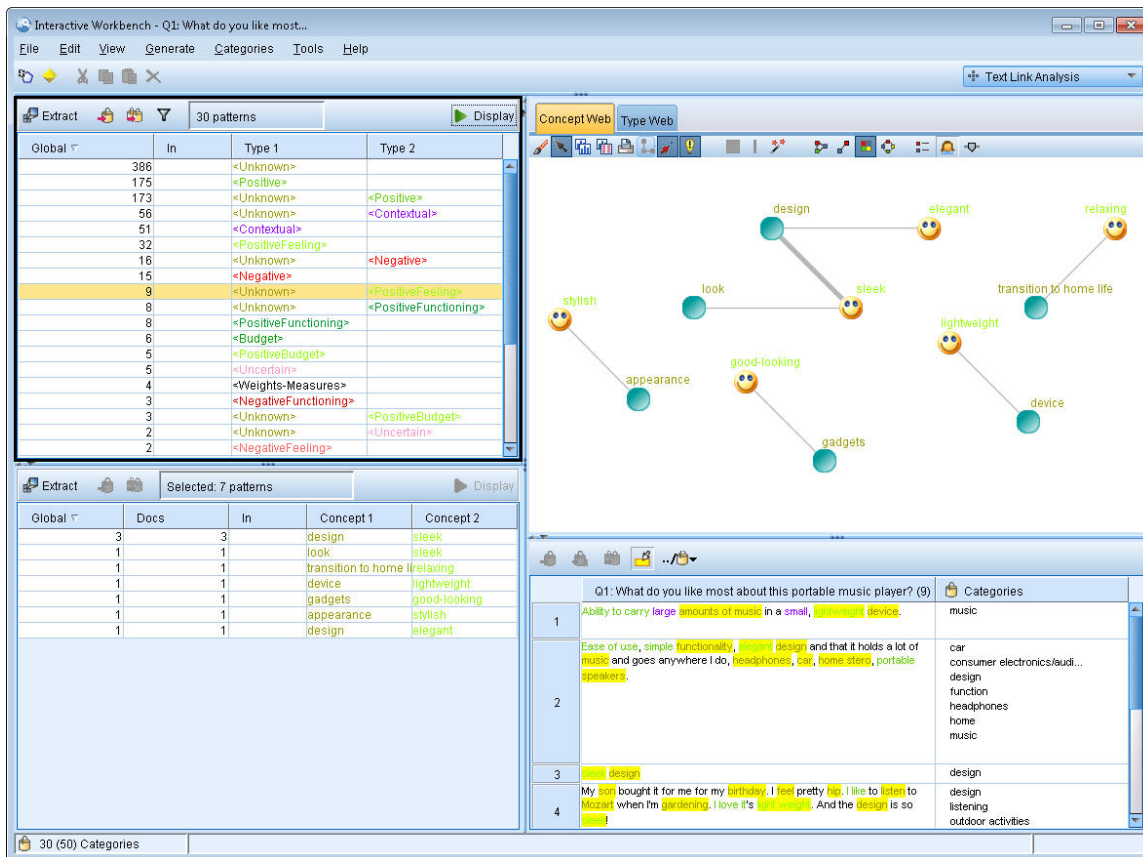


Figura 32. Vista Análisis para los enlaces de texto

Extracción de resultados del patrón TLA

El proceso de extracción genera un conjunto de conceptos y tipos, así como de patrones TLA (Análisis de enlace de texto), si está activado. Si ha extraído patrones TLA los puede ver en la vista Análisis de enlaces de texto. Cuando los resultados de extracción no están en sincronía con los recursos, los patrones de paneles de color amarillo indican que un re-extracción produciría resultados diferentes.

Debe elegir extraer estos patrones en el valor del nodo o en el cuadro de diálogo Extraer utilizando la opción **Habilitar la extracción de patrones análisis de enlace de texto**. Para obtener más información, consulte el tema “Extraer datos” en la página 82.

Nota: Hay relación entre el tamaño de conjuntos de datos y el tiempo que lleva completar el proceso de extracción. Consulte las instrucciones de instalación para conocer los estadísticos y recomendaciones sobre rendimiento. Siempre puede considerar la inserción de un nodo de ejemplo en sentido ascendente u optimizar la configuración de la máquina.

Para extraer datos

1. En los menús, seleccione **Herramientas > Extraer**. Como alternativa, pulse en el botón de la barra de herramientas **Extraer**.
2. Cambie cualquiera de las opciones que desee utilizar. Tenga en cuenta que la opción **Activar extracción de patrones de Análisis de enlace de texto** debe estar seleccionada en esta pestaña así como tener reglas TLA en su plantilla para poder extraer resultados de patrón TLA. Para obtener más información, consulte el tema “Extraer datos” en la página 82.
3. Pulse **Extraer** para empezar el proceso de extracción.

Una vez se inicia la extracción, se abre el cuadro de diálogo de progreso. Si desea anular la extracción, pulse en **Cancelar**. Cuando la extracción se haya completado, el cuadro de diálogo se cierra y los resultados aparecerán en el panel. Para obtener más información, consulte el tema “Patrones de tipo y concepto” en la página 151.

Patrones de tipo y concepto

Los patrones se componen de dos partes, una combinación de conceptos y tipos. Los patrones son muy útiles cuando se intenta descubrir opiniones sobre un asunto en particular o sobre relaciones entre conceptos. Por ejemplo, extraer el nombre del producto de la competencia podría no ser lo suficientemente interesante. En este caso, puede revisar los patrones extraídos para ver si puede encontrar ejemplos donde un documento o registro contenga un texto que expresen que el producto es bueno, malo o caro.

Los patrones puede contar de hasta seis tipos o seis conceptos. Por este motivo, las filas en ambos paneles de patrones contienen hasta seis espacios, o posiciones. Cada ranura corresponde a la posición específica de un elemento en la regla de patrón TLA tal como se define en los recursos lingüísticos. En el área de trabajo interactiva, si una ranura no contiene valores, no se muestra en la tabla. Por ejemplo, si los resultados de patrones más largos no contienen más de cuatro ranuras, los dos últimos no se muestran. Para obtener más información, consulte el tema [Capítulo 18, “Sobre las reglas de enlaces de texto”](#), en la página 211.

Cuando se extraen los resultados de patrón, se agrupan primero en el nivel tipo y después se dividen en patrones de concepto. Por esta razón, hay dos paneles de resultados diferentes: **Patrones Tipo** (superior izquierdo) y **Patrones Concepto** (inferior izquierdo). Para ver todos los patrones de concepto devueltos, seleccione todos los patrones de tipo. El panel de patrones de concepto en la parte inferior después mostrará todos los patrones de concepto hasta el valor de rango máximo (como se define el cuadro de diálogo Filtro).

Patrones tipo Este panel presenta resultados de patrón consta de uno o más tipos relacionados que coincide con una regla de patrón TLA. Los patrones tipo se muestran como <Organization> + <Location> + <Positive>, que podrían proporcionar comentarios positivos sobre una organización en una ubicación específica. La sintaxis es la siguiente:

```
<Type1> + <Type2> + <Type3> + <Type4> + <Type5> + <Type6>
```

Patrones de concepto Este panel presenta los resultados de patrón en el nivel de concepto para todos los patrones de tipo actualmente seleccionados en el panel de patrones de tipo por encima de él. Los patrones de concepto siguen una estructura como hotel + paris + wonderful. La sintaxis es la siguiente:

```
concept1 + concept2 + concept3 + concept4 + concept5 + concept6
```

Cuando los resultados de patrones utilizan menos de seis ranuras máximas, se muestra sólo el número necesario de ranuras (o columnas). Las ranuras vacías que se encuentran entre dos ranuras llenas se descartan de tal modo que el patrón <Type1>+<>+<Type2>+<>+<>+<> puede ser representado por <Type1>+<Type3>. Para un patrón de concepto, esto sería concept1+ . +concept2 (donde . representa un valor nulo).

Al igual que con los resultados de extracción en la vista Categorías y Conceptos, puede revisar los resultados aquí. Si ve perfeccionamientos que le gustaría hacer a los tipos y conceptos que componen estos patrones, los realiza en el panel de Extracción de resultados en la vista Categorías y Conceptos o directamente en el Editor de recursos y volver a extraer sus patrones. Cuando un concepto, tipo, o patrón se utiliza en una definición de categoría como es o como parte de una regla, aparece una categoría o un icono de regla en la columna **In** en la tabla de Resultados de patrón o extracción.

Nota: Si hay más resultados que puedan caber en el panel visible, puede utilizar los controles de la parte inferior del panel para desplazarse hacia delante y hacia atrás por los resultados, o escriba un número de página para ir a la misma.

Filtración de resultados TLA

Cuando está trabajando con conjuntos de datos muy grandes, el proceso de extracción puede producir millones de resultados. Para varios usuarios, puede hacer que sea más difícil revisar los resultados de manera efectiva. Puede, sin embargo, filtrar estos resultados con el fin de acercar aquellos que son más

interesantes. Puede cambiar los valores en el cuadro de diálogo Filtrar para limitar lo que se muestra patrones. Todos estos valores se utilizan en conjunto.

En la vista TLA, el cuadro de diálogo Filtrar contiene las áreas y cambios siguientes.

Filtrar por frecuencia Puede filtrar para visualizar sólo esos resultados con un cierto valor global o valor de frecuencia de documento .

- **Frecuencia global** es el número total de veces que un patrón aparece en el conjunto completo de documentos o registros y se muestra en la columna **Global**.
- **Frecuencia de documentos** es el número total de documento o registros en los que un patrón aparece y se puede ver en la columna **Documentos**.

Por ejemplo, si un patrón apareció 300 veces en 500 registros, podríamos decir que este patrón tiene una frecuencia global de 300 y una frecuencia de documento de 500.

Y por coincidencia de texto También puede filtrar para visualizar sólo los resultados que coincidan con la regla que defina aquí. Escriba el conjunto de caracteres que deben coincidir en el campo **Coincidencia de texto** y seleccione si se debe buscar este texto dentro de los nombres tipo o concepto, identificando el número de ranura o todas ellas. A continuación, seleccione la condición en la cual aplicar la coincidencia (no es necesario utilizar corchetes de ángulo para denotar el principio o al final de un nombre de tipo). Seleccione **Y** o **O** de la lista desplegable para que la regla coincida con ambas sentencias o sólo con una de ellas, y defina el segundo texto coincidente de la sentencia de la misma manera que el primero.

Tabla 36. Coincidir condiciones de texto	
Condición	Descripción
Contiene	El texto se coincide si la cadena ocurre en cualquier lugar. (opción predeterminada)
Empieza con	El texto coincide sólo si el concepto o tipo comienza con el texto especificado.
Termina con	El texto coincide sólo si el concepto o tipo termina con el texto especificado.
Coincidencia exacta	La serie entera debe coincidir con el nombre de tipo o concepto.

Resultados que se muestran en el panel de patrones

Suponga que está utilizando una versión en inglés del software; aquí se pueden ver algunos ejemplos de cómo los resultados pueden visualizarse en la barra de herramientas del panel Patrones basándose en los filtros.



Figura 33. Resultados de filtro ejemplo 1

En este ejemplo, la barra de herramientas muestra que el número de patrones devuelto era limitado debido al rango máximo especificado en el filtro. Si aparece un icono púrpura, significa que se ha llegado al número máximo de patrones. Pase el ratón por encima del icono para obtener más información. Vea la explicación anterior del filtro **Y por rango**.



Figura 34. Ejemplo 2 de resultados de filtro

En este ejemplo, la barra de herramientas muestra que resultados fueron limitados utilizando un filtro de texto de coincidencia (vea el icono de la lupa). Puede pasar por encima del icono para ver cuál es el texto de coincidencia.

Para filtrar los resultados

1. En los menús elija **Herramientas > Filtrar**. Se abre el cuadro de diálogo Filtrar.

2. Seleccione y perfeccione los filtros que desea utilizar.
3. Pulse **Aceptar** para aplicar los filtros ver los resultados nuevos.

Panel Datos

Al extraer y explorar patrones de análisis de enlace de texto, puede que desee revisar algunos de los datos con los que está trabajando. Por ejemplo, puede que desee ver los registros reales en los que un grupo de patrones fueron descubiertos. Puede revisar registros o documentos en el panel Datos, que se encuentra en la parte inferior derecha. Si no está visible de forma predeterminada, elija **Ver > Paneles > Datos** en los menús.

El panel Datos presenta una fila por documento o registro correspondiente a una selección en la vista, hasta un límite de visualización. De forma predeterminada, el número de documentos o de registros que se muestran en el panel Datos se limita a fin de hacerlo más rápido para que pueda ver sus datos. Sin embargo, puede ajustar esto en el cuadro de diálogo Opciones. Si desea más información, consulte [“Opciones: separador Sesión” en la página 75](#).

Nota: Si hay más resultados que puedan caber en el panel visible, puede utilizar los controles de la parte inferior del panel para desplazarse hacia delante y hacia atrás por los resultados, o escriba un número de página para ir a la misma.

Visualización y renovación del panel Datos

El panel Datos no renueva su visualización automáticamente, porque con grande conjuntos, la renovación automática de datos podría tardar algún tiempo en completarse. Por lo tanto, cada vez que selecciona patrones tipo o concepto en esta vista, puede pulsar **Visualización** para renovar los contenidos del panel Datos.

Documentos o registros de texto

Si los datos de texto se encuentran en forma de registros y el texto es relativamente corto en longitud, el campo de texto en el panel Datos muestra los datos de texto en su totalidad. Sin embargo, al trabajar con registros y grandes conjuntos de datos, la columna de campo de texto muestra un breve fragmento del texto y abre un panel Vista previa de texto a la derecha para mostrar más o todo el texto del registro que ha seleccionado en la tabla. Si los datos de texto están en forma de documentos individuales, el panel Datos muestra el nombre de archivo del documento. Cuando selecciona un documento, se abre el panel de Vista previa con el texto del documento seleccionado.

Colores y resaltado

Siempre que se visualizan los datos, los conceptos y descriptores que se encuentran en esos documentos o registros se resaltan en color para ayudarle a identificarlos fácilmente en el texto. La codificación por colores corresponde a los tipos a los que pertenecen los conceptos. También puede pasar el ratón por encima de los elementos con codificación por color para mostrar el concepto bajo el que se extrajo y el tipo al que se asignó. El texto que no se ha extraído aparece en negro. Generalmente, estas palabras no extraídas suelen ser conectores (y o *con*), pronombres (*mi* o *ellos*), y verbos (*es*, *tiene* o *tomar*).

Columnas del panel Datos

Mientras que la columna de campo de texto siempre está visible, también puede mostrar otras columnas. Para visualizar otras columnas, seleccione **Ver > Panel Datos** en los menús y, a continuación, seleccione la columna que desea visualizar en el panel Datos. Las siguientes columnas pueden estar disponibles para su visualización:

- **"Nombre de campo de texto" (#)/Documentos.** Añade una columna para los datos de texto desde los que se extrajeron los conceptos y el tipo. Si los datos están en documentos, la columna se denomina Documentos y sólo el nombre de archivo del documento o la vía de acceso completa es visible. Para ver el texto de dichos documentos debe buscar en el panel Vista previa de texto. El número de filas en el panel Datos se muestra entre paréntesis después de este nombre de columna. Puede haber ocasiones en que no todos los documentos o registros se muestran debido a un límite en el diálogo Opciones utilizado para aumentar la velocidad de carga. Si se alcanza el máximo, el número será seguido de **-Max**. Si desea más información, consulte [“Opciones: separador Sesión” en la página 75](#).

- **Categorías.** Enumera cada una de las categorías a las que pertenece un registro. Cuando se muestra esta columna, la actualización del panel Datos puede tardar un poco más hasta mostrar la información más actualizada.
- **Forzar dentro.** Lista las categorías en las cuales ha forzado. Puede forzarse la inclusión de documentos en una categoría mediante la selección de menús **Editar > Forzar dentro**. Si desea más información, consulte [“Forzar los documentos de las categorías”](#) en la página 140.
- **Forzar fuera.** Lista las categorías de las cuales ha eliminado un documento. Puede forzarse la exclusión de documentos de una categoría mediante la selección de menús **Editar > Forzar fuera**. Por ejemplo, esto podría utilizarse cuando el sarcasmo de un encuestado hace que una respuesta caiga en la categoría incorrecta. Si desea más información, consulte [“Forzar los documentos de las categorías”](#) en la página 140.
- **Recuentos de categorías.** Lista el número de las categorías a las que pertenece un registro.
- **Rangos de relevancia.** Proporciona un orden para cada registro de una categoría. Este orden muestra el grado de adecuación del registro en la categoría si se compara con otros registros de la misma categoría. Para ver el rango, seleccione una categoría del panel Categorías (panel superior izquierdo). Si desea más información, consulte [“Relevancia de categoría”](#) en la página 105.
- **Marcas de respuesta.** Añada una columna que muestra las marcas que puedan estar utilizándose. Pulse dentro de la columna para cambiar el tipo de marca que asigna a los documentos. Podría marcar los documentos con la marca "completo" o la marca "importante", o bien eliminar las marcas. Esto es útil para revisar la precisión de un modelo de categoría. Si desea más información, consulte [“Marcar respuestas”](#) en la página 105.

Marcar respuestas

Para ayudarle en su trabajo, puede marcar los documentos con marcas (distintivos) en el panel Datos. esta característica solo está disponible si el documento de origen contiene un ID exclusivo. Si el documento de origen no contiene un ID exclusivo, puede añadir un nodo Derivar entre el documento de origen y el nodo Text Mining.

Existen muchas razones por las cuales podría desear marcar un documento, entre ellas:

- Para marcar documentos que ha revisado manualmente, de forma que sepa donde reanudar más adelante.
- Para marcar un documento que no sabe muy bien cómo manejar

Una vez ha señalado un documento con una marca, puede continuar trabajando con el mismo. Las marcas solo tienen fines de registro. Puede elegir entre las siguientes marcas:

Tabla 37. Descripciones de las marcas

Marca	Descripción
	La marca de Completo señala los documentos que considera acabados.
	La marca de Importante señala los documentos que considera importantes.

Para marcar un documento con una marca:

1. Desde el panel Datos, pulse con el botón derecho del ratón sobre el documento que desea marcar.
2. Desde el menú de contexto, seleccione **Vista > Panel Datos > Marcas de respuesta** y seleccione a continuación el tipo de marca que desea utilizar (marca de Importante o Completo). Se asignará la marca seleccionada. Si la columna Marca del panel Datos no estaba visible, ahora aparecerá.

Para borrar las marcas:

1. Desde el panel Datos, pulse con el botón derecho del ratón sobre los documentos de los que desea eliminar la marca.
2. Desde el menú de contexto, seleccione **Marcar respuestas con** > **Borrar marcas**. Las marcas seleccionadas se eliminan.

Reglas de reasignación de tipo

Las reglas de reasignación de tipo (TRR) transforman una secuencia de tipos, macros y elementos en un nuevo concepto con un tipo específico. Específicamente, se utilizan en plantillas Opiniones para capturar opiniones con un cambio de polaridad. Por ejemplo, en la frase "not that bad", la palabra "bad" es una opinión *negativa*. Pero en este contexto, el significado real es "not bad", que es *positivo*.

Hasta la versión 18.2, este cambio de polaridad se gestionaba mediante reglas de análisis de enlace de texto (TLA) específicas:

Figura 35. Reglas TLA

A causa de los diferentes tipos de opinión (Positive, PositiveAttitude, PositiveBudget, PositiveCompetence, PositiveFeeling, PositiveFunctioning, PositiveRecommendation, Negative, NegativeAttitude, NegativeBudget, NegativeCompetence, NegativeFeeling, NegativeFunctioning, NegativeRecommendation y Contextual), se han tenido que escribir reglas TLA específicas:

- Para cada tipo. Por ejemplo:

```
"not + xxx + <NegativeBudget>" => "<PositiveBudget>"
```

O

```
"not + xxx + <PositiveAttitude>" => "<NegativeAttitude>"
```

- Para muchos contextos sintácticos. Por ejemplo:

```
* topic + negation + opinion ("hotel wasn't good")
* negation + opinion + topic ("it was not a good hotel")
* negation + opinion ("not very good")
* topic + opinion + negation + opinion ("hotel was well-located but not that good")
* 2 topics + negation + opinion ("room and swimming pool weren't always clean")
* ...
```

A partir de la versión 18.2, el nuevo enfoque es "captar" dichas secuencias (cualquier negación + cualquier palabra vacía + una opinión específica), seleccionar las palabras que aparecen en el nuevo concepto (una negación estandarizada - por ejemplo, "no" - y la opinión) y definir un tipo para este nuevo concepto (también denominado, "pseudo-término"). Este nuevo concepto puede utilizarse en las reglas TLA.

Como consecuencia, la siguiente regla coincidirá con cualquier secuencia que contenga un tema seguido de una opinión, ya sea que la opinión sea un término (comfortable) o un pseudo-término (not economical), independientemente del subtipo de opinión específico (Actitud, Presupuesto, etc).

```
#@# Bed was extremely comfortable
[pattern(190)]
name=topic + opinion_190
value=$mTopic ($mEmpty|$mToo){0,3} ($mOpinionPos|$mOpinionNeg|$Contextual)
output(1)=$1\t#1\t$3\t#3
```

Junto con el cambio de polaridad para las opiniones, también puede utilizar las TRR como ayuda en el ajuste del diccionario. Por ejemplo, supongamos que tiene un tipo llamado Anatomy con partes del cuerpo como heart, chest, breast y adrenal gland, y otro tipo llamado MedicalProcedures con procedimientos como biopsy, needle biopsy, MRI y CT scan. Sería casi imposible listar todos los procedimientos médicos correctamente asociados con un órgano. De manera que podría crear dos TRR para identificar posibles procedimiento médicos, tal como se observa en las figuras siguientes. Así, cuando se realiza la extracción, puede añadir un filtro en el tipo PotentialMedicalProcedures, revisar los términos candidatos y añadirlos después al tipo MedicalProcedures.

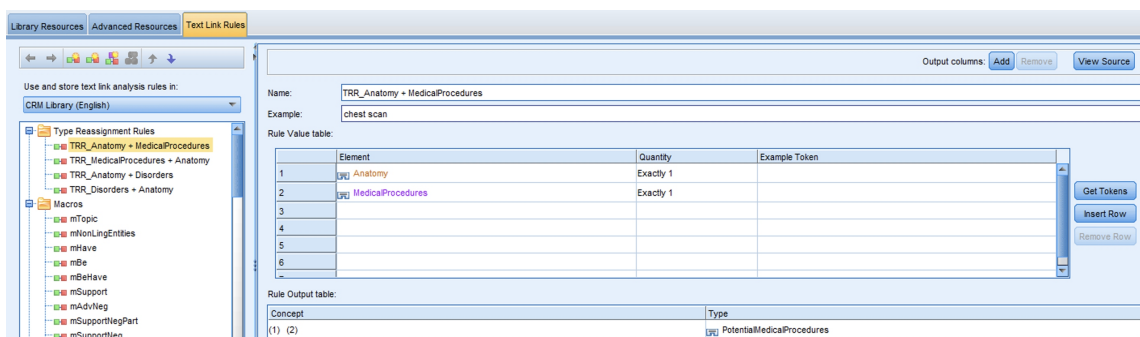


Figura 36. TRR para anatomía + procedimientos médicos

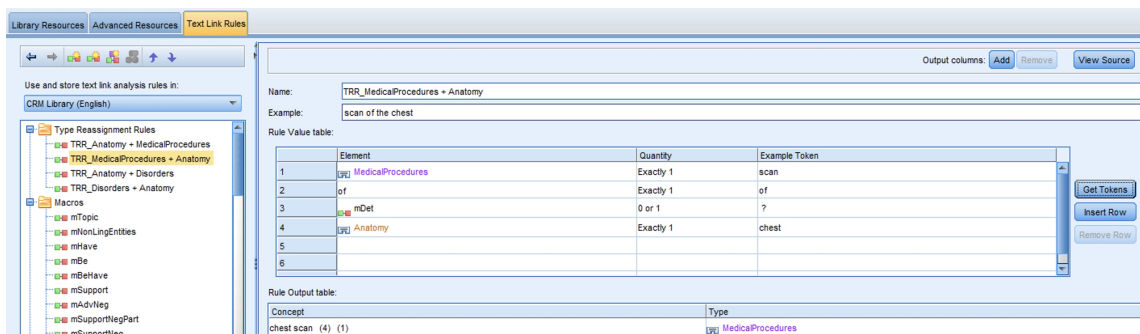


Figura 37. TRR para procedimientos médicos + anatomía

Sintaxis

```
#@# not that expensive
[typeReassignmentRule]
name=TRR_"not" NegativeBudget
value=$mAllNeg ($mAdverb|$mBe|$mHave|$mSupport|$mDet|that|more|$mQuant){0,3} $NegativeBudget
output=not $3\tPositiveBudget
```

- "name" debe ser exclusivo (TRR_"not" NegativeBudget). No se puede usar en una macro o en una regla TLA. Solo se puede utilizar el tipo definido en la salida.
- "value" es una secuencia de elementos a buscar. Los elementos pueden ser tipos (\$NegativeBudget), macros (\$mAllNeg) o señales (more). Algunos elementos pueden ser necesarios, opcionales o tener una cantidad específica.
- "output" es un único par de concepto + tipo (not \$3\tPositiveBudget). Tenga en cuenta que en la salida puede utilizar un tipo disponible (que está definido en la plantilla) o puede crear un tipo nuevo.

El tipo de salida también puede hacer referencia a un elemento coincidente (por ejemplo, #2). Esta característica es especialmente útil cuando no hay ningún cambio en el tipo entre el valor y la salida. Por ejemplo:

```
#@# could not have been any more pleased
[typeReassignmentRule]
name=TRR_ "couldn't be more" opinion
value=$mNotNeg ($mOpinionPos|$mOpinionNeg|$Contextual)
output=$2\t#2
```

Como en las reglas TLA, debe definirse una TRR más específica antes de una más genérica. Para asegurarse de que todas las TRR estén definidas en el orden correcto, puede utilizar la característica Obtener señales para probar cada TRR de la secuencia. Si una TRR no coincide, pero sí coincide con otra definición, puede subirla (o bajarla).

Casos especiales

En algunos casos, es necesario poder acceder a los elementos individuales de la secuencia y no a una TRR. Esto es especialmente aplicable a la *coordinación* sobre la *negación*. En la frase "not that fashionable or eyecatching", la coordinación "or" no permite descubrir que, en este contexto, "eyecatching" en realidad significa "not eyecatching".

En este caso, recomendamos utilizar una regla específica como:

```
#@# not that fashionable or eyecatching
[pattern(263)]
name="not" + 2 Positive_263
value=($mAdvNeg|$mSupportNeg|$mMiscNeg) @{{0,1}} $PositiveFeeling or $PositiveFeeling
output(1)=not $3\tNegativeFeeling
output(2)=not $5\tNegativeFeeling
```

Aunque la primera parte de la regla (($\$mAdvNeg|\$mSupportNeg|\$mMiscNeg$) @{{0,1}} $\$PositiveFeeling$) puede coincidir con TRR, la regla TLA tendrá prioridad.

Si escribe una regla más genérica, como en el ejemplo siguiente, las mismas restricciones que existían en la versión 18.1.1, y anteriores, siguen aplicándose. El nuevo concepto creado (pseudo-concepto) puede tener un tipo incorrecto (<Negative> en lugar de <NegativeFeeling>), y puede acabar teniendo un concepto TLA con dos tipos distintos. Una alternativa es crear el término correspondiente (no xxx) con el tipo correcto.

```
#@# not that fashionable or eyecatching
[pattern(263)]
name="not" + 2 Positive_263
value=($mAdvNeg|$mSupportNeg|$mMiscNeg) @{{0,1}} $mPos or $mPos
output(1)=not $3\tNegative
output(2)=not $5\tNegative
```

Ventajas

- La principal ventaja del uso de TRR es tener menos reglas TLA.
- Una ventaja menos obvia es que las TRR garantizan que un pseudo-término generará en la mayoría de ocasiones el tipo correcto (pero no olvide la restricción mencionada anteriormente). En el pasado, algunas "not + positiveXXX" se escribían como Negative en lugar de NegativeXXX, porque faltaban algunas reglas TLA específicas.
- Si un usuario añade un tipo de opinión específica (por ejemplo, NegativeNoise), no hay ninguna necesidad de replicar reglas TLA específicas para invertir la polaridad. El usuario solo tiene que crear la TRR correspondiente.

Capítulo 12. Visualización de gráficos

La vista Categorías y Conceptos, la vista Clústeres, y la vista Análisis de enlace de texto tienen un panel de visualización en la esquina superior derecha de la ventana. Puede utilizar este panel para explorar visualmente los datos. Los siguientes gráficos y diagramas están disponibles.

- **Vista Categorías y conceptos.** Esta vista tiene tres gráficos y diagramas: *Barra de categoría*, *Web de categoría*, y *Tabla web de categoría*. En esta vista, los gráficos solo se actualizan al pulsar **Mostrar**. Para obtener más información, consulte el tema [“Gráficos de categoría”](#) en la página 159.
- **Vista Clústeres.** Esta vista tiene dos gráficos web: *Gráfico web de concepto* y *Gráfico web de clúster*. Para obtener más información, consulte el tema [“Gráficos de clúster”](#) en la página 161.
- **Vista Análisis de enlace de texto.** Esta vista tiene dos gráficos web: *Gráfico web de concepto* y *Gráfico web de tipo*. Para obtener más información, consulte el tema [“Gráficos de Análisis de enlace de texto”](#) en la página 162.

Si desea más información sobre todas las barras de herramientas generales y paletas utilizadas para editar gráficos, consulte la sección sobre la Edición de gráficos en la ayuda en línea o en el archivo *ModelerSPOnodes.pdf*, que está disponible como parte de la descarga del producto.

Gráficos de categoría

Al crear las categorías, es importante tomarse tiempo para revisar las definiciones de categoría, los documentos o registros que contienen, y cómo se solapan las categorías. El panel de visualización ofrece varias perspectivas sobre sus categorías. El panel de visualización se encuentra en la esquina superior derecha de la vista Categorías y conceptos. Si todavía no está visible, puede acceder a este panel desde el menú Ver (**Ver > Paneles > Visualización**).

En esta vista, el panel de visualización ofrece tres perspectivas sobre las similitudes en la categorización de documento o registro. Todos estos gráficos de este panel pueden utilizarse para analizar los resultados de la categorización y ayudarle a ajustar las categorías o los informes. Al refinar categorías, puede utilizar este panel para revisar las definiciones de categoría para descubrir categorías que sean muy similares (por ejemplo, que compartan más del 75% de sus documentos o registros) o muy distintas. Si hay dos categorías demasiado similares entre sí, puede resultar útil combinar las dos categorías. Como alternativa, puede refinar las definiciones de categoría eliminando determinados descriptores de una categoría o de otra.

En función de lo que se haya seleccionado en el panel Resultados de extracción, panel Categorías, o en el recuadro de diálogo Definiciones de categoría, puede ver las interacciones correspondientes entre documentos/registros y categorías en cada una de las pestañas de este panel. Cada una presenta información similar, pero de una forma distinta o con un nivel diferente de detalle. Sin embargo, para renovar un gráfico para la selección actual, pulse **Visualizar** en la barra de herramientas del panel o recuadro de diálogo en el que haya realizado su selección.

El panel de visualización en la vista Categorías y Conceptos ofrece los diagramas y gráficos siguientes:

- **Gráfico de barras de categoría.** Una tabla y gráfico de barras presentan el solapamiento entre los documentos/registros correspondientes a su selección y las categorías asociadas. El gráfico de barras también presenta relaciones de los documentos/registros en categorías con el número total de documentos/registros. Para obtener más información, consulte el tema [“Gráfico de barras de categorías”](#) en la página 160.
- **Gráfico web de categoría.** Este gráfico presenta el solapamiento de documento/registro para las categorías a las que pertenecen los documentos/registros según la selección en los otros paneles. Para obtener más información, consulte el tema [“Gráfico de malla de categorías”](#) en la página 160.
- **Tabla web de categoría.** Esta tabla presenta la misma información que la pestaña Malla de categorías, pero en formato de tabla. La tabla contiene tres columnas que pueden ordenarse pulsando en las

cabeceras de columna. Para obtener más información, consulte el tema [“Tabla de malla de categorías” en la página 160.](#)

Para obtener más información, consulte el tema [Capítulo 9, “Categorización de datos de texto”, en la página 95.](#)

Gráfico de barras de categorías

Esta pestaña muestra una tabla y gráfico de barras que muestran el solapamiento entre los documentos/registros correspondientes a su selección y las categorías asociadas. El gráfico de barras también presenta relaciones de los documentos/registros en categorías con el número total de documentos o registros. No se puede editar el diseño de este gráfico. Sin embargo, puede ordenar las columnas pulsando en las cabeceras de las columnas.

El gráfico contiene las columnas siguientes:

- **Categoría.** Esta columna presenta el nombre de las categorías que ha seleccionado. De forma predeterminada, la categoría más común de su selección aparece en el primer lugar de la lista.
- **Barra.** Esta columna presenta, de forma visual, la relación de los documentos o registros en una determinada categoría con el número total de documentos o registros.
- **% selección.** Esta columna presenta un porcentaje basado en la relación del número total de documentos o registros para una categoría con el número total de documentos o registros representados en la selección.
- **Docs.** Esta columna presenta el número de documentos o registros en una selección para una determinada categoría.

Gráfico de malla de categorías

Esta pestaña muestra un gráfico de malla de categorías. La web presenta el solapamiento de documentos o registros para las categorías a las que pertenecen los documentos o registros en función de la selección en los otros paneles. Si existen etiquetas de categoría, estas aparecen en el gráfico. Puede elegir un diseño de gráfico (de red, circular, direccional o de cuadrícula) mediante los botones de la barra de herramientas de este panel.

En la malla, cada nodo representa una categoría. Con el ratón puede seleccionar y mover los nodos en el panel. El tamaño del nodo representa el tamaño relativo basado en el número de documentos o registros de dicha categoría en su selección. El grosor y el color de la línea entre dos categorías denotan el número de documentos o registros comunes que tienen. Si pasa el ratón por encima de un nodo en el modo Explorar, la información de la herramienta muestra el nombre (o etiqueta) de la categoría y el número global de documentos o registros en la categoría.

Nota: De forma predeterminada, la modalidad Explorar está habilitada para los gráficos en los que puede mover nodos. Sin embargo, puede cambiar al modo Editar para modificar el diseño de los gráficos, incluidos los colores, las fuentes, las leyendas, etc. Para obtener más información, consulte [“Uso de barras de herramientas y paletas de gráficos” en la página 163.](#)

Si copia los datos gráficos utilizando el botón **Copiar datos de visualización**, y los pega en una hoja de cálculo o en editor de texto, a los datos se les proporcionan cabeceras de columna como, por ejemplo, V1, V2 hasta V7. Estas columnas contienen la información siguiente:

- **V1, V2** Estos valores corresponden a las coordenadas de pantalla (X e Y, respectivamente).
- **V3, V5** Lista el concepto de categoría.
- **Tamaño, V6** Muestra el número de documentos en los que se encontraron los conceptos.
- **V7** No se utiliza actualmente.

Tabla de malla de categorías

Esta tabla presenta la misma información que la pestaña Malla de categorías, pero en formato de tabla. La tabla contiene tres columnas que pueden ordenarse pulsando en las cabeceras de columna:

- **Recuento.** Esta columna presenta el número de documentos o registros compartidos o comunes entre las dos categorías.
- **Categoría 1.** Esta columna presenta el nombre de la primera categoría seguida del número total de documentos o registros que contiene, que se muestra entre paréntesis.
- **Categoría 2.** Esta columna presenta el nombre de la segunda categoría seguida del número total de documentos o registros que contiene, que se muestra entre paréntesis.

Gráficos de clúster

Después de construir los clústeres, puede explorarlos en los gráficos web en el panel de visualización. El panel de visualización ofrece dos perspectivas de la agrupación en clúster: un gráfico de concepto web y uno de clúster web. Los gráficos en este panel pueden utilizarse para analizar los resultados de la agrupación en clúster y para ayudar a descubrir algunos conceptos y reglas que desee añadir a las categorías. El panel de visualización está ubicado en la esquina superior derecha de la vista Clústeres. Si no es ya visible, puede acceder a este panel desde el menú Ver (**Ver > Paneles > Visualización**). Mediante la selección de un clúster en el panel Clústeres, puede mostrar automáticamente los gráficos correspondientes en el panel de visualización.

Nota: De forma predeterminada, los gráficos están en modo interactivo/ selección en el cual los nodos se pueden mover. Sin embargo, puede editar el diseño de gráfico en modo Edición, incluyendo colores y fuentes, leyendas, etc. Para obtener más información, consulte el tema [“Uso de barras de herramientas y paletas de gráficos”](#) en la página 163.

La vista Clústeres tiene dos gráficos web.

- **Gráfico concepto web.** Este gráfico presenta todos los conceptos dentro del clúster seleccionado así como los conceptos vinculados fuera del clúster. Este gráfico puede ayudarle a ver en qué forma se vinculan los conceptos dentro de un clúster y cualquier enlace externo. Para obtener más información, consulte el tema [“Gráfico concepto web”](#) en la página 161.
- **Gráfico clúster web.** Este gráfico presenta los clústeres seleccionados con todos los enlaces externos entre los clústeres seleccionados en forma de líneas punteadas. Para obtener más información, consulte el tema [“Gráfico clúster web”](#) en la página 162.

Para obtener más información, consulte el tema [Capítulo 10, “Análisis de clústeres”](#), en la página 143.

Gráfico concepto web

Este separador muestra un gráfico web que muestra todos los conceptos dentro del clúster o clústeres seleccionados, así como los conceptos vinculados fuera del clúster. Este gráfico puede ayudarle a ver en qué forma se vinculan los conceptos dentro de un clúster y cualquier enlace externo. Cada concepto en un clúster es representado como un nodo, que se codifica con un color de acuerdo al color de tipo. Para obtener más información, consulte el tema [“Crear tipos”](#) en la página 189.

Los enlaces internos entre los conceptos en un clúster se dibujan y el grosor de la línea de cada enlace está directamente relacionado al recuento de documentos de cada co-ocurrencia de par de concepto o la similaridad del valor de enlace, dependiendo de la opción en la barra de herramientas de gráfico. Los enlaces externos entre los conceptos de un clúster y los conceptos fuera de este también se muestran.

Si los conceptos se seleccionan en el recuadro de diálogo Definiciones de clúster, el gráfico concepto web mostrará los conceptos y cualquier enlace interno y externo asociado a dichos conceptos. Cualquier enlace entre otros conceptos que no incluyan uno de los conceptos seleccionados no aparecerán en el gráfico.

Nota: De forma predeterminada, los gráficos están en modo interactivo/ selección en el cual los nodos se pueden mover. Sin embargo, puede editar el diseño de los gráficos en modo Editar, incluidos los colores y las fuentes, las leyendas, etc. Para obtener más información, consulte [“Uso de barras de herramientas y paletas de gráficos”](#) en la página 163.

Si copia los datos gráficos utilizando el botón **Copiar datos de visualización**, y los pega en una hoja de cálculo o en editor de texto, a los datos se les proporcionan cabeceras de columna como, por ejemplo, V1, V2 hasta V7. Estas columnas contienen la información siguiente:

- **V1, V2** Estos valores corresponden a las coordenadas de pantalla (X e Y, respectivamente).
- **V3, V6** Lista el tipo de concepto.
- **V4, V5** Muestra la etiqueta del concepto.
- **V7** No se utiliza actualmente.

Gráfico clúster web

Este separador muestra un gráfico web con los clúster seleccionados. Los enlaces externos entre los clústeres seleccionados, así como cualquier enlace entre otros clústeres, se muestran como líneas punteadas. En un gráfico clúster web, cada nodo representa un clúster completo y el grosor de las líneas trazadas entre ellos representa el número de enlaces externos entre dos clústeres.

Importante: Para mostrar un gráfico clúster web, debe tener con enlaces externos ya construidos. Los enlaces externos son enlaces entre pares de conceptos de clústeres separados (un concepto dentro de un clúster y un concepto fuera, en otro clúster).

Por ejemplo, digamos que tenemos dos clústeres. Cluster A tiene tres conceptos: A1, A2 y A3. Cluster B tiene dos conceptos: B1 y B2. Los siguientes conceptos están vinculados: A1 - A2, A1 - A3, A2 - B1 (Externo), A2 - B2 (Externo), A1 - B2 (Externo), y B1 - B2. Esto significa que en el gráfico clúster web, el grosor de la línea representa los tres enlaces externos.

Nota: De forma predeterminada, los gráficos están en modo interactivo/ selección en el cual los nodos se pueden mover. Sin embargo, puede editar el diseño de los gráficos en modo Editar, incluidos los colores y las fuentes, las leyendas, etc. Para obtener más información, consulte el tema [“Uso de barras de herramientas y paletas de gráficos”](#) en la página 163.

Gráficos de Análisis de enlace de texto

Después de extraer patrones del Análisis de enlace de texto (TLA), puede explorarlos visualmente en los gráficos web en el panel de visualización. El panel de visualización ofrece dos perspectivas sobre patrones TLA: un concepto (patrón) gráfico web y un tipo (patrón) gráfico web. Los gráficos web en este panel se pueden utilizar para representar patrones visualmente. El panel de visualización se encuentra en la esquina superior derecha del Análisis de enlace de texto. Si no es visible, puede acceder a este panel desde el menú Ver (**Ver > Paneles > Visualización**). Si no hay ninguna selección, el área de gráfico está vacía.

Nota: De forma predeterminada, los gráficos están en modo interactivo/ selección en el cual los nodos se pueden mover. Sin embargo, puede editar el diseño de los gráficos en modo Editar, incluidos los colores y las fuentes, las leyendas, etc. Para obtener más información, consulte el tema [“Uso de barras de herramientas y paletas de gráficos”](#) en la página 163.

La vista Análisis de enlace de texto tiene dos gráficos web.

- **Gráfico concepto web.** Este gráfico presenta todos los conceptos de los patrones seleccionados. La anchura de línea y los tamaños de nodo (si no se muestran los iconos de tipo) de un gráfico de concepto muestran el número de apariciones globales en la tabla seleccionada. Para obtener más información, consulte el tema [“Gráfico concepto web”](#) en la página 163.
- **Tipo de Gráfico web.** Este gráfico presenta todos los tipos de los patrones seleccionados. La anchura de línea y los tamaños de nodo (si no se muestran los iconos de tipo) del gráfico muestran el número de apariciones globales en la tabla seleccionada. Los nodos se representan ya sea mediante un tipo de color o mediante un icono. Para obtener más información, consulte el tema [“Tipo de gráfico web”](#) en la página 163.

Para obtener más información, consulte el tema [Capítulo 11, “Exploración del Análisis de enlaces de texto”](#), en la página 149.

Gráfico concepto web

Este gráfico web presenta todos los conceptos representados en la selección actual. Por ejemplo, si seleccionó un patrón tipo que tenía tres patrones de concepto coincidentes, este gráfico mostrará tres conjuntos de conceptos enlazados. La anchura de línea y tamaños de nodo en un gráfico concepto representan los recuentos de frecuencia global. El gráfico representa visualmente la misma información que lo que se ha seleccionado en los paneles de patrones. Los tipos de cada concepto se presentan ya sea por color o por un icono dependiendo de lo que seleccione en la barra de gráfico. Para obtener más información, consulte el tema [“Uso de barras de herramientas y paletas de gráficos”](#) en la página 163.

Tipo de gráfico web

Este gráfico web presenta cada tipo de patrón para la selección actual. Por ejemplo, si seleccionó dos patrones de concepto, este gráfico muestra un nodo por tipo en los patrones seleccionados y los enlaces entre los que se encuentran en el mismo patrón. La anchura de línea y los tamaños de nodo representan la frecuencia de recuentos globales para el conjunto. El gráfico representa visualmente la misma información que lo que se ha seleccionado en los paneles de patrones. Además de los nombres de tipo que aparecen en el gráfico, los tipos también se identifican ya sea por su color o por icono tipo, dependiendo de lo que haya seleccionado en la barra de herramientas de gráfico. Para obtener más información, consulte el tema [“Uso de barras de herramientas y paletas de gráficos”](#) en la página 163.

Uso de barras de herramientas y paletas de gráficos

En cada gráfico existe una barra de herramientas que ofrece acceso rápido a algunas paletas comunes en las que podrá realizar un número de acciones en los gráficos. Cada vista (Categorías y Conceptos, Clústeres y Análisis de enlace de texto) tiene una barra de herramientas ligeramente diferente. Puede elegir entre el modo de vista *Exploración* o el modo de vista *Edición*.

Mientras el modo de exploración le permite explorar analíticamente los datos y valores representados por la visualización, el modo de edición le permite cambiar el diseño y aspecto de la visualización. Por ejemplo, puede cambiar las fuentes y colores para que coincidan con la guía de estilo de su organización. Para seleccionar este modo, seleccione **Ver > Panel Visualización > Modo edición** de los menús (o pulse en el icono de la barra de herramientas).

En el modo de edición hay varias barras de herramientas que afectan a distintos aspectos del diseño de la visualización. Si no utiliza algunas de ellas, puede ocultarlas para ampliar el espacio del cuadro de diálogo en el que aparece el gráfico. Para seleccionar o anular la selección de las barras de herramientas, pulse en el nombre de la barra de herramientas o paleta correspondiente en el menú Ver.

Si desea más información sobre todas las barras de herramientas generales y paletas utilizadas para editar gráficos, consulte la sección sobre la Edición de visualizaciones en la ayuda en línea o en el archivo *ModelerSPOnodes.pdf*, que está disponible como parte de la descarga del producto.

Tabla 38. Botones de la barra de herramientas de Text Analytics	
Botón/Lista	Descripción
	Activa el modo de edición. Pase al modo de edición para cambiar el aspecto del gráfico: puede agrandar la fuente, cambiar los colores para que se adapten al estilo corporativo de su empresa, o eliminar etiquetas y leyendas.
	Activa el modo de exploración. De forma predeterminada, el modo de exploración está activado, lo que significa que puede mover y arrastrar nodos por el gráfico y pasar el ratón sobre los objetos del gráfico para revelar información adicional.

Tabla 38. Botones de la barra de herramientas de Text Analytics (continuación)

Botón/Lista	Descripción
	Seleccione un tipo de malla que mostrar para los gráficos en la vista Categorías y Conceptos, así como en la vista Análisis de enlace de texto. • Diseño circular Diseño general que puede aplicarse a cualquier gráfico. Organiza un gráfico teniendo en cuenta que los enlaces no tienen dirección y trata a todos los nodos por igual. Los nodos solo se colocan alrededor del perímetro de un círculo. • Diseño de red Diseño general que puede aplicarse a cualquier gráfico. Organiza un gráfico teniendo en cuenta que los enlaces no tienen dirección y trata a todos los nodos por igual. Los nodos se colocan libremente por el diseño. • Diseño direccional Un diseño que solo debería utilizarse para gráficos direccionales. Este enlace genera estructuras de árbol a partir de los nodos raíz hasta los nodos ramales, y las organiza por colores. Los datos jerárquicos tienden a visualizarse perfectamente con este diseño. • Diseño de cuadrícula Diseño general que puede aplicarse a cualquier gráfico. Organiza un gráfico teniendo en cuenta que los enlaces no tienen dirección y trata a todos los nodos por igual. Los nodos solo se colocan en los puntos de la cuadrícula en el espacio.
	Representación del tamaño de enlace. Seleccione qué grosor de línea se presenta en el gráfico. Esto solo es aplicable a la vista Clústeres. El gráfico de malla de Clústeres solo muestra el número de enlaces externos entre clústeres. Puede elegir entre: • Similitud El grosor indica el número de enlaces externos entre dos clústeres • Co-ocurrencia El grosor indica el número de documentos en el que tiene lugar una coocurrencia de descriptores.
	Botón de conmutación que, cuando se pulsa, muestra la leyenda. Cuando el botón no está pulsado, la leyenda queda oculta.
	Botón de conmutación que, al pulsarlo, muestra los iconos de tipo en el gráfico, en lugar de los colores de tipo. Esto solo es aplicable a la vista Análisis de enlace de texto.
	Botón de conmutación que, cuando se pulsa, muestra el Control deslizante de enlaces bajo el gráfico. Puede filtrar los resultados deslizando la flecha.
	Se mostrará el gráfico para el nivel superior de categorías seleccionadas en lugar de sus subcategorías.
	Se mostrará el gráfico para el nivel inferior de categorías seleccionadas.
	Esta opción controla cómo se muestran los nombres de las subcategorías en la salida. • Ruta de categoría completa Esta opción mostrará el nombre de la categoría y la vía de acceso completa de las categorías principales mediante la utilización de barras diagonales para separar los nombres de las categorías de los nombres de las subcategorías. • Ruta de categoría abreviada Esta opción mostrará únicamente el nombre de la categoría utilizando puntos suspensivos para mostrar el número de categorías principales para esa categoría en cuestión. • Categoría de nivel inferior Esta opción mostrará únicamente el nombre de la categoría sin mostrar la vía de acceso completa o las categorías principales.

Capítulo 13. Editor de recursos de sesión

IBM SPSS Modeler Text Analytics captura rápidamente y con precisión y extrae los conceptos clave de los datos de texto. Este proceso de extracción se basa mayoritariamente en los recursos lingüísticos para determinar de qué forma se extrae la información de los datos de texto. De forma predeterminada, estos recursos provienen de plantillas de recursos.

IBM SPSS Modeler Text Analytics se suministra con un conjunto de **plantillas de recursos** especializadas que contienen un conjunto de recursos lingüísticos y no lingüísticos en forma de bibliotecas y recursos avanzados, para ayudar a definir cómo se extraerán y manejarán los datos. Para obtener más información, consulte el tema [Capítulo 14, “Plantillas y recursos”](#), en la página 169.

En el cuadro de diálogo del nodo, puede cargar una copia de los recursos de la plantilla en el nodo. Una vez dentro de una sesión de área de trabajo interactiva, puede personalizar estos recursos específicamente para los datos de este nodo, si lo desea. Durante una sesión de área de trabajo interactiva, puede trabajar con los recursos en la vista Editor de recursos. Cuando se inicia una sesión interactiva, se lleva a cabo una extracción utilizando los recursos cargados en el cuadro de diálogo del nodo, a menos que haya almacenado en memoria caché los datos y los resultados de la extracción en el nodo.

Edición de recursos en el editor de recursos

Editor de recursos ofrece acceso al conjunto de recursos que se utilizan para generar los resultados de la extracción (conceptos, tipos y patrones) para una sesión de área de trabajo interactiva. Este editor es muy similar al de, Editor de plantillas, salvo que en el Editor de recursos está editando los recursos para esta sesión. Cuando haya terminado de trabajar en los recursos y en cualquier otro trabajo que haya hecho, puede actualizar el nodo de modelado para guardar este trabajo de modo que pueda ser restaurado en una sesión de área de trabajo interactiva posterior. Para obtener más información, consulte el tema [“Guardar y actualizar nodos de modelado”](#) en la página 77.

Si desea trabajar directamente en las plantillas utilizadas para cargar recursos en nodos, le recomendamos que utilice el Editor de plantillas. Muchas de las tareas que se pueden realizar en el interior del Editor de recursos son realizadas como están en el Editor de plantillas, como por ejemplo:

- **Trabajar con bibliotecas.** Para obtener más información, consulte el tema [Capítulo 15, “Trabajo con bibliotecas”](#), en la página 179.
- **Crear diccionarios de tipo.** Para obtener más información, consulte el tema [“Crear tipos”](#) en la página 189.
- **Añadir términos a los diccionarios.** Para obtener más información, consulte el tema [“Añadir términos”](#) en la página 190.
- **Crear sinónimos.** Para obtener más información, consulte el tema [“Definir sinónimos”](#) en la página 195.
- **Importar y exportar plantillas.** Para obtener más información, consulte el tema [“Importación y exportación de plantillas”](#) en la página 175.
- **Publicar bibliotecas.** Para obtener más información, consulte el tema [“Publicación de bibliotecas”](#) en la página 185.

Para textos en neerlandés, inglés, francés, alemán, italiano, portugués y español

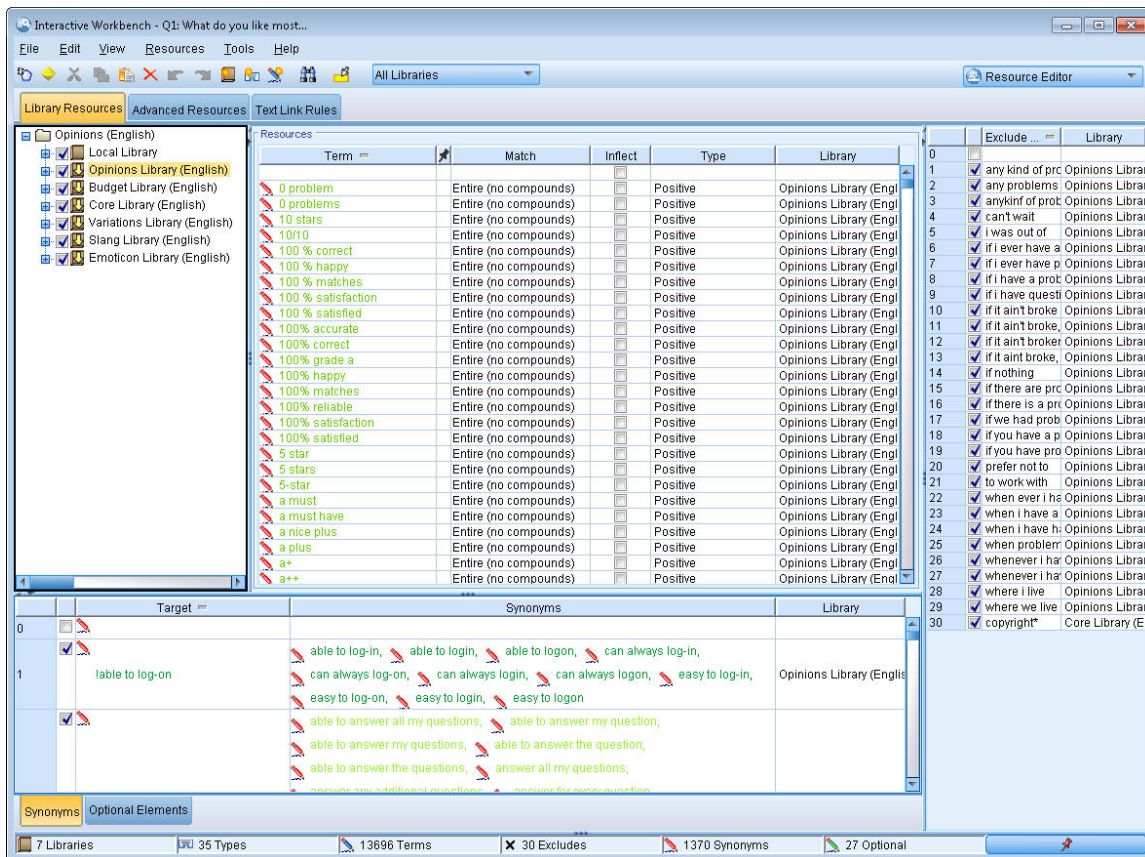


Figura 38. Vista del editor de recursos

Creación y actualización de plantillas

Cuando se realizan cambios en los recursos y prevé que los utilizará en otro momento, puede guardar los recursos en forma de plantilla. Al hacerlo, puede optar por guardar utilizando un nombre de plantilla existente o asignando un nombre nuevo. Más adelante, cuando quiera cargar esta plantilla, podrá disponer de los mismos resultados. Para obtener más información, consulte el tema [“Copiar recursos de plantillas y TAP”](#) en la página 26.

Nota: También puede publicar y compartir sus bibliotecas. Para obtener más información, consulte el tema [“Compartimiento de bibliotecas”](#) en la página 184.

Para crear (o actualizar) una plantilla

1. En los menús de la vista Editor de recursos, seleccione **Recursos > Crear plantilla de recurso**. Aparecerá el cuadro de diálogo Crear plantilla de recursos.
2. Escriba un nombre nuevo en el campo Nombre de plantilla si desea crear una plantilla nueva. Seleccione una plantilla de la tabla si desea sobrescribir una plantilla existente con los recursos actualmente cargados.
3. Pulse **Guardar** para crear la plantilla.

Importante: Puesto que las plantillas se cargan cuando las selecciona en el nodo y no cuando se ejecuta la ruta, asegúrese de que vuelve a cargar la plantilla de recursos en los otros nodos en los que se utilice si desea obtener los últimos cambios. Para obtener más información, consulte el tema [“Actualización de los recursos en el nodo después de cargar”](#) en la página 174.

Conmutar entre plantillas de recursos

Si desea sustituir los recursos actualmente cargados en la sesión por una copia de los recursos de otra plantilla, puede cambiar a esos otros recursos. Al hacerlo se sobrescribirán los recursos actualmente

cargados en la sesión. Si desea cambiar de recursos para obtener algunas reglas de patrones TLA (Análisis de enlace de texto) predefinidas, asegúrese de seleccionar una plantilla que tenga marcada la columna TLA.

El cambio de recursos resulta particularmente útil cuando desea restaurar el trabajo de la sesión (categorías, patrones y recursos) pero quiere cargar una copia actualizada de los recursos a partir de una plantilla sin perder el trabajo de la otra sesión. Puede seleccionar la plantilla cuyo contenido desea copiar en el Editor de recursos y pulsar **Aceptar**. Esto sustituye los recursos que tiene en esta Sesión. Asegúrese de actualizar el nodo de modelado del final de su sesión si desea mantener estos cambios la próxima vez que inicie la sesión de área de trabajo interactiva.

Nota: Si conmuta al contenido de otra plantilla durante una sesión interactiva, el nombre de la plantilla listada en el nodo seguirá siendo el nombre de la última plantilla cargada y copiada. Para poder beneficiarse de estos recursos o del trabajo de otra sesión, actualice el nodo de modelado antes de salir de la sesión y seleccione la opción **Utilizar trabajo de sesión** en el nodo. Para obtener más información, consulte el tema [“Guardar y actualizar nodos de modelado”](#) en la página 77.

Conmutar entre recursos

1. En los menús de la vista Editor de recursos, seleccione **Recursos > Cambiar plantillas de recursos**. Aparecerá el cuadro de diálogo Cambiar de recursos.
2. Seleccione la plantilla que desea utilizar entre las que aparecen en la tabla.
3. Pulse **Aceptar** para salir de los recursos actualmente cargados y cargar una copia de los recursos de la plantilla seleccionada. Si ha realizado cambios en los recursos y desea guardar las bibliotecas para su uso futuro, puede publicarlas, actualizarlas y compartirlas antes de realizar el cambio. Para obtener más información, consulte el tema [“Compartimiento de bibliotecas”](#) en la página 184.

Capítulo 14. Plantillas y recursos

IBM SPSS Modeler Text Analytics captura rápidamente y con precisión y extrae los conceptos clave de los datos de texto. Este proceso de extracción se basa mayoritariamente en los recursos lingüísticos para determinar de qué forma se extrae la información de los datos de texto. Para obtener más información, consulte el tema [“Cómo funciona la extracción” en la página 5](#). Puede ajustar estos recursos en la vista Editor de recursos.

Cuando se instala el software, también se obtiene un conjunto de recursos especializados. Estos recursos incluidos permiten que se beneficie de muchos años de investigación y ajustes para idiomas específicos y aplicaciones específicas. Puesto que no siempre los recursos incluidos se adaptan a la perfección al contexto de sus datos, puede editar estas plantillas de recursos o incluso crear y utilizar bibliotecas personalizadas exclusivamente adaptadas a los datos de su organización. Estos recursos vienen en diversas formas y cada uno puede utilizarse en la sesión. Los recursos pueden encontrarse en los siguientes lugares:

- **Plantilla de recursos.** Las plantillas se componen de un conjunto de bibliotecas, tipos y algunos recursos avanzados que juntos conforman un conjunto especializado de recursos que se adapta a un dominio o contexto particular, como las opiniones sobre productos.
- **Paquetes de análisis de texto (TAP).** Además de los recursos almacenados en una plantilla, los TAP unen además uno o más conjuntos de categorías especializados que se han generado utilizando esos recursos, de modo que tanto las categorías como los recursos se almacenan y reutilizan conjuntamente. Para obtener más información, consulte el tema [“Uso de los paquetes de análisis de texto” en la página 135](#).
- **Bibliotecas.** Las bibliotecas se utilizan como los cimientos sobre los que se basan tanto los TAP como las plantillas. También pueden añadirse individualmente a los recursos en la sesión. Cada biblioteca está compuesta por varios diccionarios que se utilizan para definir y administrar listas de tipos, sinónimos y exclusiones. Aunque las bibliotecas también se entregan individualmente, se empaquetan juntas en plantillas y en los TAP. Para obtener más información, consulte el tema [Capítulo 15, “Trabajo con bibliotecas”](#), en la página 179.

Nota: durante la extracción, también se utilizan algunos recursos internos compilados. Estos recursos compilados contienen un gran número de definiciones que complementan a los tipos de la biblioteca Core. Estos recursos compilados no se pueden editar.

El Editor de recursos ofrece acceso al conjunto de recursos que se utilizan para generar los resultados de la extracción (conceptos, tipos y patrones). Existe una serie de tareas que puede realizar en el Editor de recursos, e incluyen:

- **Trabajar con bibliotecas.** Para obtener más información, consulte el tema [Capítulo 15, “Trabajo con bibliotecas”](#), en la página 179.
- **Crear diccionarios de tipo.** Para obtener más información, consulte el tema [“Crear tipos” en la página 189](#).
- **Añadir términos a los diccionarios.** Para obtener más información, consulte el tema [“Añadir términos” en la página 190](#).
- **Crear sinónimos.** Para obtener más información, consulte el tema [“Definir sinónimos” en la página 195](#).
- **Actualizar los recursos en los TAP.** Para obtener más información, consulte el tema [“Actualización de los paquetes de análisis de texto” en la página 137](#).
- **Crear plantillas.** Para obtener más información, consulte el tema [“Creación y actualización de plantillas” en la página 166](#).
- **Importar y exportar plantillas.** Para obtener más información, consulte el tema [“Importación y exportación de plantillas” en la página 175](#).
- **Publicar bibliotecas.** Para obtener más información, consulte el tema [“Publicación de bibliotecas” en la página 185](#).

El editor de plantillas frente al editor de recursos

Existen dos métodos principales para trabajar con y editar las plantillas, bibliotecas y sus recursos. Puede trabajar con recursos lingüísticos en el Editor de plantillas o en el Editor de recursos.

Editor de plantillas

El Editor de plantillas permite crear y editar plantillas de recursos sin una sesión de entorno de trabajo interactivo e independiente de un nodo o ruta específicos. Puede utilizar este editor para crear o editar plantillas de recursos antes de cargarlas en el nodo de Análisis de enlace de texto y en el nodo de modelado de Text Mining.

Se puede acceder a Editor de plantillas a través de la barra de herramientas principal de IBM SPSS Modeler desde el menú **Herramientas > Editor de plantillas de análisis de texto**

Editor de recursos

El Editor de recursos, accesible desde una sesión de entorno de trabajo interactivo, permite trabajar con los recursos en el contexto de un nodo y un conjunto de datos específicos. Cuando añade un nodo de modelado Text Mining a una ruta, puede cargar una copia del contenido de la plantilla de recursos o una copia de un paquete de análisis de texto (conjuntos de categorías y recursos) para controlar cómo se extrae el texto para la minería de textos. Cuando se inicia una sesión de entorno de trabajo interactivo, además de crear categorías, extraer patrones de análisis de enlace de texto y crear modelos de categorías, también pueden ajustarse los recursos correspondientes a los datos de esa sesión en la vista integrada del Editor de recursos. Para obtener más información, consulte el tema [“Edición de recursos en el editor de recursos”](#) en la página 165.

Cuando trabaja con los recursos en una sesión de entorno de trabajo interactivo, dichos cambios se aplican únicamente a esa sesión. Si desea guardar su trabajo (recursos, categorías, patrones, etc.) para poder continuar en una sesión posterior, debe actualizar el nodo de modelado. Para obtener más información, consulte el tema [“Guardar y actualizar nodos de modelado”](#) en la página 77.

Si desea guardar los cambios en la plantilla original (cuyo contenido se copió en el nodo de modelado) para que la plantilla actualizada pueda cargarse en otros nodos, puede crear una plantilla a partir de los recursos. Para obtener más información, consulte el tema [“Creación y actualización de plantillas”](#) en la página 166.

Nota: Si realiza cambios en plantillas o bibliotecas y las guarda en un directorio de copia de seguridad y, después, actualiza la versión de IBM SPSS Modeler Text Analytics, se le dará la opción de importar las plantillas y bibliotecas personalizadas. La primera vez que ejecute una ruta SPSS Modeler Text Analytics o abra el Editor de recursos después de una actualización, las plantillas y bibliotecas predeterminadas se copian en la máquina. Se muestra un aviso **Plantillas guardadas** o un aviso **Bibliotecas guardadas** (o ambos) con una lista de las plantillas y/o bibliotecas que se han actualizado como parte de la actualización del producto, y se le proporciona la opción de importar las plantillas y bibliotecas personalizadas del directorio en el cual las ha guardado. Tras pulsar **Aceptar** en el mensaje de aviso, puede abrir el diálogo **Gestionar plantillas de recursos** o el diálogo **Gestionar bibliotecas** en cualquier momento para elegir qué plantillas o bibliotecas personalizadas desea importar.

La interfaz Editor

Las operaciones que realiza en el Editor de plantillas o Editor de recursos se concentran en torno a la administración y ajuste de los recursos lingüísticos. Estos recursos se almacenan en forma de plantillas y bibliotecas. Para obtener más información, consulte el tema [“Diccionarios de tipo”](#) en la página 187.

Pestaña Recursos de la biblioteca

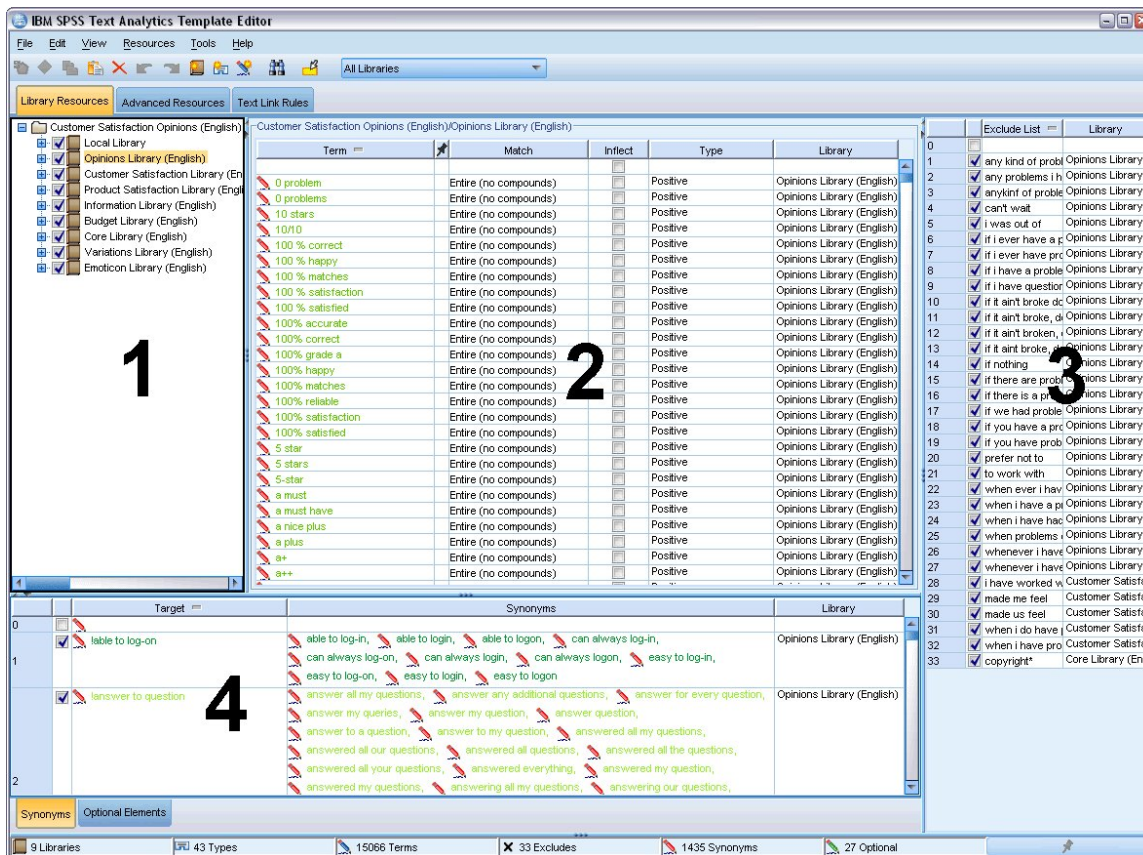


Figura 39. Editor de plantillas de minería de textos

La interfaz está organizada en cuatro partes:

1. Panel de árbol de bibliotecas. Situada en el ángulo superior izquierdo, este panel muestra un árbol con las bibliotecas. Puede activar y desactivar las bibliotecas de este árbol y filtrar las vistas en los otros paneles seleccionando una biblioteca del árbol. Puede realizar muchas operaciones en este árbol utilizando los menús contextuales. Si expande una biblioteca en el árbol, podrá ver el conjunto de tipos que contiene. También puede filtrar esta lista a través del menú **Ver** si desea centrarse únicamente en una biblioteca determinada.

2. Lista de términos del panel Diccionarios de tipo. Situado a la derecha del árbol de bibliotecas, este panel muestra las listas de términos de los diccionarios de tipo de las bibliotecas seleccionadas en el árbol. Un **diccionario de tipo** es una colección de términos que se agrupan bajo una misma etiqueta, tipo o nombre. Cuando el motor de extracción lee los datos de texto, compara las palabras que encuentra en el texto con los términos definidos en los diccionarios de tipo. Si un concepto extraído aparece como término en un diccionario de tipo, se asignará ese nombre de tipo. Puede considerar que el diccionario de tipo es un diccionario específico de términos que tienen algo en común. Por ejemplo, el tipo <Location> de la biblioteca base contiene conceptos como new orleans, great britain y new york. Todos estos términos representan ubicaciones geográficas. Una biblioteca puede contener uno o más diccionarios de tipo. Para obtener más información, consulte el tema “Diccionarios de tipo” en la página 187.

3. Panel de diccionario Excluir. Situado en el lado derecho, este panel muestra la colección de términos que se excluirán de los resultados de extracción finales. Los términos que aparecen en este diccionario de exclusión no aparecen en el panel Resultados extraídos. Los términos excluidos pueden almacenarse en la biblioteca que usted elija. Sin embargo, el panel de diccionario de exclusión muestra todos los términos excluidos para todas las bibliotecas visibles en el árbol de bibliotecas. Consulte el tema “Diccionarios de exclusión” en la página 197 para obtener más información.

4. Panel de diccionario de sustituciones. Situado en la parte inferior izquierda, este panel muestra los sinónimos y los elementos opcionales, cada uno en su propia pestaña. Los sinónimos y los elementos

opcionales ayudan a agrupar términos similares bajo un concepto principal o destino en los resultados de extracción finales. Este diccionario puede contener sinónimos conocidos y sinónimos definidos por el usuario y elementos, así como los errores ortográficos más comunes emparejados con la ortografía correcta. Las definiciones de sinónimos y los elementos opcionales pueden almacenarse en la biblioteca que elija. Sin embargo, el panel del diccionario de sustitución muestra todos los contenidos de todas las bibliotecas visibles en el árbol de bibliotecas. Mientras que este panel muestra todos los sinónimos o elementos opcionales de todas las bibliotecas, las sustituciones para todas las bibliotecas del árbol se muestran conjuntamente en este panel. Una biblioteca puede contener tan solo un diccionario de sustitución. Consulte el tema [“Diccionarios de sustitución/sinónimos”](#) en la [página 194](#) para obtener más información.

Notas:

- Si desea filtrar esta ventana de manera que solo se vea la información que pertenece a una única biblioteca, puede cambiar la vista de la biblioteca mediante la lista desplegable de la barra de herramientas. Contiene una entrada de nivel superior llamada **Todas las bibliotecas** así como una entrada adicional para cada biblioteca individual. Para obtener más información, consulte el tema [“Visión de bibliotecas”](#) en la [página 182](#).

Pestaña Recursos avanzados

Los recursos avanzados están disponibles en la segunda pestaña de la vista del editor. Puede revistar y editar los recursos avanzados en esta pestaña. Para obtener más información, consulte el tema [Capítulo 17, “Acerca de los recursos avanzados”](#), en la [página 199](#).

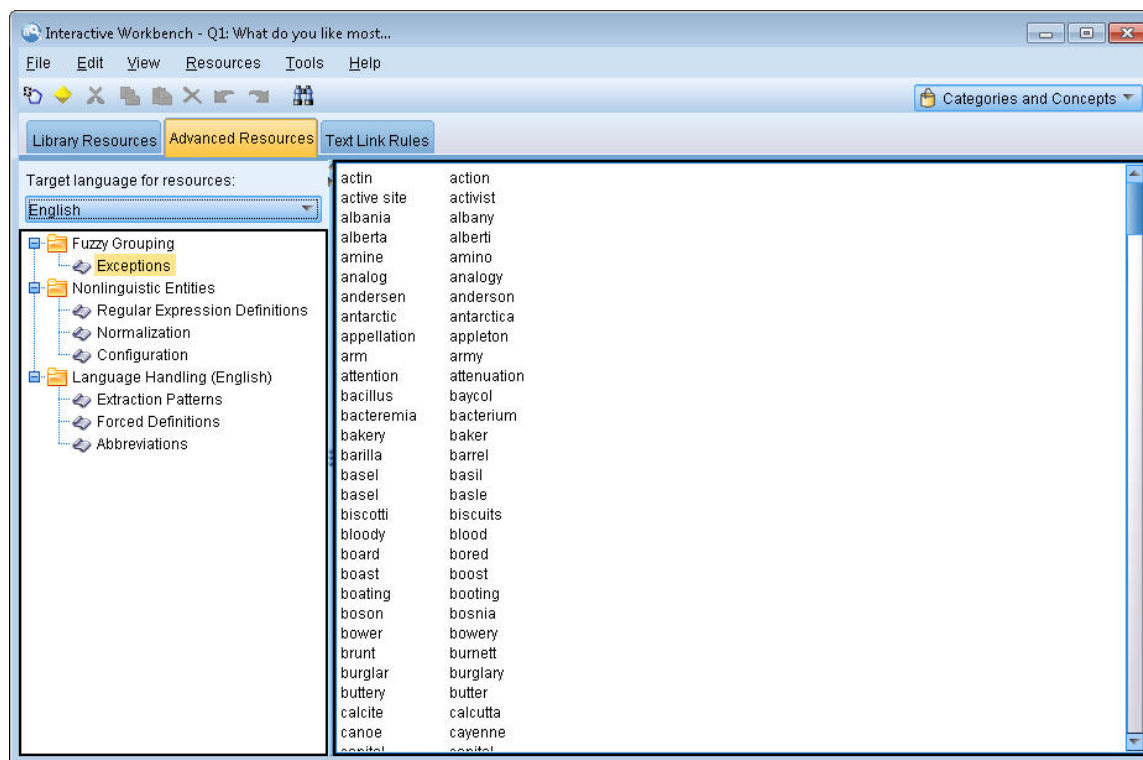


Figura 40. Editor de plantillas de minería de textos: pestaña Recursos avanzados

Pestaña Reglas de enlace de texto

A partir de la versión 14, las reglas de análisis de enlace de texto son editables en su propia pestaña de la vista del editor. Puede trabajar en el editor de reglas, crear sus propias reglas e incluso ejecutar simulaciones para ver cómo impactan sus reglas en los resultados TLA. Para obtener más información, consulte el tema [Capítulo 18, “Sobre las reglas de enlaces de texto”](#), en la [página 211](#).

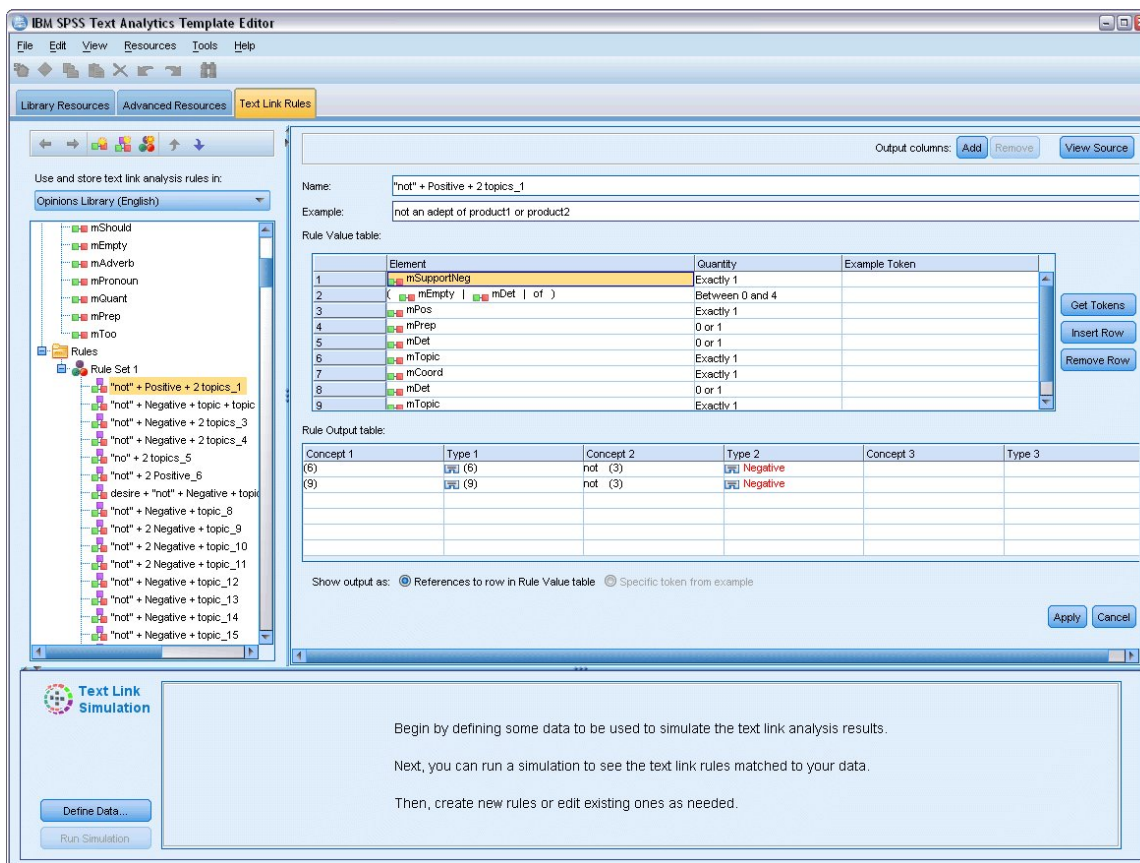


Figura 41. Editor de plantillas de minería de textos - Pestaña Reglas de enlace de texto

Apertura de plantillas

Cuando inicia el Editor de plantillas, se le indica que abra una plantilla. De igual modo, puede abrir una plantilla desde el menú Archivo. Si desea una plantilla que contenga algunas reglas de análisis de enlace de texto (TLA), asegúrese de que selecciona una plantilla que tenga un icono en la columna TLA. El idioma en el que se creó la plantilla se muestra en la columna Idioma.

Si desea importar una plantilla que no se muestra en la tabla, o si desea exportar una plantilla, puede utilizar los botones del cuadro de diálogo Abrir plantilla. Para obtener más información, consulte el tema “Importación y exportación de plantillas” en la página 175.

Para abrir una plantilla

1. En los menús de Editor de plantillas, seleccione **Archivo > Abrir plantilla de recurso**. Aparecerá el cuadro de diálogo Abrir plantilla de recursos.
2. Seleccione la plantilla que desea utilizar entre las que aparecen en la tabla.
3. Pulse **Aceptar** para abrir la plantilla. Si en ese momento tiene otra plantilla abierta en el editor, al pulsar en Aceptar se saldrá de dicha plantilla y se mostrará la que haya seleccionado aquí. Si ha realizado cambios en los recursos y desea guardar las bibliotecas para su uso futuro, puede publicarlas, actualizarlas y compartirlas antes de abrir otra. Para obtener más información, consulte el tema “Compartimiento de bibliotecas” en la página 184.

Guardado de plantillas

En el Editor de plantillas, puede guardar los cambios realizados en una plantilla. Puede optar por guardar utilizando un nombre de plantilla existente o asignando un nombre nuevo.

Si realiza cambios en una plantilla que ya haya cargado en un nodo anteriormente, tendrá que volver a cargar el contenido de la plantilla en el nodo para obtener los últimos cambios. Para obtener más información, consulte el tema [“Copiar recursos de plantillas y TAP”](#) en la página 26.

O si utiliza la opción **Utilizar trabajo interactivo guardado** en la pestaña Modelo del nodo de Text Mining, lo que significa que está utilizando recursos de una sesión de entorno de trabajo interactivo anterior, necesitará cambiar a los recursos de esta plantilla desde la sesión de entorno de trabajo interactivo. Para obtener más información, consulte el tema [“Conmutar entre plantillas de recursos”](#) en la página 166.

Nota: También puede publicar y compartir sus bibliotecas. Para obtener más información, consulte el tema [“Compartimiento de bibliotecas”](#) en la página 184.

Para guardar una plantilla

1. En los menús de Editor de plantillas, seleccione **Archivo > Guardar plantilla de recurso**. Aparecerá el cuadro de diálogo Guardar plantilla de recursos.
2. Escriba un nombre nuevo en el campo Nombre de plantilla si desea guardarla como plantilla nueva. Seleccione una plantilla de la tabla si desea sobrescribir una plantilla existente con los recursos actualmente cargados.
3. Escriba, si lo desea, una descripción para que se muestre un comentario u anotación en la tabla.
4. Pulse **Guardar** para guardar la plantilla.

Importante: Ya que los recursos de las plantillas o de los TAP se cargan o copian en el nodo, deberá actualizar los recursos volviéndolos a cargar si introduce cambios en una plantilla y si quiere beneficiarse de estos cambios en una ruta existente. Para obtener más información, consulte el tema [“Actualización de los recursos en el nodo después de cargar”](#) en la página 174.

Actualización de los recursos en el nodo después de cargar

De forma predeterminada, cuando añade un nodo a una ruta, se cargará e incrustará en su nodo un conjunto de recursos de una plantilla predeterminada. Y si cambia las plantillas o utiliza un TAP, al cargar los recursos, una copia de estos recursos los sobrescribirá. Ya que las plantillas y los TAP no están vinculados directamente a un nodo, cualquier cambio que introduzca en una plantilla o TAP no estará disponible de manera automática en un nodo ya existente previamente. Para poder beneficiarse de estos cambios, debería actualizar los recursos en ese nodo. Los recursos pueden actualizarse de dos formas.

Método 1: recarga de recursos en la pestaña Modelo

Si desea actualizar los recursos en el nodo utilizando una plantilla nueva o actualizada o un TAP, puede volver a cargarla en la pestaña Modelo del nodo. Al volver a cargarla, se reemplaza la copia de los recursos en el nodo por una copia más actual. Para su comodidad, aparecerá la fecha y la hora de la actualización en la pestaña Modelo junto con el nombre de la plantilla originaria. Para obtener más información, consulte el tema [“Copiar recursos de plantillas y TAP”](#) en la página 26.

Sin embargo, si trabaja con datos de una sesión interactiva en un nodo de modelado de Text Mining y ha seleccionado la opción **Utilizar trabajo de sesión** en la pestaña Modelo, se utilizarán los recursos y el trabajo de la sesión guardada y el botón **Cargar** quedará desactivado. Se desactiva porque, en un momento dado durante una sesión de área de trabajo interactiva, eligió la opción **Actualizar nodo de modelado** y guardó las categorías, los recursos y otro trabajo de la sesión. En ese caso, si desea cambiar o actualizar los recursos, puede intentar el método siguiente de cambio de recursos en el Editor de recursos.

Método 2: cambio de recursos en el Editor de recursos

Cuando quiera utilizar recursos diferentes durante una sesión interactiva, puede cambiar esos recursos utilizando el cuadro de diálogo Cambio de recursos. Esto resulta especialmente útil si desea volver a utilizar un trabajo de categoría existente pero reemplazado los recursos. En ese caso, puede seleccionar la opción **Utilizar trabajo de sesión** en la pestaña Modelo del nodo de modelado Text Mining. Si lo hace se desactivará la capacidad de volver a cargar una plantilla a través del cuadro de diálogo del nodo y se mantendrá en su lugar la configuración y los cambios que realice durante su sesión. A continuación puede iniciar la sesión de área de trabajo interactiva ejecutando la ruta y cambiar los recursos en el Editor de

recursos. Para obtener más información, consulte el tema [“Conmutar entre plantillas de recursos”](#) en la página 166.

Con el fin de mantener el trabajo de la sesión en sesiones posteriores, incluidos los recursos, deberá actualizar el nodo de modelado desde la sesión de área de trabajo interactiva para que los recursos (y otros datos) se guarden en el nodo. Para obtener más información, consulte el tema [“Guardar y actualizar nodos de modelado”](#) en la página 77.

Nota: si cambia al contenido de otra plantilla durante una sesión interactiva, el nombre de la plantilla listada en el nodo seguirá siendo el nombre de la plantilla cargada y copiada más recientemente. Para poder beneficiarse de estos recursos o del trabajo de otra sesión, actualice el nodo de modelado antes de salir de la sesión.

Administración de plantillas

Existen también algunas tareas básicas de administración que puede realizar de vez en cuando en las plantillas, como cambiarles el nombre, importar y exportar plantillas o eliminar las que hayan quedado obsoletas. Estas tareas se realizan en el cuadro de diálogo Administrar plantillas. La importación y exportación de plantillas permite compartir plantillas con otros usuarios. Para obtener más información, consulte el tema [“Importación y exportación de plantillas”](#) en la página 175.

Nota: no puede renombrar o suprimir las plantillas que están instaladas (o que se entregan) con este producto. En su lugar, si desea cambiarles el nombre, abra la plantilla instalada y cree una nueva con el nombre que desee. Puede eliminar las plantillas personalizadas, pero si intenta eliminar una de las plantillas incorporadas, se restaurará a la versión originalmente instalada.

Para cambiar el nombre de una plantilla

1. En los menús, seleccione **Recursos > Gestionar plantillas de recursos**. Aparecerá el cuadro de diálogo Administrar plantillas.
2. Seleccione la plantilla a la que desea cambiar el nombre y pulse en **Cambiar nombre**. El cuadro de nombre pasará a ser un campo editable en la tabla.
3. Teclee un nombre nuevo y pulse la tecla Intro. Se abrirá un cuadro de diálogo de confirmación.
4. Si está conforme con el cambio de nombre, pulse en **Sí**. De lo contrario, pulse en **No**.

Para eliminar una plantilla

1. En los menús, seleccione **Recursos > Gestionar plantillas de recursos**. Aparecerá el cuadro de diálogo Administrar plantillas.
2. En el cuadro de diálogo Administrar plantillas, seleccione la plantilla que desea eliminar.
3. Pulse **Eliminar**. Se abrirá un cuadro de diálogo de confirmación.
4. Pulse **Sí** para eliminar o en **No** para cancelar la solicitud. Si pulsa en **Sí**, la plantilla se elimina.

Importación y exportación de plantillas

Puede compartir plantillas con otros usuarios u otras máquinas importándolas y exportándolas. Las plantillas se guardan en una base de datos interna, pero pueden exportarse en forma de archivos *.lrt en el disco duro.

Puesto que en determinadas ocasiones querrá importar o exportar plantillas, existen diversos cuadros de diálogo que ofrecen estas capacidades.

- El cuadro de diálogo Abrir plantilla en el Editor de plantillas
- El cuadro de diálogo Cargar recursos en el nodo de modelado Text Mining y en el nodo Análisis para los enlaces de texto.
- En el cuadro de diálogo Administrar plantillas en el Editor de plantillas y en el Editor de recursos.

Para importar una plantilla

1. En el cuadro de diálogo, pulse en **Importar**. Aparecerá el cuadro de diálogo Importar plantilla.

2. Seleccione el archivo de la plantilla de recursos (*.lrt) para importar y pulse en **Importar**. Puede guardar la plantilla que está importando con otro nombre o sobrescribir la existente. El cuadro de diálogo se cierra y la plantilla aparece ahora en la tabla.

Para exportar una plantilla

1. En el cuadro de diálogo, seleccione la plantilla que desea exportar y pulse en **Exportar**. Aparecerá el cuadro de diálogo Seleccionar directorio.
2. Seleccione el directorio al que desea exportar y pulse en **Exportar**. Este cuadro de diálogo se cierra y se exporta la plantilla llevando la extensión de archivo (*.lrt)

Cómo salir del Editor de plantillas

Cuando termine de trabajar en el Editor de plantillas, puede guardar el trabajo y salir del editor.

Para salir del Editor de plantillas

1. En los menús, seleccione **Archivo > Cerrar**. Aparecerá el cuadro de diálogo Guardar y cerrar.
2. Seleccione **Guardar cambios en plantilla** para guardar la plantilla abierta antes de cerrar el editor.
3. Seleccione **Publicar bibliotecas** si quiere publicar las bibliotecas de la plantilla abierta antes de cerrar el editor. Si selecciona esta opción, se le pedirá que seleccione las bibliotecas que desea publicar. Para obtener más información, consulte el tema [“Publicación de bibliotecas” en la página 185](#).

Copia de seguridad de los recursos

Puede hacer una copia de seguridad de sus recursos periódicamente como medida de seguridad.

Importante: A la hora de restaurar, los contenidos completos de los recursos se limpiarán y sólo podrá accederse al contenido del archivo de seguridad en el producto. Esta acción incluye todos los trabajos abiertos.

Nota: sólo puede realizar copia de seguridad y restaurar con la misma versión principal de su software. Por ejemplo, si realiza copia de seguridad desde la versión 15, no podrá restaurar la copia de seguridad en la versión 16.

Para hacer una copia de seguridad de los recursos

1. En los menús, seleccione **Recursos > Herramientas de copia de seguridad > Copia de seguridad de recursos**. Aparecerá el cuadro de diálogo Copia de seguridad.
2. Introduzca un nombre para el archivo de copia de seguridad y pulse en **Guardar**. El cuadro de diálogo se cierra y se crea el archivo de copia de seguridad.

Para restaurar los recursos

1. En los menús, seleccione **Recursos > Herramientas de copia de seguridad > Restaurar recursos**. Un mensaje de alerta le advierte que al restaurar se sobrescriben todos los contenidos actuales de la base de datos.
2. Pulse **Sí** para continuar. Se abrirá el cuadro de diálogo.
3. Seleccione el archivo de copia de seguridad que desea restaurar y pulse en **Abrir**. El cuadro de diálogo se cierra y los recursos se restauran en la aplicación.

Importar archivos de recursos

Si ha introducido cambios directamente en los archivos de recursos fuera de este producto, puede importarlos a la biblioteca que quiera seleccionando esta biblioteca y procediendo con la importación. Cuando se importa a un directorio, también puede importar todos los archivos admitidos en una biblioteca abierta específica. Sólo puede importar archivos *.txt.

Cada archivo importado debe contener únicamente una entrada por línea, y si el contenido está estructurado así:

- Una lista de palabras o frases (una por línea). El archivo se importa como una lista de términos para un diccionario de tipo, donde el diccionario de tipo toma el nombre del archivo menos la extensión.
- Una lista de entradas, como `term1 <TAB> term2`, se importa a continuación como una lista de sinónimos, donde `term1` es el conjunto del término subyacente y `term2` es el término objetivo.

Importar un único archivo de recursos

1. En los menús, seleccione **Recursos > Importar archivos > Importar archivo único**. Aparecerá el cuadro de diálogo Importar archivo.
2. Seleccione el archivo que desea importar y pulse en **Importar**. El contenido del archivo se transforma en un formato interno y se añade a la biblioteca.

Importar todos los archivos de un directorio

1. En los menús, seleccione **Recursos > Importar archivos > Importar directorio completo**. Aparecerá el cuadro de diálogo Directorio de importación.
2. Seleccione la biblioteca en la que desee que se importen todos los archivos de recursos de la lista **Importar**. Si selecciona la opción **Por defecto**, se creará una biblioteca nueva con el mismo nombre del directorio.
3. Seleccione el directorio desde el que importar los archivos. Los subdirectorios no se leerán.
4. Pulse **Importar**. El cuadro de diálogo se cierra y el contenido de los archivos de recursos importados aparece ahora en el editor en forma de diccionarios y de archivos de recursos avanzados.

Capítulo 15. Trabajo con bibliotecas

Los recursos que utiliza el motor de extracción para extraer y agrupar términos de los datos de texto siempre contienen una o más bibliotecas. Puede ver el conjunto de bibliotecas en el árbol de bibliotecas ubicado en la parte superior izquierda de Editor de plantillas y Editor de recursos. Las bibliotecas se componen de tres tipos de diccionarios: Tipo, Substitución y Exclusión. Consulte el tema [Capítulo 16, “Acerca de los diccionarios de biblioteca”](#), en la página 187 para obtener más información.

La plantilla de recursos o los recursos del TAP que seleccionó incluye varias bibliotecas que le permiten empezar a extraer inmediatamente conceptos de los datos de texto. Sin embargo, puede crear y publicar sus propias bibliotecas para poder reutilizarlas. Para obtener más información, consulte el tema [“Publicación de bibliotecas”](#) en la página 185.

Por ejemplo, supongamos que suele trabajar con datos de texto relacionados con la industria del automóvil. Después de analizar los datos, decide que desea crear algunos recursos personalizados para gestionar el vocabulario o la jerga específica de dicha industria. Utilizar el Editor de plantillas, le permite crear una nueva plantilla, y en ella una biblioteca para extraer y agrupar términos automotrices. Dado que necesitará la información en esta biblioteca nuevamente, publique la biblioteca en un repositorio central, accesible en el recuadro de diálogo **Gestionar bibliotecas**, de forma que pueda reutilizarse independientemente en diferentes proyectos de .

Supongamos que también desea agrupar términos específicos de diferentes subsectores, como dispositivos electrónicos, motores, sistemas de refrigeración o incluso de un fabricante o mercado en particular. Puede crear una biblioteca para cada grupo y luego publicarlas para que puedan utilizarse en diferentes conjuntos de datos de texto. De esta forma, puede añadir las bibliotecas que mejor se correspondan con el contexto de los datos de texto.

Nota: Los recursos adicionales pueden configurarse y gestionarse en el separador Recursos avanzados. Algunos se aplican a todas las bibliotecas y administran entidades no lingüísticas, excepciones de agrupación difusa, etc. De forma adicional, puede editar las reglas patrón de análisis de enlace de texto, las cuales pertenecen a bibliotecas específicas, en el separador Reglas de enlace de texto. Para obtener más información, consulte el tema [Capítulo 17, “Acerca de los recursos avanzados”](#), en la página 199.

Bibliotecas suministradas

De forma predeterminada, se instalan varias bibliotecas con IBM SPSS Modeler Text Analytics. Puede utilizar estas bibliotecas preformateadas para acceder a miles de términos y sinónimos predefinidos, y a muchos tipos diferentes. A estas bibliotecas enviadas se les realizan ajustes según los diferentes dominios y están disponibles en varios idiomas distintos.

Existe un gran número de bibliotecas pero las que se utilizan habitualmente son las siguientes:

- **Biblioteca local.** Se utiliza para almacenar diccionarios definidos por el usuario. Es una biblioteca vacía que se añade de forma predeterminada a todos los recursos. También contiene un diccionario vacío de tipo. Es muy útil cuando se realizan cambios, o mejoras a los recursos directamente (como por ejemplo, agregar una palabra a un tipo) desde la vista de categorías y conceptos, vista de clústeres, y la vista análisis de enlace de texto. En este caso, estos cambios y reajustes se almacenan automáticamente en la primera biblioteca de la lista del árbol de bibliotecas, en el Editor de recursos; de forma predeterminada, es la *Biblioteca local*. No puede publicar esta biblioteca porque es específica para los datos de sesión. Si desea publicar el contenido, primero deberá cambiar el nombre de la biblioteca.
- **Biblioteca principal.** Se utiliza en la mayoría de los casos, pues contiene los cinco tipos básicos incorporados, que representan a personas, ubicaciones, organizaciones, productos y desconocido. Es posible que sólo vea unos cuantos términos en uno de sus diccionarios de tipo, pero los tipos que están representados en la biblioteca Core son en realidad complementos de los tipos más sólidos que se encuentran en los recursos compilados internos que se entregan junto con el producto de minería de textos. Estos recursos compilados internos contienen miles de términos por cada tipo. Por esta razón, aunque un término no pueda verse en la lista de términos del diccionario de tipo, todavía puede

extraerse y escribirse con un tipo Core. Esto explica cómo nombres como *Jorge* pueden extraerse y tipificarse como <Person> mientras que en el diccionario de tipo <Person> de la biblioteca Core solo aparece el nombre inglés *John*. Del mismo modo, si no incluye la biblioteca Core, puede que siga viendo estos tipos en los resultados de extracción, puesto que el motor de extracción utilizará los recursos compilados que contienen estos tipos.

- **Biblioteca de opiniones.** Se utiliza habitualmente para extraer opiniones y sentimientos procedentes de los datos de texto. Esta biblioteca incluye miles de palabras que representan actitudes, calificadores, y preferencias que cuando se usan en conjunción con otros términos indican una opinión sobre un tema. Esta biblioteca incluye un número de tipos, sinónimos y exclusiones incorporadas. También incluye un voluminoso conjunto de reglas de patrones que se utilizan para el análisis de enlace de texto. Para beneficiarse de las reglas de análisis de enlace de texto de esta biblioteca y de los resultados de patrones que producen, debe especificarse esta biblioteca en la pestaña Reglas de enlace de texto. Consulte el tema [Capítulo 18, “Sobre las reglas de enlaces de texto”](#), en la [página 211](#) para obtener más información.
- **Biblioteca Budget.** Se utiliza para extraer los términos relacionados con el coste de las cosas. Esta biblioteca incluye muchas palabras y frases que representan adjetivos, calificadores y juicios sobre el precio o la calidad de las cosas.
- **Biblioteca variaciones.** Se utiliza para incluir casos donde algunas variaciones del idioma requieren definiciones de sinónimos para poder agruparlas adecuadamente. Esta biblioteca solo contiene definiciones de sinónimos.

Aunque algunas de las bibliotecas enviadas fuera de las plantillas tienen un contenido similar al de algunas plantillas, éstas se han ajustado específicamente a aplicaciones determinadas y contienen recursos avanzados adicionales. Es recomendable que intente utilizar una plantilla que se haya diseñado para el tipo de datos de texto con el que está trabajando y realice sus cambios en aquellos recursos en lugar de añadir simplemente bibliotecas individuales a una plantilla más general.

Los recursos compilados también se entregan con IBM SPSS Modeler Text Analytics. Siempre se utilizan durante el proceso de extracción y contienen un gran número de definiciones complementarias a los diccionarios de tipo incorporados en las bibliotecas predeterminadas. Puesto que estos recursos están compilados, no pueden verse ni editarse. Sin embargo, puede forzar que un término especificado en los recursos compilados se coloque en otro diccionario. Para obtener más información, consulte el tema [“Forzar términos”](#) en la [página 192](#).

Creación de bibliotecas

Puede crear un número indefinido de bibliotecas. Después de crear una biblioteca nueva, puede empezar a crear diccionarios de tipo dentro de la biblioteca e introducir términos, sinónimos y exclusiones.

Para crear una biblioteca

1. En los menús, seleccione **Recursos > Nueva biblioteca**. El diálogo Propiedades de biblioteca se abre.
2. Introduzca un nombre para la biblioteca en el cuadro de texto Nombre.
3. Si lo desea, introduzca un comentario en el cuadro de texto Anotación.
4. Pulse **Publicar** si desea publicar esta biblioteca ahora antes de introducir nada más en ella. Para obtener más información, consulte el tema [“Compartimiento de bibliotecas”](#) en la [página 184](#). También puede publicar más tarde en cualquier momento.
5. Pulse **Aceptar** para crear la biblioteca. El cuadro de diálogo se cierra y la biblioteca aparece en la vista de árbol. Si expande las bibliotecas del árbol, comprobará que en la biblioteca se ha incluido automáticamente un diccionario de tipo vacío. Puede empezar a añadir términos en él inmediatamente. Para obtener más información, consulte el tema [“Añadir términos”](#) en la [página 190](#).

Añadir bibliotecas públicas

Si desea volver a utilizar una biblioteca de otro dato de sesión, puede añadirla sus recursos actuales mientras que sea una biblioteca pública. Una *biblioteca pública* es aquella que ha sido publicada. Para obtener más información, consulte el tema [“Publicación de bibliotecas”](#) en la [página 185](#).

Cuando se añade una biblioteca pública, se incluye una copia *local* en sus datos sesión . Puede realizar cambios en esta biblioteca, pero deberá volver a publicar su versión pública para poder compartir los cambios.

Cuando se añade una biblioteca pública, puede aparecer el cuadro de diálogo Resolver conflictos si se detectan conflictos entre los términos y los tipos de una biblioteca y las otras bibliotecas locales. Deberá resolver estos conflictos o aceptar las resoluciones propuestas para poder finalizar la operación. Para obtener más información, consulte el tema [“Resolución de conflictos”](#) en la página 186.

Nota: Si siempre actualiza las bibliotecas al lanzar una sesión de área de trabajo interactiva o al publicar cuando cierra una , tiene menos probabilidades de tener bibliotecas no sincronizadas. Para obtener más información, consulte el tema [“Compartimiento de bibliotecas”](#) en la página 184.

Añadir una biblioteca

1. En los menús, seleccione **Recursos > Añadir biblioteca**. Aparecerá el cuadro de diálogo Añadir biblioteca.
2. Seleccione la biblioteca o bibliotecas de la lista.
3. Pulse **Añadir**. Si se producen conflictos entre las bibliotecas recién añadidas y las que ya existían, se le pedirá que verifique las resoluciones de conflictos o que las cambie antes de finalizar la operación. Para obtener más información, consulte el tema [“Resolución de conflictos”](#) en la página 186.

Búsqueda de términos y tipos

Puede buscar en los diversos paneles del editor utilizando la característica Buscar. En el editor, puede elegir **Editar > Buscar** en los menús y aparece la barra de herramientas Buscar. Puede utilizar esta barra de herramientas para buscar una aparición cada vez. Si pulsa **Buscar** de nuevo, puede buscar apariciones subsiguientes del término de búsqueda.

Al realizar la búsqueda, el editor solo busca en la biblioteca o bibliotecas de la lista desplegable de la barra de herramientas Buscar. Si se selecciona **Todas las bibliotecas**, el programa buscará en todas las bibliotecas del editor.

Cuando se inicia una búsqueda, empieza por el área que tiene el foco. La búsqueda continúa a través de cada sección, y vuelve al principio hasta que llega a la casilla activa. Puede invertir el orden de la búsqueda mediante las flechas direccionales. También puede elegir si su búsqueda distingue o no entre mayúsculas y minúsculas.

Para buscar cadenas en la vista

1. En los menús elija **Editar > Buscar**. Aparece la barra de herramientas Buscar.
2. Escriba la cadena que desea buscar.
3. Pulse en el botón **Buscar** para empezar la búsqueda. La siguiente aparición del término o del tipo quedará resaltada.
4. Pulse otra vez en el botón para pasar de una aparición a otra.

Utilización de un asterisco en los términos

Utilizar un asterisco (*) en los términos resulta especialmente útil si se trata de un idioma aglutinante que crea nuevas palabras a partir de otras palabras, uniéndolas sin intercalar espacios entre ellas. Por ejemplo, la palabra alemana *Übernachtungspreis*, está compuesta de: *Übernachtung* + *s* + *Preis*.

Como ejemplo, si busca *preis** en términos del tipo Budget, coincidirá con los conceptos extraídos como *preiserhöhung*. Del mismo modo, **preis* coincidirá con *Übernachtung* y **preis** coincidirá con *Übernachtungspreiserhöhung*.

Visión de bibliotecas

Puede mostrar los contenidos de una biblioteca particular o de todas las bibliotecas. Esto puede resultar útil cuando trabaja con muchas bibliotecas o cuando desea revisar los contenidos de una biblioteca específica antes de publicarla. El cambio de la vista sólo incide en lo que ve en esta pestaña Recursos de la biblioteca, pero no impide el uso de las bibliotecas durante la extracción. Para obtener más información, consulte el tema “Desactivación de bibliotecas locales” en la página 182.

La vista predeterminada es **Todas las bibliotecas**, que muestra todas las bibliotecas del árbol y sus contenidos en otros paneles. Puede cambiar esta selección utilizando la lista desplegable de la barra de herramientas o mediante una selección de menú (**Vista > Bibliotecas**). Cuando se visualiza una única biblioteca, todos los elementos de otras bibliotecas desaparecen de la vista, pero aún se leen durante la extracción.

Para cambiar la vista de Biblioteca

1. En los menús de la pestaña Recursos de biblioteca, seleccione **Vista > Bibliotecas**. Se abre un menú con todas las bibliotecas locales.
2. Seleccione la biblioteca que desea ver o seleccione la opción **Todas las bibliotecas** para ver los contenidos de todas las bibliotecas. Los contenidos de la vista se filtran de acuerdo con su selección.

Administración de las bibliotecas locales

Las bibliotecas locales son las bibliotecas dentro de su sesión de área de trabajo interactiva o dentro de una plantilla, a diferencia de las bibliotecas públicas. Para obtener más información, consulte el tema “Administración de bibliotecas públicas” en la página 183. También las tareas de gestión de biblioteca local básica que es posible que desee realizar, incluyendo: volver a nombrar, inhabilitar, o suprimir una biblioteca local.

Cambio de nombre de las bibliotecas locales

Puede cambiar el nombre de las bibliotecas locales. Si cambia el nombre de una biblioteca local, elimina la asociación que tiene de la versión pública, si es que existe. Esto significa que si realiza cambios posteriores no podrá compartirlos en la versión pública. Puede volver a publicar esta biblioteca local con un nombre nuevo. Esto significa también que no podrá actualizar la versión pública original con cambios que realice en esta versión local.

Nota: No puede cambiar el nombre de una biblioteca pública.

1. En los menús, seleccione **Editar > Propiedades de biblioteca**. Aparecerá el cuadro de diálogo Propiedades de biblioteca.

Para cambiar el nombre de una biblioteca local

1. En la vista de árbol, seleccione la biblioteca a la que desea cambiar el nombre.
2. Introduzca un nombre nuevo para la biblioteca en el cuadro de texto Nombre.
3. Pulse **Aceptar** para aceptar el nombre nuevo de la biblioteca. El cuadro de diálogo se cierra y el nombre de la biblioteca se actualiza en la vista de árbol.

Desactivación de bibliotecas locales

Si desea excluir temporalmente una biblioteca del proceso de extracción, puede anular la selección del cuadro de verificación a la izquierda del nombre de la biblioteca de la vista de árbol. Esto indica que desea conservar la biblioteca pero que su contenido se pase por alto cuando se realice la comprobación de conflictos y durante el proceso de extracción.

Para desactivar una biblioteca

1. En el panel del árbol de bibliotecas, seleccione la biblioteca que desea desactivar.
2. Pulse la barra espaciadora. Se borra la marca de la casilla de verificación a la izquierda del nombre.

Eliminación de bibliotecas locales

Puede suprimir una biblioteca sin eliminar la versión pública de la misma y viceversa. Al suprimir una biblioteca local se suprimirá la biblioteca y todo su contenido sólo de la sesión. Suprimir una versión local de una biblioteca no elimina esa biblioteca de otras sesiones o la versión pública. Para obtener más información, consulte el tema [“Administración de bibliotecas públicas”](#) en la página 183.

Para eliminar una biblioteca local

1. En la vista de árbol, seleccione la biblioteca que desea eliminar.
2. En los menús, seleccione **Editar > Suprimir** para suprimir la biblioteca. La biblioteca se elimina.
3. Si nunca había publicado la biblioteca, se abre un mensaje que le pregunta si desea eliminar la biblioteca o conservarla. Pulse **Eliminar** para continuar o **Conservar** si desea conservarla.

Nota: Siempre debe quedar una biblioteca.

Administración de bibliotecas públicas

Para reutilizar las bibliotecas locales, puede publicarlas y, a continuación, trabajar con ellas y verlas a través del cuadro de diálogo Gestionar bibliotecas (**Recursos > Gestionar bibliotecas**). Para obtener más información, consulte el tema [“Compartimiento de bibliotecas”](#) en la página 184. Algunas tareas de gestión de biblioteca pública básica que es posible que desee realizar incluyen la importación exportación o supresión de una biblioteca pública. No puede cambiar el nombre de una biblioteca pública.

Importación de bibliotecas públicas

1. En el cuadro de diálogo Administrar bibliotecas, pulse en **Importar....** Aparecerá el cuadro de diálogo Importar biblioteca.
2. Seleccione el archivo de biblioteca (*.lib) que desea importar, y si también desea añadirla localmente seleccione **Añadir biblioteca al proyecto actual**.
3. Pulse **Importar**. El cuadro de diálogo se cierra. Si ya existe una biblioteca pública con el mismo nombre, se le pedirá que cambie el nombre de la biblioteca que está importando o que sobrescriba la biblioteca pública actual.

Exportación de bibliotecas públicas

Puede exportar bibliotecas públicas en formato .lib para poder compartirlas.

1. En el cuadro de diálogo Administrar bibliotecas, seleccione la biblioteca que desea exportar a la lista.
2. Pulse **Exportar**. Aparecerá el cuadro de diálogo Seleccionar directorio.
3. Seleccione el directorio al que desea exportar y pulse en **Exportar**. El cuadro de diálogo se cierra y el archivo de biblioteca (*.lib) se exporta.

Eliminación de bibliotecas públicas

Puede suprimir una biblioteca local sin eliminar la versión pública de la misma y viceversa. Sin embargo, si la biblioteca se suprime de este cuadro de diálogo, ya no se puede añadir a cualquier recursos de sesión hasta que una versión local se publique de nuevo.

Si elimina una biblioteca que estaba instalada en el producto, se restaurará la versión originalmente instalada.

1. En el cuadro de diálogo Administrar bibliotecas, seleccione la biblioteca que desea eliminar. Puede ordenar la lista pulsando en la cabecera apropiada.
2. Pulse **Suprimir** para suprimir la biblioteca. IBM SPSS Modeler Text Analytics verifica si la versión local de la biblioteca es la misma que en la biblioteca pública. En caso afirmativo, la biblioteca se elimina sin emitir ningún mensaje de alerta. Si las versiones de las bibliotecas difieren, se abre un mensaje de alerta para preguntarle si desea conservar o eliminar la versión pública.

Compartimiento de bibliotecas

Las bibliotecas le permiten trabajar con los recursos de manera que sea fácil de compartir entre varias sesiones de áreas de trabajo interactivas. Las bibliotecas pueden existir en dos estados o versiones posibles. Las bibliotecas que son editables en el editor y parte de una sesión de área de trabajo interactiva se denominan **Bibliotecas locales**. Mientras trabaja con una sesión de área de trabajo interactiva, puede realizar una gran cantidad de cambios, por ejemplo en la biblioteca *Vegetables*. Si los cambios realizados fueron útiles en otros datos, puede hacer que estos recursos estén disponibles creando una versión de **biblioteca pública** de la biblioteca *Vegetales*. Una biblioteca pública, como su nombre indica, está disponible para otros recursos en cualquier proyecto de sesión de área de trabajo interactiva.

Las bibliotecas públicas pueden verse en el cuadro de diálogo Administrar bibliotecas. Cuando haya creado esta versión de biblioteca pública, podrá añadirla a los recursos de otros contextos para poder compartir estos recursos lingüísticos personalizados.

Las bibliotecas enviadas inicialmente son bibliotecas públicas. Se pueden editar los recursos de estas bibliotecas y luego crear una versión pública nueva. Esas versiones nuevas serán accesibles en otras sesiones de área de trabajo interactiva.

A medida que trabaje con las bibliotecas y realice cambios, las versiones de la biblioteca quedarán desincronizadas. En algunos casos, una versión local puede ser más reciente que una versión pública, y en otros casos, la versión pública puede ser más reciente que la versión local. También es posible que tanto las versiones locales como públicas contengan cambios que otras no tengan, si la versión pública se actualizó desde otra sesión de área de trabajo interactiva. Si las versiones de las bibliotecas se desincronizan, puede sincronizarlas de nuevo. La sincronización de las versiones de las bibliotecas consiste en volver a publicar y/o actualizar las bibliotecas locales.

Cada vez que inicie o cierre una sesión de área de trabajo interactiva, se le solicitará sincronizar cualquier biblioteca que necesite actualización o que se necesite publicar de nuevo. También puede identificar fácilmente el estado de sincronización de la biblioteca local mediante el icono que aparece junto al nombre de la biblioteca en la vista de árbol, o bien en el cuadro de diálogo Propiedades de biblioteca. También puede optar por hacerlo en cualquier momento mediante las selecciones de menú apropiadas. En la tabla siguiente se describen los cinco estados posibles y sus iconos asociados.

Tabla 39. Estados de sincronización de la biblioteca local	
Icono	Descripción del estado de la biblioteca local
	Sin publicar — la biblioteca local nunca se ha publicado.
	Sincronizada — Las versiones de la biblioteca pública y local son idénticas. Esto también se aplica a la <i>Biblioteca local</i> , que no se puede publicar porque está diseñada para contener sólo sesión -recursos específicos.
	Fuera de la fecha — La versión de la biblioteca pública es más reciente que la versión local. Puede actualizar la versión local con los cambios.
	Más nueva — La versión de la biblioteca local es más reciente que la versión pública. Puede volver a publicar la versión local a la versión pública.
	Fuera de sincronización — Tanto las bibliotecas locales como públicas contienen cambios que otras no tienen. Debe decidir si actualizar o publicar la biblioteca local. Si decide actualizar, perderá los cambios realizados hasta la última vez que la actualizó o publicó. Si decide publicar, se sobrescribirán los cambios de la versión pública.

Nota: Si siempre actualiza las bibliotecas cuando inicia una sesión de área de trabajo interactiva o publica cuando cierra una sesión, tiene menos probabilidades de tener bibliotecas que están fuera de sincronización.

Puede volver a publicar una biblioteca en cualquier momento que piense que los cambios en la biblioteca beneficiarán otras rutas que también pueden contener esta biblioteca. Entonces, si los cambios

beneficiarán otras rutas, puede actualizar las versiones locales en esas rutas. De esta manera, puede crear rutas para cada contexto o dominio que se aplique a sus datos, mediante la creación de bibliotecas nuevas y/o añadiendo cualquier número de bibliotecas públicas a sus recursos.

Si se comparte una versión pública de una biblioteca, habrá mayores probabilidades de que las diferencias entre las versiones local y pública aumenten. Cada vez que inicie o cierre y publique desde una sesión de área de trabajo interactiva o abra o cierre una plantilla desde el Editor de plantillas, se visualiza un mensaje para permitirle publicar y/o actualizar las bibliotecas cuyas versiones no están en sincronización con aquellas en el cuadro de diálogo Gestionar Bibliotecas. Si la versión de la biblioteca pública es más reciente que la versión local, un cuadro de diálogo le preguntará si desea actualizar las que están abiertas. Puede optar por conservar la versión local tal como está en lugar de actualizarla con la versión pública, o bien fusionar las actualizaciones en la biblioteca local.

Publicación de bibliotecas

Si nunca ha publicado una biblioteca determinada, la publicación implica crear una copia pública de la biblioteca local en la base de datos. Si vuelve a publicar una biblioteca, los contenidos de la biblioteca local reemplazarán los contenidos de la versión pública existente. Después de volver a publicar, puede actualizar esta biblioteca en otros proyectos de sesiones de ruta para que sus versiones locales estén en sincronización con la versión pública. Aunque puede publicar una biblioteca, siempre hay una versión local almacenada en sesión.

Importante: Si realiza cambios en la biblioteca local y, mientras tanto, la versión pública de la biblioteca también se cambia, se considerará que su biblioteca está desincronizada. Se recomienda que empiece actualizando la versión local con los cambios públicos, realizar los cambios que desee y después publicar la versión local otra vez para que las dos versiones sean idénticas. Si primero realiza los cambios y publica, se sobrescribirán los cambios de la versión pública.

Para publicar bibliotecas locales en la base de datos

1. En los menús, seleccione **Recursos > Publicar bibliotecas**. Se abre el cuadro de diálogo Publicar bibliotecas con todas las bibliotecas que necesitan publicarse seleccionadas de forma predeterminada.
2. Marque la casilla de verificación de la izquierda de cada biblioteca que desee publicar o volver a publicar.
3. Pulse **Publicar** para publicar las bibliotecas en la base de datos Administrar bibliotecas.

Actualización de bibliotecas

Siempre que inicie o cierre una sesión de área de trabajo interactiva, puede actualizar o publicar las bibliotecas que ya no están en sincronía con la versión pública. Si la versión de la biblioteca pública es más reciente que la versión local, un cuadro de diálogo le preguntará si desea actualizar las que están abiertas. Puede optar entre conservar la versión local en lugar de actualizarla con la versión pública, y reemplazar la versión local con la de la pública. Si la versión pública de una biblioteca es más reciente que la versión local, puede actualizar la versión local para sincronizar su contenido con el de la versión pública. Actualizar significa incorporar en la versión local los cambios encontrados en la versión pública.

Nota: Si siempre actualiza las bibliotecas cuando inicia una sesión de área de trabajo interactiva o si publica al cerrar un, tiene menos posibilidades de tener bibliotecas que están fuera de sincronización. Para obtener más información, consulte el tema [“Compartimiento de bibliotecas”](#) en la página 184.

Para actualizar bibliotecas locales

1. En los menús, seleccione **Recursos > Actualizar bibliotecas**. Se abre el cuadro de diálogo Actualizar bibliotecas con todas las bibliotecas que necesitan actualizarse seleccionadas de forma predeterminada.
2. Marque la casilla de verificación de la izquierda de cada biblioteca que desee publicar o volver a publicar.
3. Pulse **Actualizar** para actualizar las bibliotecas locales.

Resolución de conflictos

Conflictos de la biblioteca local frente a la pública

Cada vez que inicie una sesión de ruta , IBM SPSS Modeler Text Analytics realiza una comparación de las bibliotecas locales y de aquellas listadas en el cuadro de diálogo Gestionar bibliotecas. Si las bibliotecas locales en su sesión no están en sincronía con las versiones publicadas, se abre el cuadro de diálogo con el aviso de advertencia de Sincronización de la biblioteca. Puede elegir entre las opciones siguientes para seleccionar las versiones de biblioteca que desea utilizar aquí:

- **Todas las bibliotecas locales en archivo.** Esta opción conserva todas las bibliotecas locales tal como están. Siempre puede volver a publicarlas o actualizarlas en otro momento.
- **Todas las bibliotecas publicadas en esta máquina.** Esta opción sustituye las bibliotecas locales que se muestran por las versiones que se encuentran en la base de datos.
- **Todas las bibliotecas más recientes.** Esta opción sustituye las bibliotecas locales más antiguas por las versiones públicas más recientes de la base de datos.
- **Otros.** Esta opción permite seleccionar manualmente las versiones que desee eligiéndolas a partir de la tabla.

Conflictos en los términos forzados

Cuando añade una biblioteca pública o actualiza una biblioteca local, pueden quedar al descubierto conflictos y entradas duplicadas entre los términos y los tipos de esta biblioteca y los términos y los tipos de las otras bibliotecas de sus recursos. Si esto ocurre, se le pedirá que verifique las resoluciones de conflictos que se proponen o que las cambie antes de finalizar la operación en el cuadro de diálogo Editar términos forzados. Para obtener más información, consulte el tema [“Forzar términos” en la página 192](#).

El cuadro de diálogo Editar términos forzados contiene las parejas de términos o tipos conflictivos. Se utilizan colores de fondo alternos para distinguir visualmente cada pareja conflictiva. Estos colores pueden cambiarse en el cuadro de diálogo Opciones. Para obtener más información, consulte el tema [“Opciones: Visualizar pestaña” en la página 75](#). El cuadro de diálogo Editar términos forzados contiene dos pestañas:

- **Duplicados.** Esta pestaña contiene los términos duplicados que se encuentran en las bibliotecas. Si aparece un icono de chincheta detrás de cada término, significa que esta aparición del término está forzada. Si aparece un icono de X de color negro, significa que esta aparición del término se pasará por alto durante la extracción porque está forzado en algún otro sitio.
- **Usuario definido.** Esta pestaña contiene una lista de los términos que se han forzado manualmente en este panel de términos del diccionario de tipo y no a través de los conflictos.

Nota: El cuadro de diálogo Editar términos forzados se abre después de añadir o actualizar una biblioteca. Si cancela este cuadro de diálogo, no significa que cancele la actualización o la adición de la biblioteca.

Para resolver conflictos

1. En el cuadro de diálogo Editar términos forzados, seleccione el botón de radio de la columna Usar correspondiente al término que desea forzar.
2. Cuando haya terminado, pulse en **Aceptar** para aplicar los términos forzados y cerrar el cuadro de diálogo. Si pulsa en **Cancelar**, se cancelarán los cambios realizados en el cuadro de diálogo.

Capítulo 16. Acerca de los diccionarios de biblioteca

Los recursos utilizados para extraer datos de texto se guardan en forma de plantillas y bibliotecas. Cada biblioteca se compone de tres diccionarios.

- El **diccionario de tipo** contiene una colección de palabras que se agrupan bajo una misma etiqueta, tipo o nombre. Cuando el motor de extracción lee los datos de texto, compara las palabras que encuentra en el texto con los términos definidos en los diccionarios de tipo. Durante la extracción, las formas declinadas de los sinónimos y los términos de tipo se agrupan bajo un término de destino llamado concepto. Los conceptos extraídos se asignan al diccionario de tipo donde aparecen como términos. Puede administrar los diccionarios de tipo en los paneles central y superior izquierdos del editor: el árbol de biblioteca y el panel de términos. Para obtener más información, consulte el tema [“Diccionarios de tipo” en la página 187](#).
- El **diccionario de sustitución** contiene una recopilación de términos definidos como sinónimos o como elementos opcionales que se utilizan para agrupar términos similares bajo un término de destino, llamado concepto en los resultados de extracción finales. Puede administrar los diccionarios de sustitución en el panel inferior izquierdo del editor utilizando la pestaña Sinónimos y la pestaña Opcional. Para obtener más información, consulte el tema [“Diccionarios de sustitución/sinónimos” en la página 194](#).
- El **diccionario de exclusión** contiene una recopilación de términos y tipos que se eliminarán de los resultados de la extracción final. Puede administrar los diccionarios de exclusión en el panel del extremo derecho del editor. Para obtener más información, consulte el tema [“Diccionarios de exclusión” en la página 197](#).

Para obtener más información, consulte el tema [Capítulo 15, “Trabajo con bibliotecas”, en la página 179](#).

Diccionarios de tipo

Un *diccionario de tipo* está compuesto por un nombre o etiqueta de tipo, y una lista de términos. Los tipos de diccionarios se gestionan en los paneles izquierdo y central en la parte superior del separador Recursos de biblioteca en el editor. Puede acceder a esta vista mediante **Vista > Editor de recursos** en los menús, si está en una sesión de entorno de trabajo interactiva. De lo contrario, puede editar los diccionarios de una plantilla específica en el Editor de plantillas.

Cuando el motor de extracción lee los datos de texto, compara las palabras que encuentra en el texto con los términos definidos en los diccionarios de tipo. Los términos son palabras o frases de los diccionarios de tipo de los recursos lingüísticos.

Cuando una palabra coincide con un término, se le asigna al nombre de tipo correspondiente a dicho término. Cuando los recursos se leen durante la extracción, los términos que se encontraron en el texto pasan una serie de procedimientos antes de que se conviertan en conceptos en el panel Resultados extraídos. Si el motor de extracción considera que los diferentes términos que pertenecen a un mismo diccionario son sinónimos, se agruparán bajo el término que aparezca con más frecuencia y se denominará como un *concepto* en el panel Resultados extraídos. Por ejemplo, si los términos *question* y *query* pueden aparecer bajo el nombre de concepto *question* al final.

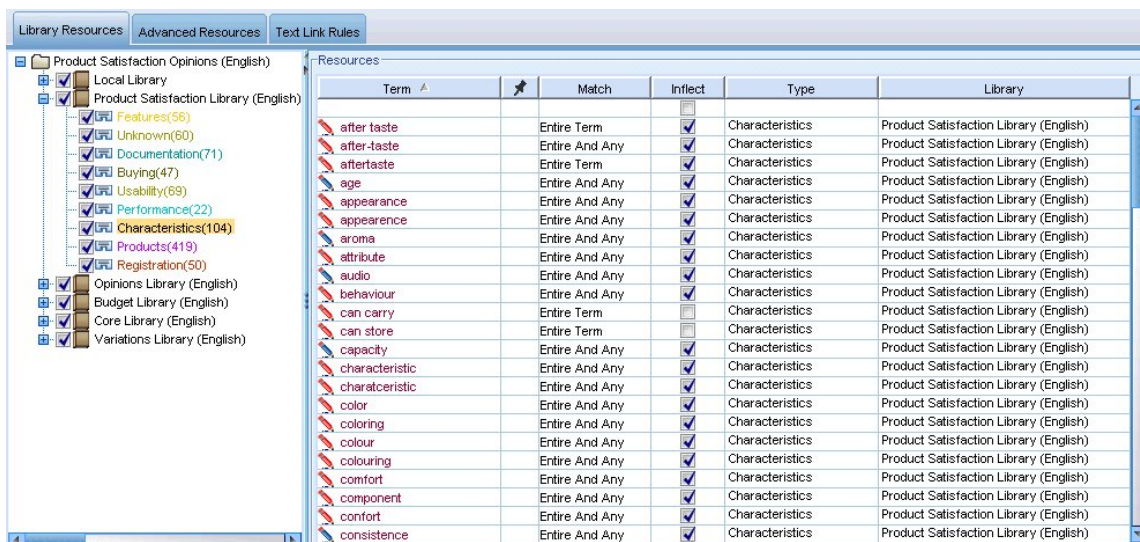


Figura 42. Árbol de bibliotecas y panel de términos

La lista de diccionarios de tipo se muestra en el panel del árbol de bibliotecas a la izquierda. El contenido de cada diccionario de tipo aparece en el panel central. Los diccionarios de tipo constituyen algo más que una simple lista de términos. La manera en que las palabras y las frases de los datos de texto se hacen coincidir con los términos definidos en los diccionarios de tipo está determinada por la opción de coincidencia que se haya definido. Una **opción de coincidencia** específica de qué manera el tema se ancla a una palabra o frase candidata en los datos de texto. Consulte el tema “Añadir términos” en la página 190 para obtener más información.

También puede ampliar los términos del diccionario de tipo especificando si desea que se generen y se añadan automáticamente al diccionario formas declinadas de los términos. Al generar formas declinadas, automáticamente se añaden los plurales de los términos en singular, y los singulares de los términos en plural, y los adjetivos en el diccionario de tipo. Para obtener más información, consulte el tema “Añadir términos” en la página 190.

Nota: Para la mayoría de idiomas, los conceptos que no se encuentran en ningún diccionario de tipo, pero que se extraen del texto, se escriben automáticamente como <Unknown>

Utilización de un asterisco en los términos

Utilizar un asterisco (*) en los términos resulta especialmente útil si se trata de un idioma aglutinante que crea nuevas palabras a partir de otras palabras, uniéndolas sin intercalar espacios entre ellas. Por ejemplo, la palabra alemana *Übernachtungspreis*, está compuesta de: *Übernachtung* + *s* + *Preis*.

Como ejemplo, si busca *preis** en términos del tipo Budget, coincidirá con los conceptos extraídos como *preiserhöhung*. Del mismo modo, **preis* coincidirá con *Übernachtung* y **preis** coincidirá con *Übernachtungspreiserhöhung*.

Tipos incorporados

IBM SPSS Modeler Text Analytics se entrega con un conjunto de recursos lingüísticos en forma de bibliotecas entregadas y recursos compilados. Las bibliotecas incluidas contienen un conjunto de diccionarios de tipo integrados que incluyen <Location>, <Organization>, <Person> y <Product>.

El motor de extracción utiliza estos diccionarios de tipo para asignar tipos a los conceptos que extrae como, por ejemplo, asignar el tipo <Location> al concepto *parís*. Aunque en los diccionarios de tipo incorporados hay un gran número de términos ya definidos, no abarcan todas las posibilidades. Por lo tanto, puede añadir términos en los diccionarios o crear los suyos propios. Para obtener una descripción de los contenidos de un determinado diccionario de tipo enviado, lea la anotación del cuadro de diálogo Propiedades de tipo. Seleccione el tipo en el árbol y elija **Editar > Propiedades** en el menú contextual.

Nota:

Además de las bibliotecas suministradas, los recursos compilados (también utilizados por el motor de extracción) contienen un gran número de definiciones complementarias a los diccionarios de tipo incorporados, pero su contenido no es visible en el producto. Sin embargo, puede forzar que un término especificado en los diccionarios compilados se coloque en otro diccionario. Si desea más información, consulte [“Forzar términos” en la página 192](#).

Crear tipos

Puede crear diccionarios de tipo para facilitar la agrupación de los términos similares. Cuando durante el proceso de extracción se descubren los términos que aparecen en este diccionario, se asignarán a este nombre de tipo y se extraerán bajo un nombre de concepto. Siempre que se crea una biblioteca, se incluye una biblioteca de tipos vacía para que pueda empezar a introducir términos inmediatamente.

Si está analizando un texto sobre comida y desea agrupar términos relacionados con verduras, puede crear su propio diccionario tipo <Vegetables>. A continuación, puede añadir términos como carrot, broccolli y spinach si considera que son términos importantes que aparecerán en el texto. A continuación, durante la extracción, si se encuentra alguno de estos términos, se extraen como conceptos y se asignan al tipo <Vegetables>.

No es necesario que defina cada forma de una palabra o expresión, porque puede optar por generar las formas declinadas de los términos. Al elegir esta opción, el motor de extracción reconoce automáticamente las formas singular y plural de los términos entre el resto de las formas como pertenecientes a este tipo. Esta opción resulta particularmente útil cuando el tipo contiene mayoritariamente sustantivos, puesto que es improbable que desee incluir las formas declinadas de verbos o adjetivos.

El cuadro de diálogo Propiedades tipo contiene los campos siguientes.

Nombre. El nombre que asigna al diccionario de tipo que está creando. Se recomienda que no utilice espacios en blanco en los nombres de tipo, sobre todo si hay dos o más nombres de tipo que empiecen con la misma palabra.

Nota: Existen algunas restricciones sobre los nombres de tipo y el uso de símbolos. Por ejemplo, no utilice símbolos como "@" o "!" en el nombre.

Coincidencia predeterminada. El atributo de coincidencia predeterminada indica al motor de extracción cómo debe hacer coincidir este término con los datos de texto. Siempre que se añade un término a este diccionario de tipo, este es el atributo de coincidencia que se asigna automáticamente. Siempre puede cambiar la opción de coincidencia manualmente en la lista de términos. Las opciones incluyen: **Término completo, Inicio, Fin, Cualquier, Inicio o Fin, Completo e Inicio, Completo y Fin, Completo y (Inicio o Fin), y Completo (sin compuestos)**. Consulte el tema [“Añadir términos” en la página 190](#) para obtener más información.

Añadir a. Este campo indica la biblioteca donde creará el nuevo diccionario de tipo.

Genera formas declinadas de forma predeterminada. Esta opción indica al motor de extracción que utilice la morfología gramatical para capturar y agrupar formas similares de los términos que añada a este diccionario, como las formas singular y plural del término. Esta opción resulta particularmente útil cuando el tipo contiene mayormente sustantivos. Cuando selecciona esta opción, todos los términos nuevos que se añaden a este tipo adoptarán automáticamente esta opción, aunque puede cambiarlo manualmente en la lista.

Color de fuente. Este campo permite distinguir los resultados de este tipo del resto de resultados de la interfaz. Si selecciona **Usar color principal**, también se utilizará el color de tipo predeterminado para este diccionario de tipo. Este color predeterminado se establece en el cuadro de diálogo Opciones. Para obtener más información, consulte el tema [“Opciones: Visualizar pestaña” en la página 75](#). Si selecciona **Personalizado**, seleccione un color en la lista desplegable.

Anotación. Este campo es opcional y puede utilizarse para introducir comentarios o descripciones.

Para crear un diccionario de tipo

1. Seleccione la biblioteca donde desea crear un diccionario de tipo nuevo.

2. En los menús, seleccione **Herramientas > Nuevo tipo**. Aparecerá el cuadro de diálogo Propiedades de tipo.
3. Escriba el nombre del diccionario de tipo en el cuadro de texto **Nombre** y elija las opciones que desee.
4. Pulse **Aceptar** para crear el diccionario de tipo. El nuevo tipo será visible en el panel del árbol de bibliotecas y aparecerá en el panel central. Puede empezar a añadir términos inmediatamente. Para obtener más información, consulte [“Añadir términos” en la página 190](#).

Nota: Estas instrucciones le muestran como realizar cambios en la Editor de recursos vista o Editor de plantillas . Tenga en cuenta que también puede realizar este tipo de ajuste directamente desde el panel Resultados de extracción , panel Datos, panel Categorías o el recuadro de diálogo Definiciones en las demás vistas.. Para obtener más información, consulte el tema [“Refinar resultados de extracción” en la página 89](#).

Añadir términos

El panel del árbol de bibliotecas muestra las bibliotecas y puede expandirse para mostrar también los diccionarios de tipo que contienen. En el panel central aparece una lista de los términos de la biblioteca o del diccionario de tipo seleccionados, según cuál sea la selección en el árbol.

En el Editor de recursos, puede añadir términos a un diccionario de tipo directamente en el panel de términos o a través del cuadro de diálogo Añadir nuevos términos. Los términos que añada pueden ser palabras simples o compuestas. En la parte superior de la lista siempre encontrará una fila en blanco para que pueda añadir un término nuevo.

Nota: Estas instrucciones le muestran como realizar cambios en la Editor de recursos vista o Editor de plantillas . Tenga en cuenta que también puede realizar este tipo de ajuste directamente desde el panel Resultados de extracción , panel Datos, panel Categorías o el recuadro de diálogo Definiciones en las demás vistas.. Para obtener más información, consulte el tema [“Refinar resultados de extracción” en la página 89](#).

Columna Término

En esta columna, introduzca palabras simples o compuestas en la casilla. El color en el que aparece el término depende del color del tipo donde se ha almacenado o forzado el término. Puede cambiar los colores de tipo en el cuadro de diálogo Propiedades de tipo. Para obtener más información, consulte el tema [“Crear tipos” en la página 189](#).

Columna Forzar

En esta columna, al colocar un icono de chincheta en esta casilla, está indicando al motor de extracción que pase por alto el resto de las apariciones de este mismo término en otras bibliotecas. Para obtener más información, consulte el tema [“Forzar términos” en la página 192](#).

Columna Coincidencia

En esta columna, seleccione una opción de coincidencia para indicar al motor de extracción cómo debe hacer coincidir este término con los datos de texto. Consulte la tabla para ver los ejemplos pertinentes. Puede cambiar el valor predeterminado editando las propiedades de tipo. Para obtener más información, consulte el tema [“Crear tipos” en la página 189](#). En los menús, seleccione **Editar > Cambiar coincidencia**. A continuación figuran las opciones básicas de coincidencia, puesto que también pueden combinarse:

- **Inicio.** Si el término en el diccionario coincide con la primera palabra en un concepto extraído del texto, se asignará este tipo. Por ejemplo, si especifica *apple*, *apple tart* se emparejará.
- **Final.** Si el término en el diccionario coincide con la última palabra en un concepto extraído del texto, se asigna este tipo. Por ejemplo, si especifica *apple*, *cider apple* se emparejará.

- **Cualquiera.** Si el término en el diccionario coincide con cualquier palabra de un concepto extraído del texto, se asignará este tipo. Por ejemplo, si especifica *apple*, la opción **Cualquiera** será *apple tart*, *cider apple* y *cider apple tart* de la misma manera.
- **Término completo.** Si el concepto completo extraído del texto coincide con el término exacto del diccionario, se asignará este tipo. Si se añade un término como **Término completo, Completo e inicio, Completo y fin, Completo y todo o Completo (sin compuestos)** se forzará la extracción de un término.

Además, puesto que el tipo <Person> solo extrae nombres con dos partes, como *edith piafo* o *mohandas gandhi*, puede añadir explícitamente los nombres propios en este diccionario de tipo si intenta extraer un nombre propio sin que se mencione ningún apellido. Por ejemplo, si desea captar todas las instancias de *edith* como nombre propio, debe añadir *edith* al tipo <Person> utilizando **Término completo o Completo e inicio**.

- **Completo (sin compuestos).** Si el concepto completo extraído del texto coincide con el término exacto en el diccionario, se asignará este tipo y la extracción se detendrá para impedir que la extracción haga coincidir el término con un compuesto más largo. Por ejemplo, si especifica *apple*, la opción **Entero (no compuesto)** tecleará *apple* y no extraerá el compuesto *apple sauce*, a menos que se vea forzado en otro lugar.

En la tabla siguiente, suponga que el término *apple* está en un diccionario de tipo. En función de la opción de coincidencia, en esta tabla se muestran los conceptos que se extraerían y tipificarían si se encontraran en el texto.

Tabla 40. Ejemplos de coincidencias				
Opciones de coincidencia para el término:  apple	Conceptos extraídos			
	apple	Tarta apple	maduro apple	casera Tarta apple
Término completo	✓			
Iniciar		✓		
END			✓	
Inicio a fin		✓	✓	
Completo e inicio	✓	✓		
Completo y fin	✓		✓	
Completo y (Inicio o Fin)	✓	✓	✓	
Cualquiera		✓	✓	✓
Completo y Cualquiera	✓	✓	✓	✓
Completo (sin compuestos)	✓	<i>nunca extraído</i>	<i>nunca extraído</i>	<i>nunca extraído</i>

Columna Flexionar

En esta columna, seleccione si el motor de extracción debe generar formas declinadas de este término durante la extracción, de manera que se agrupen juntas. El valor predeterminado de esta columna está

definido en Propiedades de tipo, pero puede cambiar esta opción en cada caso individual directamente en la columna. En los menús, seleccione **Editar > Cambiar conjugación**.

Columna Tipo

En esta columna, seleccione un diccionario de tipo de la lista desplegable. La lista de tipos se filtra según la opción que haya seleccionado en el panel del árbol de bibliotecas. El primer tipo de la lista siempre es el tipo predeterminado seleccionado en el panel del árbol de bibliotecas. En los menús, seleccione **Editar > Cambiar tipo**.

Columna Biblioteca

En esta columna, se muestra la biblioteca en la que está almacenado el término. Puede arrastrar y soltar un término en otro tipo en el panel del árbol de bibliotecas para cambiarlo de biblioteca.

Añadir un único término al diccionario de tipo

1. En el panel del árbol de bibliotecas, seleccione el diccionario de tipo al que desea añadir el término.
2. En la lista de términos del panel central, escriba el término en la primera casilla disponible y defina las opciones que desee para dicho término.

Añadir varios términos a un diccionario de tipo

1. En el panel del árbol de bibliotecas, seleccione el diccionario de tipo al que desea añadir los términos.
2. En los menús, seleccione **Herramientas > Nuevos términos**. Aparecerá el cuadro de diálogo Añadir nuevos términos.
3. Especifique los términos que desee añadir al diccionario de tipo seleccionado escribiéndolos o copiando y pegando el conjunto de términos. Si especifica varios términos, deberá separarlos mediante el delimitador definido en el cuadro de diálogo Opciones, o añadir cada término en una línea nueva. Consulte el tema [“Opciones de configuración”](#) en la [página 74](#) para obtener más información.
4. Pulse **Aceptar** para añadir los términos al diccionario. La opción de coincidencia se establece automáticamente en la opción predeterminada para esta biblioteca de tipo. El cuadro de diálogo se cierra y los nuevos términos aparecen en el diccionario.

Forzar términos

Si desea que un término se asigne a un tipo determinado, puede añadirlo al diccionario de tipo correspondiente. Sin embargo, si hay varios términos con el mismo nombre, el motor de extracción debe conocer qué tipo se va a utilizar. Por lo tanto, se le indicará que seleccione el tipo que desea utilizar. A esta acción se le llama *forzar* un término dentro de un tipo. Esta opción es más útil al alterar temporalmente la asignación de tipo de un diccionario compilado (interno, no editable). En general, se recomienda evitar los términos duplicados.

La acción de forzar no *eliminará* el resto de las apariciones de este término, sino que el motor de extracción las pasará por alto. Más adelante podrá cambiar qué aparición debe utilizarse mediante el forzado o la anulación del forzado de un término. Es posible que también deba forzar un término dentro de un diccionario de tipo cuando añada o actualice una biblioteca pública.

Puede ver qué términos están forzados o cuáles se pasan por alto en la columna de Forzar, la segunda columna del panel de términos. Si aparece un icono de chincheta, significa que esta aparición del término está forzada. Si aparece un icono de X de color negro, significa que esta aparición del término se pasará por alto durante la extracción porque está forzado en algún otro sitio. Además, cuando fuerza un término, éste aparece del color del tipo dentro del que se ha forzado. Esto significa que si ha forzado un término que está en Type 1 y Type 2 en Type 1, en cualquier momento que vea este término en la ventana, aparecerá en el color de font definido para Type 1.

Para cambiar el estado pulse dos veces con el ratón en el icono. Si el término aparece en algún otro lugar, se abre el cuadro de diálogo Resolver conflictos para que pueda seleccionar cuál de las apariciones debe utilizarse.

Renombrar tipos

Puede cambiar el nombre de un diccionario de tipo o cambiar otras opciones de diccionario editando las propiedades de tipo.

Importante: Se recomienda que no utilice espacios en blanco en los nombres de tipo, sobre todo si hay dos o más nombres de tipo que empiecen con la misma palabra. También se recomienda no cambiar el nombre de los tipos en las bibliotecas Core u Opinions, ni cambiar sus atributos de coincidencia predeterminada.

Renombrar un tipo

1. En el panel del árbol de bibliotecas, seleccione el diccionario de tipo al que desea cambiar el nombre.
2. Pulse el botón derecho del ratón y elija **Propiedades de tipo** en el menú contextual. Aparecerá el cuadro de diálogo Propiedades de tipo.
3. Escriba el nombre nuevo del diccionario de tipo en el cuadro de texto Nombre.
4. Pulse **Aceptar** para aceptar el nombre nuevo. El nombre de tipo nuevo aparece visible en el panel del árbol de bibliotecas.

Cómo mover tipos

Puede arrastrar un diccionario de tipo y soltarlo en otra ubicación de una biblioteca o en otra biblioteca del árbol.

Para reordenar un tipo en una biblioteca

1. En el panel del árbol de bibliotecas, seleccione el diccionario de tipo que desea mover.
2. En los menús, seleccione **Editar > Subir** para subir el diccionario de tipo una posición en el panel de árbol de biblioteca o **Editar > Bajar** para bajarlo una posición.

Mover un tipo a otra biblioteca

1. En el panel del árbol de bibliotecas, seleccione el diccionario de tipo que desea mover.
2. Pulse el botón derecho del ratón y elija **Propiedades de tipo** en el menú contextual. Aparecerá el cuadro de diálogo Propiedades de tipo. (También puede arrastrar y soltar el tipo en otra biblioteca).
3. En el cuadro de lista Añadir a, seleccione la biblioteca a la que desea mover el diccionario de tipo.
4. Pulse en **Aceptar**. El cuadro de diálogo se cierra, y el tipo se encuentra ahora en la biblioteca que ha seleccionado.

Inhabilitar y suprimir tipos

Si desea eliminar temporalmente un diccionario de tipo, puede desactivarlo quitándole la marca de verificación de la casilla de la izquierda del nombre del diccionario en el panel del árbol de bibliotecas. Esto indica que desea conservar el diccionario en la biblioteca, pero que desea que se pasen por alto sus contenidos durante una comprobación de conflictos y durante el proceso de extracción.

También puede eliminar permanentemente los diccionarios de tipo de una biblioteca.

Para inhabilitar un diccionario de tipo

1. En el panel del árbol de bibliotecas, seleccione el diccionario de tipo que desea desactivar.

2. Pulse la barra espaciadora. Se borra la marca de la casilla de verificación a la izquierda del nombre del tipo.

Para suprimir un diccionario de tipo

1. En el panel del árbol de bibliotecas, seleccione el diccionario de tipo que desea eliminar.
2. En los menús, seleccione **Editar > Suprimir** para suprimir el diccionario de tipos.

Diccionarios de sustitución/sinónimos

Un *diccionario de sustitución* es una recopilación de términos que ayuda a agrupar términos similares bajo un término de destino. Los diccionarios de sustitución se administran en el panel inferior de la pestaña Recursos de la biblioteca. Puede acceder a esta vista mediante **Vista > Editor de recursos** en los menús, si está en una sesión de entorno de trabajo interactiva. De lo contrario, puede editar los diccionarios de una plantilla específica en el Editor de plantillas.

Puede definir dos formas de sustituciones en este diccionario: *sinónimos* y *elementos opcionales*. Puede pulsar en las pestañas de este panel para conmutarlas.

Después de ejecutar una extracción en los datos de texto, puede encontrar varios conceptos que son sinónimos o formas declinadas de otros conceptos. Mediante la identificación de elementos opcionales y sinónimos, puede forzar que el motor de extracción asigne esos términos a un único término de destino.

La sustitución del uso de sinónimos y elementos opcionales reduce el número de conceptos en el panel Resultados de extracción combinándolos en conceptos más significativos y representativos con cuentas Doc. con mayor frecuencia

Sinónimos

Los sinónimos asocian dos o más palabras con el mismo significado. Los sinónimos también pueden utilizarse para agrupar términos con sus abreviaturas, o para agrupar palabras que suelen escribirse mal con la ortografía correcta. Puede definir estos sinónimos en la pestaña Sinónimos.

Una definición de sinónimo se compone de dos partes. La primera es un término **Objetivo**, que es el término bajo el cual desea que el motor de extracción agrupe todos los términos sinónimos. A menos que utilice este término de destino como sinónimo de otro término de destino, o a menos que se excluya, es probable que se convierta en el concepto que aparece en el panel Resultados extraídos. La segunda es la lista de sinónimos que se agrupará bajo el término de destino.

Por ejemplo, si desea que *automobile* se sustituya por *vehicle*, *automobile* es el sinónimo y *vehicle* es el término objetivo.

Puede especificar cualquier palabra en la columna **Sinónimo**, pero si la palabra no se encuentra durante la extracción y el término tiene una opción de coincidencia con Entire, no se puede realizar ninguna sustitución. Sin embargo, no es necesario extraer el término de destino para que se agrupen los sinónimos bajo este término.

Elementos opcionales

Los Elementos opcionales identifican palabras opcionales de un término compuesto que puede pasarse por alto durante la extracción para poder mantener juntos los términos parecidos aunque aparezcan con ligeras diferencias en el texto. Los elementos opcionales son palabras simples que, si se eliminan de un término compuesto, pueden crear una coincidencia con otro término. Estas palabras simples pueden aparecer en cualquier lugar dentro del término compuesto (al principio, en el centro o al final). Puede definir estos elementos opcionales en la pestaña Opcional.

Por ejemplo, para agrupar los términos *ibm* y *ibm corp*, debe declarar que *corp* debe tratarse como un elemento opcional en este caso. En otro ejemplo, si designa el término *access* como un elemento opcional y durante la extracción se encuentran tanto *internet access speed* como *internet speed*, se agruparán bajo el término que se produce con más frecuencia.

Definir sinónimos

En la pestaña Sinónimos, puede especificar una definición de sinónimo en la línea vacía de la parte superior de la tabla. Empiece definiendo el término de destino y sus sinónimos. También puede seleccionar la biblioteca en la que desea que se guarde esta definición. Durante la extracción, todas las apariciones de los sinónimos se agruparán bajo el término de destino de la extracción final. Para obtener más información, consulte el tema [“Añadir términos”](#) en la [página 190](#).

Por ejemplo, si los datos de texto incluyen mucha información de telecomunicaciones, puede tener estos términos: `cellular phone`, `wireless phone` y `mobile phone`. En este ejemplo, es posible que desee definir `cellular` y `mobile` como sinónimos de `wireless`. Si define estos sinónimos, cada aparición extraída de `cellular phone` y `mobile phone` se tratará como el mismo término que `wireless phone` y aparecerá conjuntamente en la lista de términos.

Cuando esté creando sus diccionarios de tipo, puede especificar un término y luego pensar en tres o cuatro sinónimos del mismo. En ese caso, puede escribir todos los términos y luego el término de destino en el diccionario de sustitución, y a continuación arrastrar los sinónimos.

La sustitución de sinónimos también se aplica a las formas declinadas (como los plurales) del sinónimo. En función del contexto puede imponer límites en la forma en que se sustituyen los términos. Pueden utilizarse determinados caracteres para aplicar límites sobre el proceso de la sinonimia:

- **Signo de exclamación (!).** Cuando el signo de exclamación precede directamente al sinónimo `! synonym`, esto indica que no se sustituirán las formas conjugadas del sinónimo por el término de destino. Sin embargo, un signo de exclamación que precede directamente al término objetivo `! target-term` significa que no desea que ninguna parte del término objetivo compuesto o variantes reciba más sustituciones.
- **Asterisco (*).** Un asterisco colocado directamente después de un sinónimo, como por ejemplo `synonym*`, significa que desea que esta palabra se sustituya por el término objetivo. Por ejemplo, si ha definido `manage*` como sinónimo y `management` como objetivo, `associate managers` se sustituirá por el término objetivo `associate management`. También puede añadir un espacio y un asterisco después de la palabra (`synonym *`) como `internet *`. Si ha definido el objetivo como `internet` y los sinónimos como `internet *` y `web *`, `internet access card` y `web portal` se sustituirán por `internet`. No puede comenzar una palabra o una cadena con el comodín de asterisco en este diccionario.
- **Signo de intercalación (^).** Un signo de intercalación y un espacio que precede al sinónimo, como por ejemplo `^ synonym`, significa que la agrupación de sinónimos solo se aplica cuando el término empieza por el sinónimo. Por ejemplo, si define `^ wage` como el sinónimo y `income` como el objetivo, y ambos términos se extraen, se agruparán bajo el término `income`. Sin embargo, si se extraen `minimum wage` y `income`, no se agruparán juntos, ya que `minimum wage` no empieza por `wage`. Debe colocarse un espacio entre este símbolo y el sinónimo.
- **Signo de dólar (\$).** Un espacio y un signo de dólar después del sinónimo, como por ejemplo `synonym $`, significa que la agrupación de sinónimos solo se aplica cuando el término finaliza con el sinónimo. Por ejemplo, si define `cash $` como el sinónimo y `money` como el objetivo, y ambos términos se extraen, se agruparán bajo el término `money`. Sin embargo, si se extraen `cash cow` y `money`, no se agruparán juntos, ya que `cash cow` no termina por `cash`. Debe colocarse un espacio entre este símbolo y el sinónimo.
- **Signo de intercalación (^) y signo de dólar (\$).** Si se utilizan conjuntamente el signo de intercalación y el signo de dólar, como por ejemplo `^ synonym $`, un término coincide con el sinónimo solo si se trata de una coincidencia exacta. Esto significa que no puede aparecer ninguna palabra delante ni detrás del sinónimo del término extraído para que pueda realizarse la agrupación de sinónimos. Por ejemplo, es posible que desee definir `^ van $` como sinónimo y `truck` como objetivo para que solo `van` esté agrupado con `truck`, mientras que `marie van` permanecerá sin cambios. Además, siempre que defina un sinónimo utilizando los símbolos de intercalación y de dólar y esta palabra aparezca en cualquier lugar de texto de origen, el sinónimo se extraerá automáticamente.

Añadir una entrada de sinónimo

1. Con el panel de sustitución visualizado, pulse en la pestaña **Sinónimos** en el ángulo inferior izquierdo.
2. En la línea vacía de la parte superior de la tabla, escriba el término de destino en la columna Objetivos. El término de destino que ha escrito aparece en color. Este color representa el tipo en el que el término aparece o se fuerza, si se da el caso. Si el término aparece en negro, significa que no está en ningún diccionario de tipo.
3. Pulse la segunda celda a la derecha del objetivo y especifique el conjunto de sinónimos. Separe cada entrada utilizando el delimitador global tal como está definido en el cuadro de diálogo Opciones. Para obtener más información, consulte el tema [“Opciones de configuración”](#) en la página 74. Los términos que especifique aparecerán en color. Este color representa el tipo en el que aparece el término. Si el término aparece en negro, significa que no está en ningún diccionario de tipo.
4. Pulse la última celda para seleccionar la biblioteca en la cual desea almacenar esta definición de sinónimo.

Nota: Estas instrucciones le muestran como realizar cambios en la Editor de recursos vista o Editor de plantillas . Tenga en cuenta que también puede realizar este tipo de ajuste directamente desde el panel Resultados de extracción , panel Datos, panel Categorías o el recuadro de diálogo Definiciones en las demás vistas.. Para obtener más información, consulte el tema [“Refinar resultados de extracción”](#) en la página 89.

Definir elementos opcionales

En la pestaña Opcional, puede definir elementos opcionales para la biblioteca que desee. Estas entradas se agrupan juntas para cada biblioteca. Tan pronto como se añade una biblioteca al panel del árbol de bibliotecas, se añade una línea vacía de elemento opcional en la pestaña Opcional.

Todas las entradas pasan a estar en minúsculas automáticamente. El motor de extracción hará coincidir las entradas en mayúsculas y minúsculas del texto.

Nota: Los términos se delimitan mediante el delimitador definido en el diálogo Opciones. Para obtener más información, consulte el tema [“Opciones de configuración”](#) en la página 74. Si el elemento opcional que está especificado incluye el mismo delimitador como parte del término, lo debe preceder una barra inclinada invertida.

Añadir una entrada

1. Con el panel de sustitución visualizado, pulse en la pestaña Opcional en el ángulo inferior izquierdo del editor.
2. Pulse en la celda de la columna de Elementos opcionales correspondiente a la biblioteca a la que desea añadir esta entrada.
3. Especifique el elemento opcional. Separe cada entrada utilizando el delimitador global tal como está definido en el cuadro de diálogo Opciones. Consulte el tema [“Opciones de configuración”](#) en la página 74 para obtener más información.

Desactivación y eliminación de sustituciones

Puede eliminar una entrada de forma temporal desactivándola del diccionario. Al desactivar una entrada, ésta se pasará por alto durante la extracción.

También puede eliminar las entradas obsoletas en el diccionario de sustitución.

Para desactivar una entrada

1. En el diccionario seleccione la entrada que desea desactivar.
2. Pulse la barra espaciadora. Se borra la marca de la casilla de verificación a la izquierda de la entrada.

Nota: También puede desmarcar la casilla de verificación a la izquierda de la entrada para inhabilitarla.

Para eliminar una entrada de sinónimo

1. En el diccionario seleccione la entrada que desea eliminar.
2. En los menús, seleccione **Editar > Suprimir** o pulse la tecla **Suprimir** en el teclado. La entrada desaparece del diccionario.

Para eliminar una entrada de elemento opcional

1. En el diccionario pulse dos veces en la entrada que desea eliminar.
2. Elimine manualmente el término.
3. Pulse Intro para aplicar el cambio.

Diccionarios de exclusión

Un *diccionario de exclusión* es una lista de palabras, frases o cadenas parciales. Se pasarán por alto o se excluirán de la extracción los términos que tengan alguna coincidencia o contengan una entrada en el diccionario de exclusión. Los diccionarios de exclusión se administran en el panel derecho del editor. Por lo general, los términos que añada a esta lista serán palabras o frases de relleno que se utilizan en el texto para conferir continuidad, pero que no aportan información relevante al texto y que además pueden cargar innecesariamente los resultados de la extracción. Si añade estos términos al diccionario de exclusión, tendrá la seguridad de que no se extraerán nunca.

Los diccionarios de exclusión se administran en el panel superior derecho de la pestaña Recursos de la biblioteca del editor. Puede acceder a esta vista mediante **Vista > Editor de recursos** en los menús, si está en una sesión de entorno de trabajo interactiva. De lo contrario, puede editar los diccionarios de una plantilla específica en el Editor de plantillas.

En el diccionario de exclusión puede introducir una palabra, frase o cadena parcial en la línea vacía en la parte superior de la tabla. Puede añadir cadenas de caracteres al diccionario de exclusión como una o más palabras o incluso como palabras parciales utilizando el asterisco como comodín. Las entradas declaradas en el diccionario de exclusión se utilizarán para impedir que los conceptos se extraigan. Si una entrada también se declara en algún otro lugar de la interfaz, como en un diccionario de tipo, se muestra con un signo de tachado en los otros diccionarios, lo que indica que actualmente está excluida. Esta cadena no tiene que aparecer en los datos de texto ni declararse como parte de ningún diccionario de tipo que vaya a aplicarse.

Nota: Si añade un concepto al diccionario de exclusión que también actúa como el objetivo en una entrada de sinónimo, el objetivo y todos sus sinónimos también se excluirán. Para obtener más información, consulte el tema [“Definir sinónimos”](#) en la página 195.

Utilizar comodines (*)

utilizar el comodín asterisco para indicar que desea que la entrada de exclusión sea tratada como una serie parcial. Todos los términos el motor de extracción que encuentre con una palabra que empiece o termine por una cadena especificada en el diccionario de exclusión se excluirán de la extracción final. Sin embargo, existen dos casos en los que el uso de comodines no está permitido:

- Carácter de guión (-) precedido por un comodín asterisco, como *-
- Carácter de apóstrofe (') precedido por un asterisco, como *'s

Tabla 41. Ejemplos de entradas de exclusión		
Entrada	Ejemplo	Resultados
word	<i>Siguiente</i>	No se extraerá ningún concepto (ni sus términos) si contienen la palabra <i>siguiente</i> .
frase	<i>Por ejemplo</i>	No se extraerá ningún concepto (ni sus términos) si contienen la frase <i>por ejemplo</i> .
parcial	<i>copyright*</i>	Excluirá los conceptos (o sus términos) coincidentes o que contengan las variaciones de la palabra <i>copyright</i> , como <i>copyrighted</i> , <i>copyrighting</i> , <i>copyrights</i> , o <i>copyright 2010</i> .

Tabla 41. Ejemplos de entradas de exclusión (continuación)

Entrada	Ejemplo	Resultados
parcial	*ware	Excluirá los conceptos (o sus términos) coincidentes o que contengan las variaciones de la palabra <i>ware</i> , como <i>freeware</i> , <i>shareware</i> , <i>software</i> , <i>hardware</i> , <i>beware</i> , o <i>silverware</i> .

Añadir entradas

- En la línea vacía en la parte superior de la tabla, escriba un término. El término que especifique aparecerá en color. Este color representa el tipo en el que aparece el término. Si el término aparece en negro, significa que no está en ningún diccionario de tipo.

Inhabilitar entradas

Puede eliminar temporalmente una entrada desactivándola en el diccionario de exclusión. Al desactivar una entrada, ésta se pasará por alto durante la extracción.

- En el diccionario de exclusión seleccione la entrada que desea desactivar.
- Pulse la barra espaciadora. Se borra la marca de la casilla de verificación a la izquierda de la entrada.

Nota: También puede desmarcar la casilla de verificación a la izquierda de la entrada para inhabilitarla.

Suprimir entradas

Puede eliminar las entradas que ya no necesite en el diccionario de exclusión.

- En el diccionario de exclusión seleccione la entrada que desea eliminar.
- En los menús, seleccione **Editar > Suprimir**. La entrada desaparece del diccionario.

Capítulo 17. Acerca de los recursos avanzados

Además de los diccionarios de tipo, de exclusión y de sustitución, también puede trabajar con una serie de opciones de recursos avanzados como configuración de agrupación difusa o definiciones de tipo no lingüístico. Puede trabajar con estos recursos en la pestaña Recursos avanzados en la vista Editor de plantillas o Editor de recursos.

En la pestaña Recursos avanzados puede editar la siguiente información:

- **Idioma de destino para recursos.** Utilizado para seleccionar el idioma para el que se crearán y ajustarán los recursos. Para obtener más información, consulte el tema [“Idioma de destino para Recursos”](#) en la página 201.
- **Agrupación inexacta (excepciones).** Se utiliza para excluir parejas de palabras del algoritmo de agrupación difusa (corrección de error ortográfico). Para obtener más información, consulte el tema [“Agrupación difusa”](#) en la página 201.
- **Entidades no lingüísticas.** Se utiliza para activar y desactivar las entidades no lingüísticas que pueden extraerse, así como las expresiones normales y las reglas de normalización que se aplican durante la extracción. Para obtener más información, consulte el tema [“Entidades no lingüísticas”](#) en la página 202.
- **Manejo del idioma.** Se utiliza para declarar los métodos especiales para estructurar frases (patrones de extracción y definiciones forzadas) y para utilizar abreviaturas en el idioma seleccionado. Para obtener más información, consulte el tema [“Gestión de idiomas”](#) en la página 206.

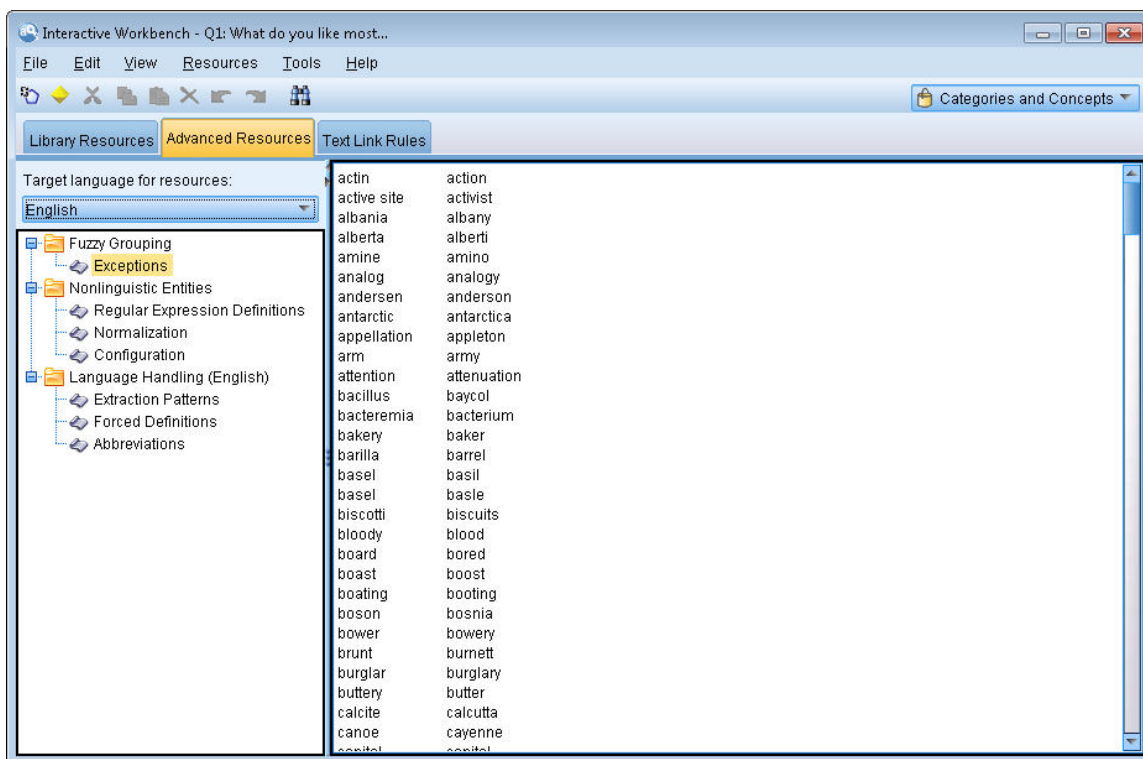


Figura 43. Editor de plantillas de minería de textos: pestaña Recursos avanzados

Nota: Puede utilizar la barra de herramientas Buscar/Reemplazar para buscar información rápidamente o para realizar cambios globales en una sección. Para obtener más información, consulte [“Reemplazo”](#) en la página 200.

Para Editar recursos avanzados

1. Localice y seleccione la sección de recursos que desea editar. Los contenidos aparecerán en el panel derecho.
2. Utilice el menú o los botones de la barra de herramientas para cortar, copiar o pegar el contenido, si es preciso.
3. Edite el archivo o archivos que desea cambiar utilizando las reglas de formato de esta sección. Los cambios se guardan en cuanto los lleva a cabo. Utilice las flechas de deshacer o rehacer de la barra de herramientas para invertir la acción sobre los cambios anteriores.

Buscar

En algunos casos, puede que necesite localizar información rápidamente en una sección determinada. Por ejemplo, si lleva a cabo un análisis de enlace de texto, puede tener cientos de definiciones de macros y patrones. Con la característica Buscar, puede encontrar rápidamente una regla específica. Para buscar información en una sección, puede utilizar la barra de herramientas Buscar.

Para utilizar la característica Buscar

1. Localice y seleccione la sección de recursos que desea buscar. Los contenidos aparecen en el panel derecho del editor.
2. En los menús, seleccione **Editar > Buscar**. La barra de herramientas Buscar aparecerá en el ángulo superior derecho del cuadro de diálogo Editar recursos avanzados.
3. Escriba la cadena de palabras que desea buscar en el cuadro de texto. Puede utilizar los botones de la barra de herramientas para controlar las mayúsculas/minúsculas, la coincidencia parcial y la dirección de la búsqueda.
4. Pulse **Buscar** para iniciar la búsqueda. Si se encuentra una coincidencia, el texto se resalta en la ventana.
5. Pulse **Buscar** otra vez para buscar la siguiente coincidencia.

Nota: Al trabajar en la pestaña Reglas de enlace de texto, la opción Buscar sólo está disponible cuando ve el código fuente.

Reemplazo

En algunos casos, puede que necesite realizar actualizaciones globales en los recursos avanzados. La característica Reemplazar puede ayudarle a realizar actualizaciones uniformes en el contenido.

Para utilizar la característica Reemplazar

1. Localice y seleccione la sección de recursos que desea buscar y reemplazar. Los contenidos aparecen en el panel derecho del editor.
2. En los menús, seleccione **Editar > Sustituir**. Aparecerá el cuadro de diálogo Reemplazar.
3. En el cuadro de texto **Buscar**, escriba la cadena de palabras que desea buscar.
4. En el cuadro de texto **Reemplazar por**, escriba la cadena que desea utilizar en lugar del texto de búsqueda.
5. Seleccione **Coincidir sólo palabra completa** si desea buscar o reemplazar únicamente palabras completas.
6. Seleccione **Coincidir mayúsculas y minúsculas** si desea buscar o reemplazar únicamente palabras que coincidan con las mayúsculas y minúsculas exactamente.
7. Pulse **Buscar siguiente** para buscar una coincidencia. Si se encuentra una coincidencia, el texto se resalta en la ventana. Si no desea reemplazar esta coincidencia, pulse en **Buscar siguiente** de nuevo hasta que encuentre una coincidencia que desee reemplazar.
8. Pulse **Reemplazar** para sustituir la coincidencia seleccionada.

9. Pulse **Reemplazar** para sustituir todas las coincidencias de la sección. Se abrirá un mensaje indicando el número de reemplazos realizados.

10. Cuando haya terminado de reemplazar, pulse en **Cerrar**. El cuadro de diálogo se cierra.

Nota: Si ha cometido un error de sustitución, puede deshacer la sustitución cerrando el recuadro de diálogo y eligiendo **Editar > Deshacer** en los menús. Deberá realizar esta acción una vez por cada cambio que desee deshacer.

Idioma de destino para Recursos

Los recursos se crean para un idioma de texto determinado. El idioma en el que se adaptan estos recursos se define en la pestaña Recursos avanzados. Si fuese necesario puede ir a otro idioma seleccionándolo en el recuadro combinado **Idioma de destino para los recursos**. Además, el idioma aparecerá aquí como el idioma de cualquier paquete de análisis de texto que cree con estos recursos.

Importante: No necesitará cambiar el idioma de sus recursos. Al hacerlo se podrían crear problemas cuando sus recursos ya no coincidan con el idioma de extracción. Aunque raramente se utiliza, puede cambiar un idioma si ha planificado utilizar la opción de idioma ALL durante la extracción porque esperaba tener texto en más de un idioma. Al cambiar el idioma, puede acceder, por ejemplo, los recursos del manejo del idioma para patrones de extracción, abreviaturas y definiciones forzadas para el idioma secundario que le interese. Sin embargo, tenga en cuenta que antes de publicar o guardar los cambios del recurso que haya hecho o ejecutar otra extracción, establezca el idioma nuevamente en el idioma primario que le interese extraer.

Agrupación difusa

En el nodo de Text Mining y el diálogo Configuración de extracción, si selecciona **Acomodar la ortografía a un límite mínimo de caracteres raíz de**, significa que ha activado el algoritmo de agrupación difusa.

La agrupación difusa ayuda a agrupar las palabras que comúnmente se escriben mal o que tienen una ortografía similar pasando temporalmente por alto todas las vocales (excepto la primera) o consonantes dobles o triples de las palabras extraídas, y luego comparándolas para comprobar si son las mismas. Durante el proceso de extracción, la característica de agrupación difusa se aplicará a los términos extraídos, y los resultados se comparan para determinar si se han encontrado coincidencias. En caso afirmativo, los términos originales se agrupan juntos en la lista de extracción final. Se agrupan bajo el término que aparece más veces en los datos.

Nota: Si los dos términos de comparación están asignados a tipos diferentes, excluyendo el tipo <Unknown>, la técnica de agrupación difusa no se aplicará a esta pareja de términos. En otras palabras, para poder aplicar la técnica, los términos deben pertenecer al mismo tipo o al tipo <Unknown>.

Si ha activado esta característica y descubre que dos palabras con ortografía similar se han agrupado incorrectamente, puede excluir dichas palabras de la agrupación difusa. Para ello escriba las parejas mal agrupadas en la sección Excepciones de la pestaña Recursos avanzados. Para obtener más información, consulte el tema [Capítulo 17, “Acerca de los recursos avanzados”, en la página 199](#).

En el ejemplo siguiente se muestra el proceso de la agrupación difusa. Si se ha activado la agrupación difusa, estas palabras parecen ser iguales y se emparejan de la forma siguiente:

color -> colr	mountain -> montn
colour -> colr	montana -> montn
modeling -> modlng	furniture -> furntr
modelling -> modlng	furnature -> furntr

En el ejemplo anterior, lo más probable es que desee impedir que mountain y montana se agrupen. Por lo tanto, puede incluir estas palabras en la sección Excepciones de la siguiente manera:

mountain	montana
----------	---------

Importante: En algunos casos, las excepciones de agrupación difusa no finalizan el proceso de emparejamiento de 2 palabras porque se han aplicado algunas reglas de sinónimos. En ese caso, es

posible que desee intentar especificar sinónimos utilizando el comodín de signo de exclamación (!) para prohibir que las palabras se conviertan en sinónimos en la salida. Para obtener más información, consulte [“Definir sinónimos”](#) en la página 195.

Reglas de formato para excepciones de agrupación difusa

- Defina solo una pareja de excepción por línea.
- Utilice palabras simples o compuestas.
- Utilice solamente caracteres en minúsculas para las palabras. Las palabras en mayúsculas se pasarán por alto.
- Utilice un carácter TAB para separar cada palabra de un par.

Entidades no lingüísticas

Cuando trabaja con determinados tipos de datos, puede ser de gran interés extraer fechas, números de la seguridad social, porcentajes u otras entidades no lingüísticas. Estas entidades están explícitamente declaradas en el archivo de configuración, donde puede activar o desactivar las entidades. Para obtener más información, consulte el tema [“Configuración”](#) en la página 205. A fin de optimizar la salida del motor de extracción, la entrada del procesamiento no lingüístico se normaliza para agrupar a las entidades parecidas de acuerdo a formatos predefinidos. Para obtener más información, consulte el tema [“Normalización”](#) en la página 205.

Nota: Puede activar y desactivar la extracción de entidades no lingüísticas en la configuración de extracción.

Entidades no lingüísticas disponibles

Pueden extraerse las entidades no lingüísticas de la tabla siguiente. El nombre del tipo está entre paréntesis.

<i>Tabla 42. Entidades no lingüísticas que se pueden extraer</i>	
Direcciones	(<Address>)
Aminoácidos	(<Aminoacid>)
Divisas	(<Currency>)
Fechas	(<Date>)
Retardo	(<Delay>)
Digits	(<Digit>)
Direcciones de correo electrónico	(<email>)
Direcciones de HTTP/URL	(<url>)
Dirección IP	(<IP>)
Organizaciones	(<Organization>)
Porcentajes	(<Percent>)
Productos	(<Product>)
Proteínas	(<Gene>)
Números de teléfono	(<PhoneNumber>)
Times	(<Time>)
Seguridad social de EEUU	(<SocialSecurityNumber>)

Tabla 42. Entidades no lingüísticas que se pueden extraer (continuación)

Pesos y medidas	(<Weights-Measures>)
-----------------	----------------------

Limpieza del texto para proceso

Antes de que se realice la extracción de entidades no lingüísticas, el texto de entrada se limpia. Durante este paso, se realizan los cambios temporales siguientes para que las entidades no lingüísticas puedan identificarse y extraerse como tales:

- Cualquier secuencia de dos o más espacios se sustituye por un espacio único.
- Las tabulaciones se sustituyen por espacios.
- Los caracteres de secuencia o caracteres únicos de final de línea se sustituyen por un espacio, mientras que las secuencias múltiples de final de línea se marcan como final de un párrafo. El final de línea puede indicarse mediante retornos de carro (CR) y cambio de línea (LF), o incluso ambos a la vez.
- Los códigos HTML y XML se pasan por alto temporalmente y se ignoran.

Definiciones de expresiones regulares

Cuando se extraen entidades no lingüísticas, puede editar o añadir a la expresión regular aquellas definiciones que se utilizan para identificar las expresiones regulares. Esto se hace en la sección **Definiciones de expresiones regulares** de la pestaña Recursos avanzados. Para obtener más información, consulte el tema [Capítulo 17, “Acerca de los recursos avanzados”](#), en la página 199.

El archivo está dividido en varias secciones. La primera sección se llama [macros]. Además de dicha sección, puede existir una sección adicional para cada entidad no lingüística. Puede añadir secciones a este archivo. En cada sección, las reglas están numeradas (*regexp1*, *regexp2*, etc.). Estas reglas se deben numerar de forma secuencial desde 1–n. Toda interrupción en la numeración hará que el proceso de este archivo quede suspendido.

En determinados casos, una entidad puede depender del idioma. Se considera que una entidad depende del idioma si toma un valor que no sea 0 como parámetro de idioma en el archivo de configuración. Para obtener más información, consulte el tema [“Configuración”](#) en la página 205. Cuando una entidad depende del idioma, el idioma se debe utilizar como prefijo del nombre de la sección, como por ejemplo [english/PhoneNumber]. Esta sección contendrá reglas que se apliquen solo a los números de teléfono ingleses si se asigna el valor de 2 al idioma de la entidad PhoneNumber.

Importante: Si realiza cambios en este archivo o en cualquier otro en el editor, y el motor de extracción deja de funcionar como se esperaba, utilice la opción **Restablecer originales** en la barra de herramientas para restaurar el archivo con el contenido original enviado con el producto. Es necesario estar familiarizado con las expresiones regulares para trabajar con este archivo. Si necesita ayuda adicional en esta área, póngase en contacto con IBM Corp. para obtener ayuda.

Caracteres especiales. [] {} () \ * + ? | ^ \$

Todos los caracteres coinciden consigo mismos excepto por los siguientes caracteres especiales, que se utilizan con un propósito específico en una expresión: . [{ () \ * + ? | ^ \$ Para utilizar estos caracteres de esta forma, deben estar precedidos de una barra inclinada invertida (\) en la definición.

Por ejemplo, si intentaba extraer direcciones web, el carácter de punto y aparte es muy importante en la entidad, por lo que debe colocar una barra inclinada invertida, así:

```
www\[a-z]+\.[a-z]+
```

Operadores y cuantificadores de repetición ? + * {}

Para que las definiciones sean más flexibles, puede utilizar varios comodines que sean estándares en las expresiones regulares. Son * ? +

- **Asterisco** * indica que hay *ceros* o *más* elementos de la cadena precedente. Por ejemplo, ab*c coincide con "ac", "abc", "abbbc", etc.

- El signo más + indica que hay *uno o más* elementos de la cadena precedente. Por ejemplo, `ab+c` coincide con `"abc"`, `"abbc"`, `"abbbc"`, pero no con `"ac"`.
- Signo de interrogación ? indica que hay *cero o un* elementos de la cadena precedente. Por ejemplo, `modell?ing` coincide tanto con `"modelado"` como con `"modelado"`.
- Limitar la repetición mediante llaves {} indica los límites de la repetición. Por ejemplo:
`[0-9]{n}` coincide con un dígito que se repite exactamente *n* veces. Por ejemplo, `[0-9]{4}` coincidirá con `"1998"`, pero no con `"33"` o con `"19983"`.
`[0-9]{n,}` coincide con un dígito repetido *n o más* veces. Por ejemplo, `[0-9]{3,}` coincidirá con `"199"` o con `"1998"`, pero no con `"19"`.
`[0-9]{n,m}` coincide con un dígito que se repite entre *n y m veces, inclusive*. Por ejemplo, `[0-9]{3,5}` coincidirá con `"199"`, `"1998"` o `"19983"`, pero no con `"19"` ni con `"199835"`.

Guiones y espacios opcionales

En algunos casos, puede que necesite incluir un espacio opcional en una definición. Por ejemplo, si desea extraer monedas como `"pesos uruguayos"`, `"peso uruguayo"`, `"pesos uruguay"`, `"peso uruguay"`, `"pesos"` o `"peso"`, tendrá que afrontar el hecho de que puede haber dos palabras separadas por un espacio. En este caso, esta definición debe escribirse como `(uruguayan |uruguay )?pesos?`. Puesto que `uruguayo` o `uruguay` van seguidos de un espacio cuando se utiliza con `pesos/peso`, el espacio opcional debe definirse dentro de la secuencia opcional `(uruguayan |uruguay )`. Si el espacio no estaba en la secuencia opcional como, por ejemplo, `(uruguayan|uruguay)? pesos?`, no coincidiría con `"pesos"` o `"peso"` ya que el espacio sería necesario.

Si busca una serie de objetos que incluyen guiones (-) en una lista, el guión debe estar definido al final. Por ejemplo, si busca una coma (,) o un guión (-), utilice `[ , -]` pero nunca `[- , ]`.

Orden de las cadenas en listas y macros

Siempre debe definir la secuencia más larga antes que la más corta; de lo contrario, nunca se detectará la más larga, puesto que la coincidencia se producirá en la secuencia más corta. Por ejemplo, si estuviera buscando cadenas `"trillón"` o `"trillo"`, entonces `"trillón"` debe definirse antes que `"trillo"`. Así, por ejemplo, `(billion|bill)` y no `(bill|billion)`. Esto también se aplica a las macros, puesto que estas son listas de cadenas.

Orden de las reglas en la sección de definición

Defina una regla por línea. En cada sección, las reglas están numeradas (`regexp1`, `regexp2`, etc.). Estas reglas se deben numerar de forma secuencial desde 1–*n*. Toda interrupción en la numeración hará que el proceso de este archivo quede suspendido. Para desactivar una entrada, coloque un símbolo de número (#) al principio de cada línea que se utiliza para definir la expresión regular. Para activar una entrada, elimine el símbolo de número (#) del principio de la línea.

En cada sección, las reglas más específicas deben definirse antes que las más generales para asegurar un proceso apropiado. Por ejemplo, si se está buscando una fecha en el formato `"mes año"` y en el formato `"mes"`, se debe definir la regla `"mes año"` antes que la regla `"mes"`. Este es un ejemplo de cómo debe definirse:

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

y no

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

Uso de macros en las reglas

Cuando se utiliza una secuencia específica en varias reglas, puede utilizar una macro. Así pues, si necesita cambiar la definición de esta secuencia, deberá cambiarla solo una vez, y no es necesario que lo haga en todas las reglas a las que haga referencia. Por ejemplo, imagine que tiene la macro siguiente:

```
MONTH=((january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)
(\.)?)
```

Siempre que se refiera al nombre de la macro, debe ir entre \$(), como en: `regex1=$(MONTH)`

Todas las macros deben definirse en la sección `[macros]`.

Normalización

Cuando se extraen entidades no lingüísticas, las entidades que se encuentran se normalizan para agrupar entidades parecidas según los formatos predefinidos. Por ejemplo, los símbolos de moneda y sus equivalentes en palabras se consideran lo mismo. Las entradas de normalización se almacenan en la sección **Normalización** de la pestaña Recursos avanzados. Para obtener más información, consulte el tema [Capítulo 17, “Acerca de los recursos avanzados”](#), en la página 199. El archivo está dividido en varias secciones.

Importante: Este archivo solo deben utilizarlo usuarios experimentados. Es altamente improbable que necesite cambiar este archivo. Si necesita ayuda adicional en esta área, póngase en contacto con IBM Corp. para obtener ayuda.

Reglas de formato para la normalización

- Añada solo una entrada de normalización por línea.
- Respete estrictamente las secciones de este archivo. No pueden añadirse secciones nuevas.
- Para desactivar una entrada, coloque un símbolo de número (#) al principio de la línea. Para activar una entrada, elimine el símbolo de número (#) del principio de la línea.

Fechas en inglés en normalización

De forma predeterminada las fechas en una plantilla en inglés se reconocen en un formato de fecha de estilo norteamericano; es decir: mes, día, año. Si necesita cambiarla al formato día, mes, año, inhabilite la línea "formato:EEUU" (al añadir # al principio de la línea) y habilite "formato:Reino Unido" (al eliminar # de esa línea).

Configuración

Puede activar y desactivar los tipos de entidades no lingüísticas que desee extraer en el archivo de configuración de entidades no lingüísticas. Al desactivar las entidades que no necesita, puede disminuir el tiempo de proceso necesario. Esto se hace en la sección **Configuración** de la pestaña Recursos avanzados. Para obtener más información, consulte el tema [Capítulo 17, “Acerca de los recursos avanzados”](#), en la página 199. Si se habilita la extracción no lingüística, el motor de extracción lee este archivo de configuración durante el proceso de extracción para determinar que tipos de entidades no lingüísticas deben extraerse.

La sintaxis de este archivo es la siguiente:

```
#name<TAB>Language<TAB>Code
```

Tabla 43. Sintaxis del archivo de configuración	
Etiqueta de columna	Descripción
#name	Término por el que se hará referencia a las entidades no lingüísticas en los otros dos archivos requeridos para la extracción de entidades no lingüísticas. Los nombres que se utilizan aquí son sensibles a mayúsculas y minúsculas.

Tabla 43. Sintaxis del archivo de configuración (continuación)

Etiqueta de columna	Descripción
Language	El idioma de los documentos . Se recomienda seleccionar el idioma específico; sin embargo, existe una opción Cualquiera . Las posibles opciones son: 0 = Cualquiera que se utilice cuando un regexp no es específico de un idioma y puede utilizarse en varias plantillas con distintos idiomas, por ejemplo un IP/URL/direcciones de correo electrónico; 1 = francés; 2 = inglés; 4 = alemán; 5 = español; 6 = holandés; 8 = portugués; 10 = italiano.
Code	Código de categoría léxica. La mayoría de las entidades toman un valor de "s" excepto en unos pocos casos. Los valores posibles son: s = palabras excluidas; a = adjetivo; n = sustantivo. Si está activado, las entidades no lingüísticas se extraen en primer lugar y se aplican los patrones de extracción para identificar su rol en un contexto más amplio. Por ejemplo, a los porcentajes se les asigna un valor de "a." Supongamos que se extrae 30% como entidad no lingüística. Se identificaría como un adjetivo. Entonces, si el texto contenía "30% de incremento salarial," la entidad no lingüística "30%" se ajusta al patrón de categoría léxica "ann" (adjetivo sustantivo sustantivo).

Orden en la definición de entidades

El orden en que se declaran las entidades en este archivo es relevante y afecta a la forma en que se extraerán. Se aplican en el orden de la lista. Si cambia el orden, cambiará el resultado. Las entidades no lingüísticas más específicas deben definirse antes que las más generales.

Por ejemplo, la entidad no lingüística "Aminoacid" está definida por:

```
regexp1=($(AA) -?$(NUM))
```

donde \$(AA) corresponde a "(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)", que son específicas secuencias de 3 letras que corresponden a aminoácidos particulares.

Por otro lado, la entidad no lingüística "Gene" es más general y está definida por:

```
regexp1=p[0-9]{2,3}
regexp2=[a-z]{2,4}-?[0-9]{1,3}-?[r]
regexp3=[a-z]{2,4}-?[0-9]{1,3}-?p?
```

Si se define "Gene" antes que "Aminoacid" en la sección Configuración, "Aminoacid" nunca será coincidente, ya que regexp3 de "Gene" siempre coincidirá primero.

Reglas de formato para la configuración

- Utilice un carácter TAB para separar cada entrada de una columna.
- No elimine ninguna línea.
- Respete la sintaxis que se muestra en la tabla anterior.
- Para desactivar una entrada, coloque un símbolo de número (#) al principio de la línea. Para activar una entidad, elimine el símbolo de número (#) del principio de la línea.

Gestión de idiomas

Todos los idiomas actuales tienen maneras especiales de expresar ideas, estructurar frases y utilizar abreviaturas. En la sección Gestión de idioma, puede editar patrones de extracción, forzar definiciones para dichos patrones y declarar abreviaturas para el idioma que ha seleccionado en la lista desplegable de Idiomas.

- Patrones de extracción

- Definiciones forzadas
- Abreviaturas

Patrones de extracción

Al extraer información de los documentos, el motor de extracción aplica un conjunto de patrones de extracción de piezas de discurso a una "pila" de palabras en el texto para identificar términos candidatos (palabras y frases) para la extracción. Puede añadir o modificar los patrones de extracción.

Las categorías léxicas incluyen elementos gramaticales, como sustantivos, adjetivos, participios, determinantes, preposiciones, conjunciones, nombres propios, iniciales y partículas. Un patrón de extracción de categorías léxicas está compuesto por una serie de estos elementos. En los productos de minería de textos de IBM Corp., cada categoría léxica está representada por un único carácter para facilitar la definición de sus patrones. Por ejemplo, un adjetivo se representa mediante la letra minúscula *a*. El conjunto de códigos soportados aparece de forma predeterminada en la parte superior de cada sección de patrones de extracción predeterminados junto con un conjunto de patrones y ejemplos de cada patrón para ayudarle a comprender cada código que se utiliza.

Reglas de formato para los patrones de extracción

- Un patrón por línea.
- Utilice *#* al principio de una línea para desactivar un patrón.

El orden en el que aparecen los patrones de extracción es muy importante, porque el motor de extracción lee una secuencia determinada de palabras solo una vez, y se le asigna al primer patrón de extracción para el que el motor encuentra una coincidencia.

Partes de códigos de categoría léxica soportados

A continuación figura una tabla de todas las partes de códigos de categoría léxica soportados definidos en el diccionario compilado de inglés.

Todas las partes de categoría léxica que se utilizan en una plantilla en particular aparecen listados en la parte superior de **Recursos avanzados > Patrones de extracción**.

La principal diferencia entre la plantilla de recursos básicos y la plantilla de opiniones es que cuando se utilizan determinantes ("d") y preposiciones ("c") mínimos en los básicos, en las opiniones se utilizan sus equivalentes ampliados ("e" y "r"). "0" y "1" tienen un uso limitado en todas las plantillas de opiniones. Consulte **Recursos avanzados > Gestión de idioma (Inglés) > Definiciones forzadas y patrones de extracción**.

Otras plantillas en inglés pueden utilizar algunas categorías léxicas que no figuran en el diccionario (por ejemplo, "w" y "W", en la plantilla de Inteligencia de mercado). Pero en tal caso, dichas categorías léxicas se asignan a palabras específicas debajo de **Recursos avanzados > Definiciones forzadas**.

Tabla 44. Partes de códigos de categoría léxica soportados		
Código	Descripción	Ejemplo
a	adjetivo	abdominal, azul...
A	no utilizado	no utilizado
b	adverbio	frecuentemente, a menudo, mucho...
B	no utilizado	no utilizado
c	preposición	"de"
C	código interno de palabras con errores ortográficos	
d	determinante	"el"

Tabla 44. Partes de códigos de categoría léxica soportados (continuación)

Código	Descripción	Ejemplo
D	no utilizado	no utilizado
e	determinante ampliado	el, un, mi, tu...
E	no utilizado	no utilizado
f	nombre	John, Mary...
F	no utilizado	no utilizado
g	no utilizado	no utilizado
G	adjetivo de nacionalidad	francés, americano...
h	no utilizado	no utilizado
H	no utilizado	no utilizado
i	letras únicas de inicial seguidas de "."	"a.", "w." y algunas letras únicas como "w" (utilizada para extraer nombres de personas como John W. Doe)
I	no utilizado	no utilizado
j	no utilizado	no utilizado
J	no utilizado	no utilizado
k	no utilizado	no utilizado
K	no utilizado	no utilizado
l	no utilizado	no utilizado
L	no utilizado	no utilizado
m	nombre o desconocido	perro, ibm
M	no utilizado	no utilizado
n	nombre	perro
N	unos pocos nombres propios	ibm
o	coordinación	"y", "&"
O	no utilizado	no utilizado
p	participio pasado	abandonado, interpretado...
P	no utilizado	no utilizado
q	no utilizado	no utilizado
Q	calificador	caro, pequeña, buena...
r	preposición ampliada	de, entre, contra, desde...
R	no utilizado	no utilizado
s	palabra vacía	cualquier palabra que no desee extraer
S	no utilizado	no utilizado
t	tratamiento	Sr., Sra., condesa, presidente...
T	no utilizado	no utilizado

Tabla 44. Partes de códigos de categoría léxica soportados (continuación)

Código	Descripción	Ejemplo
u	desconocido por definición, no está en el diccionario	
U	no utilizado	no utilizado
v	verbo	comer, come, comieron, comiendo...
V	verbo en infinitivo	comer, ...
w	no utilizado	no utilizado
W	no utilizado	no utilizado
x	auxiliar	ser
X	no utilizado	no utilizado
y	partícula	von, di, de, ... (utilizado para extraer nombres de persona: John von Doe)
Y	no utilizado	no utilizado
z	no utilizado	no utilizado
Z	no utilizado	no utilizado
0	adverbio de opinión	Sólo en opiniones. Consulte Recursos avanzados > Gestión de idioma (Inglés) > Definiciones forzadas .
1	"a" en opiniones	Consulte Recursos avanzados > Gestión de idioma (Inglés) > Definiciones forzadas .
2	no utilizado	no utilizado
3	no utilizado	no utilizado
4	no utilizado	no utilizado
5	no utilizado	no utilizado
6	no utilizado	no utilizado
7	no utilizado	no utilizado
8	no utilizado	no utilizado
9	no utilizado	no utilizado

Definiciones forzadas

Al extraer información de documentos, el motor de extracción examina el texto e identifica la parte de discurso para cada palabra que encuentra. En algunos casos, una palabra puede ajustarse a varios roles diferentes en función del contexto. Si desea forzar una palabra para que tome un rol de categoría léxica o para excluirla completamente del proceso, puede hacerlo en la sección **Definición forzada** de la pestaña Recursos avanzados. Para obtener más información, consulte el tema [Capítulo 17, “Acerca de los recursos avanzados”](#), en la página 199.

Para forzar un rol de categoría léxica para una palabra determinada, debe añadir una línea en esta sección empleando esta sintaxis:

```
term:code
```

Tabla 45. Descripción de la sintaxis

Entrada	Descripción
term	Nombre de término.
code	Código de un solo carácter que representa el rol de categoría léxica. Puede enumerar hasta seis códigos de categorías léxicas diferentes por unitérmino. Además, puede impedir la extracción de una palabra en palabras/frases compuestas utilizando el código en minúsculas, como por ejemplo <code>additional:s</code> .

Reglas de formato para definiciones forzadas

- Una línea por palabra.
- Los términos no pueden tener el carácter de dos puntos.
- Utilice el carácter en minúscula `s` como código de categoría léxica para impedir que una palabra se extraiga.
- Utilice un máximo de seis códigos de categorías léxicas por línea. Los códigos soportados de categorías léxicas se muestran en la sección sobre Patrones de extracción. Para obtener más información, consulte el tema [“Patrones de extracción”](#) en la página 207.
- Utilice un asterisco (*) como carácter comodín al final de una cadena para conseguir coincidencias parciales. Por ejemplo, si especifica `add*:s`, las palabras como `add`, `additional`, `additionally`, `addendum` y `additive` nunca se extraen como un término o como parte de un término de palabra compuesto. Sin embargo, si se declara de manera explícita una coincidencia de palabra como término en un diccionario compilado o en las definiciones forzadas, sí que se extraerá. Por ejemplo, si especifica tanto `add*:s` como `addendum:n`, todavía se extraerá `addendum` si se encuentra en el texto.

Abreviaturas

Cuando el motor de extracción está procesando texto, generalmente considera los puntos como una indicación de que la frase ha terminado. Esto suele ser correcto; sin embargo, este manejo de los puntos no se aplica en el caso de las abreviaturas.

Si extrae términos del texto y detecta que no se han manejado bien determinadas abreviaturas, deberá declararlas de manera explícita en esta sección.

Nota: si la abreviatura ya aparece en una definición de sinónimo o se define como un término en un diccionario de tipo, no hay ninguna necesidad de añadir la entrada de abreviatura aquí.

Reglas de formato para las abreviaturas

- Defina una abreviatura por línea.

Capítulo 18. Sobre las reglas de enlaces de texto

El análisis de enlaces de texto (TLA) es un patrón que hace coincidir tecnologías que se utilizan para extraer las relaciones que se encuentran en los textos utilizando un conjunto de reglas. Cuando se habilita el análisis de enlaces de texto para la extracción, los datos de texto se comparan con estas reglas. Cuando se encuentra una coincidencia, el patrón de análisis de enlaces de texto se extrae y presenta. Estas reglas se definen en la pestaña Reglas de enlaces de texto.

Por ejemplo, el extraer conceptos que representan ideas simples sobre una organización puede no resultarle lo suficientemente interesante, pero al utilizar el TLA, también podría aprender sobre los enlaces entre distintas organizaciones o las personas asociadas a la organización. El TLA también se puede utilizar para extraer opiniones de temas sobre cómo se sienten las personas con respecto a un producto o experiencia en particular.

Para beneficiarse del TLA, debe tener recursos que contengan reglas de enlaces de texto (TLA). Cuando selecciona una plantilla, puede ver cuáles son las plantillas que contienen reglas de TLA si tienen o no un icono en la columna TLA.

Los patrones de análisis de enlaces de texto se encuentran en los datos de texto durante la fase de coincidencia de patrón o durante el proceso de extracción. Durante esta fase, se comparan las reglas con los datos de texto y cuando se encuentra una coincidencia, esta información se extrae como un patrón. Hay momentos cuando es posible que desee obtener más del análisis de enlaces de texto o cambiar cómo se compara algo. En estos casos, puede refinar las reglas para adaptarlas a sus necesidades particulares. Esto se lleva a cabo en la pestaña Reglas de enlaces de textos.

Nota: A partir de la versión 8.2, están disponibles las Reglas de reasignación de tipo (TRR). Las TRR transforman una secuencia de tipos, macros y elementos en un nuevo concepto con un tipo específico. Se pueden utilizar en plantillas Opiniones para capturar opiniones con un cambio de polaridad. Para obtener más información, consulte [“Reglas de reasignación de tipo”](#) en la página 155.

Dónde trabajar en las reglas de enlaces de texto

Puede editar y crear reglas directamente en el separador Reglas de enlace de texto en la vista Editor de plantillas o Editor de recursos. Para ayudarle a ver cómo es posible que las reglas hagan coincidir un texto, puede ejecutar una simulación en este separador. Durante la simulación, se ejecuta una extracción solo en los datos de simulación de ejemplo y las reglas de enlace de texto se aplican para ver si los patrones coinciden. Las reglas que coincidan con el texto se muestran en el panel de simulaciones. Basándose en las coincidencias, puede elegir editar las reglas y macros para modificar la forma en que el texto se compara.

A diferencia de otros recursos avanzados, las reglas TLA pertenecen a bibliotecas específicas, por lo tanto, solo puede utilizarlas una biblioteca a la vez. Desde el Editor de plantillas o Editor de recursos, vaya al separador **Reglas de enlace de texto**. En este separador puede especificar la biblioteca en la plantilla que contiene las reglas TLA que desea editar o utilizar. Por este motivo, recomendamos encarecidamente que almacene todas las reglas en una biblioteca, a menos que exista una razón muy específica por la que no se desee esto.

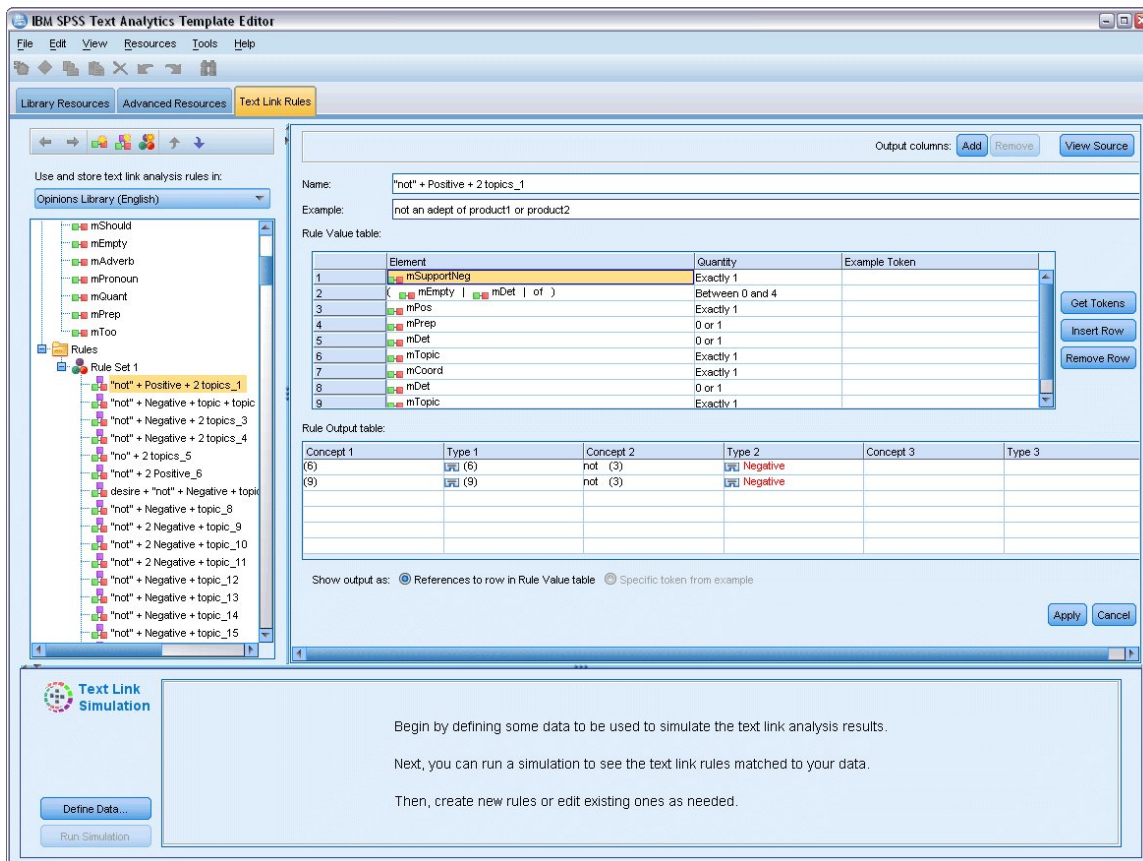


Figura 44. Pestaña Reglas de enlace de texto

Dónde comenzar

Hay varias maneras de comenzar a trabajar en el editor del separador Reglas de enlace de texto:

- Comience por simular los resultados con algunos textos de ejemplo, y edite o cree reglas de coincidencia basadas en cómo el conjunto de reglas actuales extrae los patrones de los datos de simulación.
- Cree una nueva regla desde cero o edite una ya existente.
- Trabajar directamente en la vista de origen.

Cuándo editar o crear reglas

Aunque las reglas de análisis de enlace de texto que se entregan con cada plantilla son a menudo adecuadas para la extracción de varias relaciones simples o complejas del texto, hay ocasiones en que quizás desee realizar algunos cambios a dichas reglas para crear algunas propias. Por ejemplo:

- Para capturar una idea o relación que no se extraiga con las reglas existentes mediante la creación de una nueva regla o macro.
- Para cambiar el comportamiento predeterminado de un tipo añadido a los recursos. Esto requiere generalmente que edite una macro como mTopic o mNonLingEntities. Para obtener más información, consulte el tema “Macros especiales: mTopic, mNonLingEntities, SEP” en la página 218.
- Para añadir nuevos tipos a reglas y macros de análisis de enlace de texto existentes. Por ejemplo, si cree que el tipo <Organization> es demasiado amplio, podría crear nuevos tipos para organizaciones de diversos sectores empresariales como <Pharmaceuticals>, <Car Manufacturing>, <Finance>, etc. En este caso, debe editar las reglas de análisis de enlaces de texto y/o crear una macro para que estos nuevos tipos se tomen en cuenta y se los procese según corresponda.

- Para añadir tipos a una regla de análisis de enlaces de texto existente. Por ejemplo, digamos que tiene una regla que captura el siguiente texto `john doe called jane doe` pero desea que esta regla que captura las comunicaciones telefónicas también capture intercambios de correo electrónico. Puede añadir el tipo de entidad no lingüística para el correo electrónico a la regla de modo que también capture texto como: `johndoe@ibm.com emailed janedoe@ibm.com`.
- Para modificar ligeramente una regla existente, en lugar de crear una nueva. Por ejemplo, digamos que tiene una regla que coincide con el siguiente texto `xyz is very good`, pero desea que esta regla también capture `xyz is very, very good`.

Simulación de resultados del análisis de enlace de texto

Para ayudar a definir nuevas reglas de enlace de texto o ayudar a entender cómo ciertas oraciones se comparan durante el análisis de enlace de texto, a menudo es útil tomar un fragmento de texto de ejemplo y ejecutar una simulación. Durante la simulación, solo se ejecuta una extracción en los datos de simulación de ejemplo utilizando el conjunto actual de recursos lingüísticos y los valores actuales de extracción. El objetivo es obtener los resultados simulados y utilizarlos para mejorar las reglas, crear unas nuevas o comprender mejor cómo se produce la coincidencia. Por cada fragmento de texto (frase, palabra u oración dependiendo del contexto), una salida de simulación muestra la colección de señales y las reglas de TLA que han descubierto un patrón en dicho texto. Una **señal** se define como cualquier palabra o frase de palabras identificada durante el proceso de extracción.

A diferencia de otros recursos avanzados, las reglas TLA pertenecen a bibliotecas específicas, por lo tanto, solo puede utilizarlas una biblioteca a la vez. Desde el Editor de plantillas o Editor de recursos, vaya al separador **Reglas de enlace de texto**. En este separador puede especificar la biblioteca en la plantilla que contiene las reglas TLA que desea editar o utilizar. Por este motivo, recomendamos encarecidamente que almacene todas las reglas en una biblioteca, a menos que exista una razón muy específica por la que no se desee esto.

Importante: Recomendamos encarecidamente que si utiliza un archivo de datos, se asegure de que el texto que contiene sea corto para poder minimizar el tiempo de proceso. El objetivo de la simulación es ver cómo un fragmento de texto se interpreta y entender cómo las reglas coinciden con este texto. Esta información le ayudará a escribir y editar las reglas. Utilice el nodo de análisis de enlace de texto o ejecute una ruta con sesión interactiva con la extracción de TLA habilitada para obtener resultados para un conjunto de datos más completo. Esta simulación es solo para fines de prueba y creación de reglas.

Definición de datos para la simulación

Para ayudarle a ver cómo las reglas podrían hacer coincidir el texto, puede ejecutar una simulación mediante la utilización de datos de ejemplo. El primer paso es definir los datos.

Definición de datos

1. Pulse **Definir datos** en el panel de simulaciones en la parte inferior del separador **Reglas de enlace de texto**. De forma alternativa, si no se han definido datos previamente, seleccione **Herramientas > Ejecutar simulación** desde los menús. Se abrirá el asistente Datos de simulación.
2. Especifique el tipo de datos mediante la selección de una de las siguientes acciones:
 - **Pegar o ingresar texto directamente** Se proporciona un cuadro de texto para que pegue texto del portapapeles o que ingrese manualmente el texto que desee para ser procesado. Puede ingresar una oración por línea, o utilizar puntuación, como puntos o comas, para dividir la oración. Una vez que haya ingresado el texto, puede comenzar la simulación pulsando **Ejecutar simulación**.
 - **Especificar un origen de datos de archivo** Esta opción indica que desea procesar un archivo que contiene texto. Pulse **Siguiente** para proceder con el asistente en el cual puede definir el archivo que procesará. Una vez que el archivo ha sido seleccionado, puede comenzar la simulación pulsando **Ejecutar simulación**. Están soportados los siguientes tipos de archivo: `.txt` y `.text`. El archivo de datos que seleccionó se lee como "tal cual" durante la simulación. Todo el archivo se trata de la misma manera como si hubiera conectado un nodo Lista de archivos a un nodo Minería de textos.

Importante: Recomendamos encarecidamente que si utiliza un archivo de datos, se asegure de que el texto que contiene sea corto para poder minimizar el tiempo de proceso. El objetivo de la simulación es ver cómo un fragmento de texto se interpreta y entender cómo las reglas coinciden con este texto. Esta información le ayudará a escribir y editar las reglas. Utilice el nodo de análisis de enlace de texto o ejecute una ruta con sesión interactiva con la extracción de TLA habilitada para obtener resultados para un conjunto de datos más completo. Esta simulación es solo para fines de prueba y creación de reglas.

3. Para iniciar el proceso de simulación, pulse **Ejecutar simulación**. Aparecerá un diálogo de progreso. Si está en una sesión interactiva, los valores de extracción utilizados durante la simulación son los que están actualmente seleccionados en la sesión interactiva (consulte **Herramientas > Valores de extracción** en la vista Conceptos y categorías). Si se encuentra en Editor de plantillas, los valores de extracción utilizados durante la simulación son los valores de extracción predeterminados, que son los mismos que se muestran en el separador Experto de un nodo de Análisis de enlace de textos. Para obtener más información, consulte [“Descripción de los resultados de simulación” en la página 214](#).

Descripción de los resultados de simulación

Para ayudarle a ver de qué forma las reglas hacen coincidir textos, puede ejecutar una simulación mediante la utilización de datos de muestra y revisar los resultados. Desde allí, puede modificar el conjunto de reglas para que se ajusten mejor a sus datos. Cuando el proceso de simulación y extracción se ha completado, se le mostrarán los resultados de la simulación.

Para cada "oración" identificada durante la extracción, se le presentarán varias partes de la información incluida la "oración" exacta, el desglose de los elementos encontrados en la oración de texto de salida, y, finalmente, cualquier regla que coincida con el texto en dicha oración. Con "**oración**", nos referimos a una palabra, frase o cláusula, dependiendo de cómo el extractor dividió el texto en trozos legibles.

Un **elemento** se define como cualquier palabra o frase de palabras identificadas durante el proceso de extracción. Por ejemplo, en la frase *Mi tío vive en Nueva York*, podrían encontrarse los elementos siguientes durante la extracción: *mi*, *tío*, *vive*, *en*, y *nueva york*. Además, *tío* podría extraerse como un concepto y escrito como <Unknown>, y *nueva york* también podría extraerse como concepto y escribirse como <Location>. Todos los conceptos son elementos pero no todos los elementos son conceptos. También pueden ser elementos otros macros, cadenas literales y espacios entre palabras. Solo aquellas palabras o frases de palabras que se escriben pueden ser conceptos.

Cuando trabaja en el editor de recursos o sesión interactiva, está trabajando a nivel de concepto. Las reglas TLA son más granulares, y los elementos individuales de una oración pueden utilizarse en la definición de una regla incluso si nunca se extraen o se escriben. Utilizar elementos, los cuales no son conceptos, le ofrece más flexibilidad a las reglas para capturar relaciones complejas en el texto.

Si tiene más de una oración en los datos de simulación, puede desplazarse hacia delante y hacia atrás en los resultados si pulsa **Siguiente** y **Previo**.

En los casos donde una oración no coincida con ninguna regla TLA en la biblioteca seleccionada (consulte el nombre de la biblioteca sobre el árbol en este separador), los resultados se consideran como no coincidentes y se habilitan los botones **Siguiente sin coincidir** y **Previa sin coincidir** para hacerle saber que las reglas no han encontrado una coincidencia de un texto, y para permitirle navegar a dichas instancias rápidamente.

Después de crear nuevas reglas, editarlas o modificar los recursos o los valores de extracción quizás desee volver a ejecutar una simulación. Para volver a ejecutar una simulación, pulse **Ejecutar simulación** en el panel de simulaciones y los mismos datos de entrada se utilizarán nuevamente.

Los siguientes campos y tablas se muestran en los resultados de simulación:

Texto de entrada. La "oración" real identificada mediante el proceso de extracción desde los datos de simulación que definió en el asistente. Por oración, nos referimos a una palabra, frase o cláusula, dependiendo de cómo el extractor dividió el texto en trozos legibles.

Vista de sistema. Una colección de elementos que el proceso de extracción ha identificado.

- **Elemento de texto de entrada.** Cada elemento encontrado en el texto de entrada. Los elementos fueron definidos anteriormente en este tema.
- **Escrito como.** Si un elemento ha sido identificado como un concepto y se ha escrito, entonces el nombre asociado (tal como <Unknown>, <Person>, <Location>) se muestra en esta columna.
- **Macro coincidente.** Si un elemento coincidió con una macro existente, entonces el nombre de macro asociado se muestra en esta columna.

Reglas coincidentes con texto de entrada. Esta tabla muestra cualquier regla TLA que coincidió con el texto de entrada. Por cada regla coincidente, verá el nombre de la regla en la columna **Regla de salida** y los valores de salida asociados de la regla (pares Concepto + Tipo). Puede efectuar una doble pulsación en el nombre de regla coincidente en el panel de editor que se encuentra sobre el panel de simulaciones.

Botón Crear regla. Si pulsa este botón en el panel de simulaciones, se abrirá una nueva regla en el panel de editor que se encuentra sobre el panel de simulaciones. Se tomará el texto de entrada como ejemplo. Del mismo modo, cualquier elemento que se haya escrito o que haya coincidido con una macro durante la simulación se insertará de forma automática en la columna Elementos en la **tabla Valores de regla**. Si un elemento se escribió y coincidió con una macro, el valor de ésta es el que se utilizará en la regla para simplificarla. Por ejemplo, la oración “*Me gusta la pizza*” podría escribirse durante la simulación como <Unknown> y coincidir con la macro mTopic si utilizó los recursos de inglés básico. En este caso mTopic se utilizará como el elemento en la regla creada. Para obtener más información, consulte el tema “[Cómo trabajar con reglas de enlace de texto](#)” en la página 219.

Navegación por reglas y macros en el árbol

Cuando se realice el análisis de enlace de texto durante la extracción, se utilizarán las reglas de enlace de texto almacenadas en la biblioteca seleccionada en el separador **Reglas de enlace de texto**.

A diferencia de otros recursos avanzados, las reglas TLA pertenecen a bibliotecas específicas, por lo tanto, solo puede utilizarlas una biblioteca a la vez. Desde el Editor de plantillas o Editor de recursos, vaya al separador **Reglas de enlace de texto**. En este separador puede especificar la biblioteca en la plantilla que contiene las reglas TLA que desea editar o utilizar. Por este motivo, recomendamos encarecidamente que almacene todas las reglas en una biblioteca, a menos que exista una razón de peso o específica por la que no se desee esto.

Puede especificar en qué biblioteca desea trabajar en el separador Reglas de enlace de texto seleccionando la biblioteca en la lista desplegable **Utilizar y almacenar reglas de análisis de enlace de texto en:** en este separador. Cuando se realice el análisis de enlace de texto durante la extracción, se utilizarán las reglas de enlace de texto almacenadas en la biblioteca seleccionada en el separador **Reglas de enlace de texto**. Por lo tanto, si ha definido las reglas de enlace de texto (reglas de TLA) en más de una biblioteca, solo la primera biblioteca en la que se encuentran las reglas TLA se utilizará para el análisis de enlace de texto. Por este motivo, recomendamos encarecidamente que almacene todas las reglas en una biblioteca, a menos que exista una razón muy específica por la que no se desee esto.

Cuando selecciona una macro o regla en el árbol, su contenido se visualiza en el panel del editor a la derecha. Si pulsa con el botón derecho del ratón en cualquier elemento del árbol, se abrirá un menú contextual para mostrar las otras tareas posibles, como por ejemplo:

- Crear una nueva macro en el árbol y abrirla en el editor a la derecha.
- Crear una nueva regla en el árbol y abrirla en el editor a la derecha.
- Crear un nuevo conjunto de reglas en el árbol.
- Cortar, copiar y pegar elementos para simplificar la edición.
- Suprimir macros, reglas y conjuntos de reglas para eliminarlos de los recursos.
- Inhabilitar macros, reglas y conjuntos de reglas para indicar que se deben ignorar durante el proceso.
- Mover las reglas hacia arriba o abajo para influir en el orden de proceso.

Avisos en el árbol

Los avisos se visualizan con un triángulo amarillo en el árbol y están ahí para informarle de que puede haber un problema. Sitúe el puntero del ratón sobre la regla o macro defectuosa para visualizar una

descripción emergente. En la mayoría de los casos, verá algo como: **Aviso: no se ha proporcionado ningún ejemplo; especifique un ejemplo**, por lo tanto debe especificar un ejemplo.

Si le falta un ejemplo o si el ejemplo no coincide con la regla, no podrá utilizar la función Obtener señales, por lo que le recomendamos que especifique solo un ejemplo por regla.

Cuando la regla aparece resaltada en amarillo significa que un tipo o macro es desconocida para el editor de TLA. El mensaje será parecido a: **Aviso: tipo o macro desconocida**. Esto es para informarle de que un elemento que sería definido por \$something en la vista de origen, por ejemplo \$myType, no es un tipo heredado en la biblioteca ni tampoco una macro.

Para actualizar el comprobador de sintaxis debe conmutar a otra regla o macro; no es necesario recompilar nada. Así, por ejemplo, si la regla A muestra un aviso porque falta el ejemplo, debe añadir un ejemplo, pulsar una regla superior o inferior y después volver a la regla A para comprobar que ahora esté correcto.

Cómo trabajar con macros

Las macros pueden simplificar el aspecto de las reglas de análisis de enlaces de texto, lo cual le permite agrupar tipos, otras macros y series literales (palabra) junto con un operador OR (|). La ventaja de utilizar macros es que no solo puede reutilizar macros en varias reglas de análisis de enlaces de texto para simplificarlas, sino que también le permite realizar actualizaciones en una macro en lugar de tener que realizar actualizaciones en todas las reglas de análisis de enlaces de texto. La mayoría de las reglas de TLA suministradas contienen macros predefinidas. Las macros aparecen en la parte superior del árbol en el panel que está más a la izquierda del separador Reglas de enlace de texto.

Los siguientes campos y tablas se muestran en los resultados de simulación:

Nombre. Un nombre exclusivo que identifica esta macro. Es recomendable que le ponga una m minúscula como prefijo a los nombres de macros para ayudarlo a identificar las macros rápidamente en las reglas. Cuando hace referencia manualmente a las macros en las reglas (mediante la edición en línea o en la vista de origen) tiene que utilizar el prefijo de carácter \$, de modo que el proceso de extracción sepa buscar este nombre especial. Sin embargo, si arrastra y suelta el nombre de la macro o lo añade mediante los menús contextuales, el producto lo reconocerá automáticamente como una macro y no se añadirá ningún \$.

Tabla **Valor de macro**.

- Un número de filas que representa todos los valores posibles que esta macro puede representar. Estos valores distinguen entre mayúsculas y minúsculas.
- Estos valores pueden incluir uno o una combinación de tipos, series literales, espacios entre palabras o macros. Para obtener más información, consulte el tema [“Elementos soportados para reglas y macros” en la página 224](#).
- Para especificar un valor para un elemento de una macro, efectúe una doble pulsación en la fila en la que desea trabajar. Aparece un recuadro de texto editable en el que puede especificar una referencia de tipo, una referencia de macro, una serie literal o un espacio entre palabras. De forma alternativa, pulse con el botón derecho del ratón en la celda para visualizar un menú contextual ofreciendo listas de macros comunes, nombres de tipo y nombres de tipo no lingüísticos. Para hacer referencia a un tipo o una macro debe anteponer un carácter ‘\$’ al nombre de tipo o macro, como por ejemplo, \$mTopíc para la macro mTopíc. Cuando se combinan argumentos, debe utilizar paréntesis () para agruparlos, y el carácter | para indicar un OR booleano.
- Puede añadir o eliminar filas de la tabla Valor de macro utilizando los botones a su derecha.
- Especifique cada elemento en su propia fila. Por ejemplo, si desea crear una macro que represente una de 3 series literales como am OR was OR is, debe especificar cada serie literal en una fila separada en la vista, y la tabla Macro contendrá 3 filas.

Creación y edición de macros

Puede crear nuevas macros o editar las existentes. Siga las directrices y descripciones para el editor de macros. Para obtener más información, consulte el tema [“Cómo trabajar con macros” en la página 216](#).

Creación de nuevas macros

1. En los menús, seleccione **Herramientas > Nueva macro**. Como alternativa, pulse el icono Nueva macro en la barra de herramientas del árbol para abrir una nueva macro en el editor.
2. Especifique un nombre exclusivo y defina los elementos de valor de la macro.
3. Pulse **Aplicar** cuando haya terminado para comprobar si hay errores.

Edición de macros

1. Pulse el nombre de la macro en el árbol. La macro se abrirá en el panel del editor a la derecha.
2. Efectúe los cambios que desee.
3. Pulse **Aplicar** cuando haya terminado para comprobar si hay errores.

Inhabilitación y supresión de macros

Inhabilitación de macros

Si desea que se ignore una macro durante el proceso, puede inhabilitarla. Al hacerlo puede generar avisos o errores en las reglas que aún hacen referencia a esta macro inhabilitada. Tenga cuidado al suprimir e inhabilitar macros.

1. Pulse el nombre de la macro en el árbol. La macro se abrirá en el panel del editor a la derecha.
2. Pulse con el botón derecho del ratón en el nombre.
3. En los menús contextuales, elija **Inhabilitar**. El icono de la macro se vuelve gris y la propia macro no se puede editar.

Supresión de macros

Si desea deshacerse de una macro, puede suprimirla. Al hacerlo puede causar errores en las reglas que aún hacen referencia a esta macro. Tenga cuidado al suprimir e inhabilitar macros.

1. Pulse el nombre de la macro en el árbol. La macro se abrirá en el panel del editor a la derecha.
2. Pulse con el botón derecho del ratón en el nombre.
3. En los menús contextuales, elija **Suprimir**. La macro desaparecerá de la lista.

Comprobación de errores, guardado y cancelación

Aplicación de cambios de macro

Si pulsa fuera del editor de macros o si pulsa **Aplicar**, se explora automáticamente la macro en busca de errores. Si se encuentra un error, deberá arreglarlo antes de pasar a otra parte de la aplicación.

Sin embargo, si se detectan errores menos graves, solo se emite un aviso. Por ejemplo, si la regla contiene definiciones incompletas o no referenciadas a tipos u otros macros, se muestra un mensaje de aviso. Una vez que pulsa **Aplicar**, los avisos sin corregir provocan que aparezca un icono de aviso a la izquierda del nombre de macro en el Árbol de macros y reglas en el panel izquierdo.

Aplicar una macro no significa que la macro se guarda permanentemente. La aplicación provocará que el proceso de validación busque errores y avisos.

Cómo guardar recursos en una sesión de área de trabajo interactiva

1. Para guardar los cambios que ha realizado a los recursos durante una sesión de área de trabajo interactiva y poder obtenerlos la próxima vez que ejecute la ruta, debe:
 - Actualizar el nodo de modelado para asegurarse de que puede obtener estos mismos recursos la próxima vez que ejecute la ruta. Para obtener más información, consulte el tema [“Guardar y actualizar nodos de modelado”](#) en la página 77. A continuación, guarde la ruta. Para guardar la ruta, hágalo en la ventana principal de IBM SPSS Modeler después de actualizar el nodo de modelado.
2. Para guardar los cambios que ha realizado a los recursos durante una sesión de área de trabajo interactiva y poder utilizarlos en otras rutas, puede:

- Actualizar la plantilla que ha utilizado o hacer una nueva. Para obtener más información, consulte el tema [“Creación y actualización de plantillas”](#) en la página 166. Esto no guardará los cambios para el nodo actual (vea el paso anterior)
- O bien, actualizar el TAP que ha utilizado. Para obtener más información, consulte el tema [“Actualización de los paquetes de análisis de texto”](#) en la página 137.

Cómo guardar recursos en el Editor de plantillas

1. En primer lugar, publique la biblioteca. Para obtener más información, consulte el tema [“Publicación de bibliotecas”](#) en la página 185.
2. A continuación, guarde la plantilla a través de **Archivo > Guardar plantilla de recurso** en los menús.

Cómo cancelar cambios de macro

1. Si desea descartar los cambios, pulse **Cancelar**.

Macros especiales: mTopic, mNonLingEntities, SEP

La plantilla Opiniones (y plantillas similares), así como las plantillas Recursos básicos se suministran con dos macros especiales denominadas mTopic y mNonLingEntities.

mTopic

De forma predeterminada, la macro mTopic agrupa todos los tipos suministrados en la plantilla que son probable que estén conectados con una opinión, como los siguientes tipos de biblioteca *principales*: <Person>, <Organization>, <Location>, etc., siempre que el tipo no sea un tipo de opinión (por ejemplo, <Negative> o <Positive>) o un tipo definido como una entidad no lingüística en los Recursos avanzados.

Siempre que se crea un nuevo tipo en una plantilla Opiniones (o similar), el producto presupone que a menos que este tipo esté especificado en otra macro o en la sección de entidades no lingüísticas del separador Recursos avanzados, se tratará de la misma forma que a los demás tipos definidos en la macro mTopic.

Digamos que ha creado nuevos tipos en los recursos desde una plantilla Opiniones: <Vegetables> y <Fruit>. Sin necesidad de realizar cambios, los nuevos tipos se tratan como tipos mTopic de modo que puede descubrir automáticamente las opiniones positivas, negativas, neutrales y contextuales sobre los nuevos tipos. Durante la extracción, por ejemplo, la frase *"Me encanta el brócoli, pero detesto el pomelo"* generaría los dos patrones de resultados siguientes:

broccoli <Vegetables> + like <Positive>

grapefruit <Fruit> + dislike <Negative>

Sin embargo, si desea procesar esos tipos de forma diferente a los otros tipos en mTopic, puede añadir el nombre de tipo a una macro existente como mPos, que agrupa todos los tipos de opinión positiva, o crear una nueva macro a la que más adelante puede hacerse referencia en uno o más reglas.

Importante: Si crea un nuevo tipo como por ejemplo <Vegetables>, este nuevo tipo se incluirá como un tipo en mTopic, sin embargo, este nombre de tipo no se verá de forma explícita en la definición de la macro.

mNonLingEntities

De forma similar, si añade nuevas entidades no lingüísticas en la sección **Entidades no lingüísticas** del separador Recursos avanzados, se procesarán automáticamente como mNonLingEntities a menos que se especifique lo contrario. Para obtener más información, consulte el tema [“Entidades no lingüísticas”](#) en la página 202.

SEP

También puede utilizar la macro predefinida SEP, que corresponde al separador global definido en la máquina local, generalmente una coma (,).

Cómo trabajar con reglas de enlace de texto

Una regla de análisis de enlace de texto es una consulta booleana que se utiliza para realizar una comparación en una oración. Las reglas de análisis de enlace de texto contienen uno o más de los argumentos siguientes: tipos, macros, series literales o espacios entre palabras. Debe tener al menos una regla de análisis de enlace de texto para poder extraer los resultados de TLA.

Las siguientes áreas y campos se muestran en el separador Reglas de enlace de texto, Editor de reglas:

Campo **Nombre**. El nombre exclusivo para la regla de enlace de texto.

Campo **Ejemplo**. Opcionalmente, puede incluir una oración de ejemplo o secuencia de palabras que sería capturada por esta regla. Se recomienda utilizar ejemplos. En este editor, podrá generar señales de este texto de ejemplo para ver cómo coincide con la regla y cómo se emitirá. Un **elemento** se define como cualquier palabra o frase de palabras identificadas durante el proceso de extracción. Por ejemplo, en la frase *Mi tío vive en Nueva York*, podrían encontrarse los elementos siguientes durante la extracción: *mi*, *tío*, *vive*, *en*, y *nueva york*. Además, *tío* podría extraerse como un concepto y escrito como <Unknown>, y *nueva york* también podría extraerse como concepto y escribirse como <Location>. Todos los conceptos son elementos pero no todos los elementos son conceptos. También pueden ser elementos otros macros, cadenas literales y espacios entre palabras. Solo aquellas palabras o frases de palabras que se escriben pueden ser conceptos.

Tabla Valor de regla. Esta tabla contiene los elementos de la regla que se utilizan para la coincidencia de una regla con una oración. Puede añadir o eliminar filas de la tabla utilizando los botones a su derecha. La tabla consta de 3 columnas:

- Columna **Elemento**. Especifique los valores como uno o una combinación de tipos, series literales, espacios entre palabras (<Any Token>) o macros. Para obtener más información, consulte el tema “Elementos soportados para reglas y macros” en la página 224. Efectúe una doble pulsación en la celda del elemento para especificar la información directamente. De forma alternativa, pulse con el botón derecho del ratón en la celda para visualizar un menú contextual ofreciendo listas de macros comunes, nombres de tipo y nombres de tipo no lingüísticos. Tenga en cuenta que si especifica la información en la celda escribiéndola, debe anteponer un carácter ‘\$’ al nombre de tipo o macro, como por ejemplo, \$mTopic para la macro mTopic. El orden en el que crea las filas de elementos es crucial para ver cómo se comparará la regla con el texto. Cuando se combinan argumentos, debe utilizar paréntesis () para agruparlos, y el carácter | para indicar un OR booleano. Tenga en cuenta que los valores distinguen entre mayúsculas y minúsculas.
- Columna **Cantidad**. Esto indica el número mínimo y máximo de veces que se debe encontrar el elemento para que se produzca una coincidencia. Por ejemplo, si desea definir un espacio o una serie de palabras, entre otros dos elementos de cualquier lugar de 0 a 3 palabras, puede elegir **Entre 0 y 3** de la lista o especificar los números directamente en el recuadro de diálogo. El valor predeterminado es ‘Exactamente 1’. En algunos casos le interesará hacer que un elemento sea opcional. Si este es el caso, tendrá una cantidad mínima de 0 y una cantidad máxima mayor que 0 (es decir, 0 ó 1, entre 0 y 2). Tenga en cuenta que el primer elemento de una regla no puede ser opcional, lo que significa que no puede tener una cantidad de 0.
- Columna **Señal de ejemplo**. Si pulsa **Obtener señales**, el programa desglosa el texto **Ejemplo** en señales y las utiliza para llenar esta columna con aquellas que coincidan con los elementos que ha definido. También puede ver estas señales en la tabla de salida si lo desea.

Tabla de salida de regla Cada fila de esta tabla define cómo aparecerá el resultado del patrón TLA en los resultados. La salida de regla puede producir patrones de hasta seis pares de columna Concepto/Tipo, cada una de las que representa un *intervalo*. Por ejemplo, el patrón de tipo <Location> + <Positive> es un patrón de dos ranuras que significa que está formado por 2 pares de columnas Concepto/Tipo.

Nota: Los términos de la columna **Elemento** de la tabla **Valores de regla**, o en cualquiera de las columnas **Concepto** de la tabla **Salida de reglas** no pueden empezar por ninguno de los siguientes caracteres: ` , #, %, ^, *, _ , - , : , < , > , / , \ o " .

Así como el lenguaje nos da la libertad de expresar las mismas ideas básicas de muchas maneras distintas, se puede tener un número de reglas definidas para capturar la misma idea básica. Por ejemplo, el texto “*París es un lugar que me encanta*” y el texto “*Me gusta muchísimo París y Florencia*” representan

la misma idea básica, que le gusta París, pero se expresan de forma distinta y se requerirían dos reglas diferentes para capturar ambos. Sin embargo, es más fácil trabajar con los resultados del patrón si se agrupan las ideas similares. Por esta razón, aunque puede tener 2 reglas diferentes para capturar estas 2 frases, puede definir la misma salida para ambas reglas, como por ejemplo, el tipo de patrón <Location> + <Positive> de modo que represente ambos textos. Y de esta forma, puede ver que la salida no siempre imita la estructura u orden de las palabras encontradas en el texto original. Además, este tipo de patrón podría coincidir con otras frases y podría generar patrones de conceptos como: paris + like y tokyo + like.

Para ayudarle a definir la salida rápidamente con menos errores, puede utilizar el menú contextual para elegir el elemento que desea ver en la salida. Como alternativa, también puede arrastrar y soltar elementos de la tabla Valor de regla en la salida. Por ejemplo, si tiene una regla que contiene una referencia a la macro mTopic en la fila 2 de la tabla Valor de regla y desea que ese valor esté en la salida, puede simplemente arrastrar/soltar el elemento para mTopic en el primer par de columna en la tabla Salida de regla. De este modo se llenarán automáticamente el Concepto y Tipo para el par que ha seleccionado. O bien, si desea que la salida comience con el tipo definido por el tercer elemento (fila 3) de la tabla de valor de regla, arrastre ese tipo de la tabla Valor de regla a la celda **Tipo 1** de la tabla de salida. La tabla se actualizará para mostrar la referencia de la fila en paréntesis (3).

De forma alternativa, puede especificar estas referencias manualmente en la tabla mediante una doble pulsación en la celda de cada columna **Concepto** a la que desea dar salida y especificando el símbolo \$ seguido del número de fila, como por ejemplo \$2 para hacer referencia al elemento definido en la fila 2 de la tabla Valor de regla. Cuando especifica la información manualmente, también debe definir la columna **Tipo**, especifique el símbolo # seguido del número de fila, como por ejemplo #2 para hacer referencia al elemento definido en la fila 2 de la tabla Valor de regla.

Además, se puede incluso combinar los métodos. Supongamos que tenía el tipo <Positive> en la fila 4 de la tabla Valor de regla. Podría arrastrarlo a la columna Type 2 y, a continuación, hacer doble clic en la celda de la columna Concept 2 y, a continuación, especificar manualmente la palabra 'no' delante de ella. A continuación, la columna de salida leerá not (4) en la tabla o si estaba en la modalidad de edición o en la modalidad de origen not \$4. A continuación, podría pulsar con el botón derecho del ratón en la columna Tipo 1 y seleccionar, por ejemplo, la macro denominada mTopic. Esta salida podría generar un patrón de concepto como: car + bad.

La mayoría de las reglas solo tienen una fila de salida, pero a veces es posible y se desea que haya más de una salida. En este caso, defina una salida por fila en la tabla Salida de regla.

Importante: Tenga en cuenta que otras operaciones lingüísticas se realizan durante la extracción del patrón TLA. Entonces, cuando la salida lea t\$3\t#3, significa que el patrón mostrará el concepto final del tercer elemento y el tipo final del tercer elemento después que se hayan aplicado todos los procesos lingüísticos (sinónimos y otras agrupaciones).

- **Mostrar salida como.** De forma predeterminada, se selecciona la opción **Referencias a la fila en la tabla Valor de regla** y se muestra la salida mediante las referencias numéricas a la fila tal como se define en la tabla Valor de regla. Si ha pulsado anteriormente Obtener señales y tiene señales en la columna Señales de ejemplo en la tabla Valor de regla, puede optar por ver la salida para estas señales específicas seleccionando la opción.

Nota: Si no se muestran suficientes pares de salida de concepto/tipo en la tabla de salida, puede añadir otro par pulsando el botón Añadir en la barra de herramientas del editor. Si actualmente se muestran 3 pares y pulsa añadir, se añaden 2 columnas más (Concepto 4 y Tipo 4) a la tabla. Esto significa que ahora verá 4 pares en la tabla de salida para todas las reglas. También puede eliminar los pares no utilizados siempre y cuando ninguna otra regla en el conjunto de reglas de esta biblioteca utilice ese par.

Regla de ejemplo

Supongamos que los recursos contienen la siguiente regla de análisis de enlace de texto y que ha habilitado la extracción de resultados de TLA:

Output columns: Add Remove View Source

Name:

Example:

Rule Value table:

	Element	Quantity	Example Token
1	mSupportNeg	Exactly 1	isn't
2		0 or 1	
3	(anything ((any a one) thing ?))	Exactly 1	anything
4		Between 0 and 2	that i
5	mNeg	Exactly 1	disliked
6	(about with in)	Exactly 1	about
7		0 or 1	
8	mDet	0 or 1	the

Get Tokens
Insert Row
Remove Row

Rule Output table:

Concept 1	Type 1	Concept 2	Type 2	Concept 3	Type 3
product (9)	Products (9)	no dislike (5)	Positive		

Show output as: ☐ References to row in Rule Value table ☒ Specific token from example

Apply Cancel

Figura 45. Separador Reglas de enlace de texto: Editor de reglas

Al realizar la extracción, el motor de extracción leerá cada frase e intentará la coincidencia de la secuencia siguiente:

Tabla 46. Ejemplo de secuencia de extracción	
Elemento (fila)	Descripción de los argumentos
1	El concepto de uno de los tipos representados por las macros mPos o mNeg o del tipo <Uncertain>.
2	Un concepto escrito como uno de los tipos representados por la macro mTopic.
3	Una de las palabras representadas por la macro mBe.
4	Un elemento opcional, 0 o 1 palabras, también denominado espacio entre palabras o <Any Token>
5	Un concepto escrito como uno de los tipos representados por la macro mTopic.

La tabla de salida muestra que todo lo que se desea de esta regla es un patrón donde cualquier concepto o tipo correspondiente a la macro mTopic definida en la fila 5 de la **tabla de valor de regla** + cualquier concepto o tipo correspondiente a mPos, mNeg o <Uncertain>, tal como se ha definido en la fila 1 de la **tabla de valor de regla**. Esto puede ser sausage + like o <Unknown> + <Positive>.

Creación y edición de reglas

Puede crear nuevas reglas o editar las existentes. Siga las directrices y descripciones para el editor de reglas. Para obtener más información, consulte el tema [“Cómo trabajar con reglas de enlace de texto”](#) en la página 219.

Creación de nuevas reglas

1. En los menús, seleccione **Herramientas > Nueva regla**. Como alternativa, pulse el icono Nueva regla en la barra de herramientas del árbol para abrir una nueva regla en el editor.

2. Especifique un nombre exclusivo y defina los elementos de valor de regla.
3. Pulse **Aplicar** cuando haya terminado para comprobar si hay errores.

Edición de reglas

1. Pulse el nombre de regla en el árbol. La regla se abrirá en el panel del editor a la derecha.
2. Efectúe los cambios que desee.
3. Pulse **Aplicar** cuando haya terminado para comprobar si hay errores.

Inhabilitación y supresión de reglas

Inhabilitación de reglas

Si desea que se ignore una regla durante el proceso, puede inhabilitarla. Tenga cuidado al suprimir e inhabilitar reglas.

1. Pulse el nombre de regla en el árbol. La regla se abrirá en el panel del editor a la derecha.
2. Pulse con el botón derecho del ratón en el nombre.
3. En los menús contextuales, elija **Inhabilitar**. El icono de la regla se vuelve gris y la propia regla no se puede editar.

Eliminación de reglas

Si desea deshacerse de una regla, puede suprimirla. Tenga cuidado al suprimir e inhabilitar reglas.

1. Pulse el nombre de regla en el árbol. La regla se abrirá en el panel del editor a la derecha.
2. Pulse con el botón derecho del ratón en el nombre.
3. En los menús contextuales, elija **Suprimir**. La regla desaparecerá de la lista.

Comprobación de errores, guardado y cancelación

Aplicación de cambios de regla

Si pulsa fuera del editor de reglas o si pulsa **Aplicar**, se explora automáticamente la regla en busca de errores. Si se encuentra un error, deberá arreglarlo antes de pasar a otra parte de la aplicación.

Sin embargo, si se detectan errores menos graves, solo se emite un aviso. Por ejemplo, si la regla contiene definiciones incompletas o no referenciadas a tipos o macros, se muestra un mensaje de aviso. Una vez que pulsa **Aplicar**, los avisos sin corregir provocan que aparezca un icono de aviso a la izquierda del nombre de regla en el árbol en el panel izquierdo.

Aplicar una regla no significa que la regla se guarda permanentemente. La aplicación provocará que el proceso de validación busque errores y avisos.

Cómo guardar recursos en una sesión de área de trabajo interactiva

1. Para guardar los cambios que ha realizado a los recursos durante una sesión de área de trabajo interactiva y poder obtenerlos la próxima vez que ejecute la ruta, debe:
 - Actualizar el nodo de modelado para asegurarse de que puede obtener estos mismos recursos la próxima vez que ejecute la ruta. Para obtener más información, consulte el tema [“Guardar y actualizar nodos de modelado”](#) en la [página 77](#). A continuación, guarde la ruta. Para guardar la ruta, hágalo en la ventana principal de IBM SPSS Modeler después de actualizar el nodo de modelado.
2. Para guardar los cambios que ha realizado a los recursos durante una sesión de área de trabajo interactiva y poder utilizarlos en otras rutas, puede:
 - Actualizar la plantilla que ha utilizado o hacer una nueva. Para obtener más información, consulte el tema [“Creación y actualización de plantillas”](#) en la [página 166](#). Esto no guardará los cambios para el nodo actual (vea el paso anterior)
 - O bien, actualizar el TAP que ha utilizado. Para obtener más información, consulte el tema [“Actualización de los paquetes de análisis de texto”](#) en la [página 137](#).

Cómo guardar recursos en el Editor de plantillas

1. En primer lugar, publique la biblioteca. Para obtener más información, consulte el tema [“Publicación de bibliotecas”](#) en la página 185.
2. A continuación, guarde la plantilla a través de **Archivo > Guardar plantilla de recurso** en los menús.

Cómo cancelar cambios de regla

1. Si desea descartar los cambios, pulse **Cancelar** en el panel del editor.

Orden de proceso para reglas

Al realizar un análisis de enlace de texto durante la extracción, se comparará una "oración" (oración, palabra, frase) con cada regla, una por una, hasta que se encuentre una coincidencia o se hayan agotado todas las reglas. La posición en el árbol dicta el orden en que se prueban las reglas. El método recomendado indica que se deberían ordenar las reglas de la más específica a la más genérica. Las más específicas deberían estar en la parte superior del árbol. Para cambiar el orden de una regla específica o conjunto de reglas, seleccione **Subir** o **Bajar** en el menú contextual Árbol de macro y reglas o las flechas hacia arriba y hacia abajo en la barra de herramientas.

Si está en la *vista de origen*, no puede cambiar el orden de las reglas desplazándolas en el editor. Cuanto más arriba aparece la regla en la vista de origen, antes se procesa. Recomendamos encarecidamente la reordenación de las reglas solo en el árbol, para evitar problemas de copiar/pegar.

Importante: En versiones anteriores de IBM SPSS Modeler Text Analytics, se le solicitaba tener un ID de regla numérico y único. A partir de la versión 18.4.0, solo puede indicar el orden de proceso desplazando una regla hacia arriba o abajo en el árbol o por su posición en la vista de origen.

Por ejemplo, suponga que el texto contiene las dos frases siguientes:

Me encantan las anchoas

Me encantan las anchoas y los pimientos verdes

Además, suponga que existen dos reglas de análisis de enlace de texto con los valores siguientes:

A			
	Element	Quantity	Example Token
1	Positive	Exactly 1	
2	mDet	0 or 1	
3	mTopic	Exactly 1	
4			
5			
6			
7			

B			
	Element	Quantity	Example Token
1	Positive	Exactly 1	
2	mDet	0 or 1	
3	mTopic	Exactly 1	
4	(SEP and or)	1 or 2	
5	mDet	0 or 1	
6	mTopic	Exactly 1	
7			

Figura 46. 2 reglas de ejemplo

En la vista de origen, los valores de regla pueden ser parecidos a lo siguiente:

A: value = \$Positive \$mDet? \$mTopic

B: value = \$Positive \$mDet? \$mTopic (\$SEP|and|or){1,2} \$mDet? \$mTopic

Si la regla **A** está más arriba en el árbol (más cerca de la parte superior) que la regla **B**), la regla **A** se procesará en primer lugar y la sentencia *I love anchovies and green peppers* coincidirá primero con

\$Positive \$mDet? \$mTopic, y generará una salida de patrón incompleta (anchovies + like) ya que coincidía con una regla que no estaba buscando 2 coincidencias de \$mTopic.

Por lo tanto, para capturar la verdadera esencia del texto la regla más específica, en este caso **B**, debe colocarse más arriba en el árbol que la más genérica, en este caso la regla **A**.

Cómo trabajar con conjuntos de reglas (varios pasos)

Un conjunto de reglas es una forma útil de agrupar un conjunto de reglas relacionado en el Árbol de macros y reglas para realizar el proceso de varios pasos. Un conjunto de reglas no tiene definición propia que no sea un nombre, y se utiliza para organizar las reglas en grupos significativos. En algunos contextos, el texto es demasiado rico y variado para procesarse en una sola pasada. Por ejemplo, al trabajar con datos de inteligencia y seguridad, el texto puede contener enlaces entre individuos que se descubren mediante métodos de contacto (*x denominado y*), relaciones de familia (*x cuñado de y*), intercambio de dinero (*x le transfirió \$100 a y*), etc. En este caso, es útil crear conjuntos especializados de reglas de análisis de enlace de texto, cada uno de los cuales se centra en un determinado tipo de relación como una para descubrir contactos, otra para descubrir miembros de la familia, etc.

Para crear un conjunto de reglas, seleccione “Crear conjunto de reglas” del menú contextual del Árbol de macros y reglas o de la barra de herramientas. A continuación, cree nuevas reglas directamente bajo un nodo de conjunto de reglas en el árbol o mueva las reglas existentes a un conjunto de reglas.

Cuando ejecuta una extracción utilizando recursos en los que las reglas se agrupan en conjuntos de reglas, el buscador de extracción está obligado a realizar varios pasos a través del texto para poder comparar diferentes tipos de patrones en cada paso. De este modo, puede compararse una "oración" con una regla en cada conjunto de reglas, mientras que un conjunto de reglas solo puede compararse con una única regla.

Nota: Puede añadir hasta 512 reglas por conjunto de reglas.

Creación de nuevos conjuntos de reglas

1. En los menús, seleccione **Herramientas > Nuevo conjunto de reglas**. Como alternativa, pulse el icono Nuevo conjunto de reglas en la barra de herramientas del árbol. Aparece un conjunto de reglas en el árbol de reglas.
2. Añada nuevas reglas a este conjunto de reglas o mueva las reglas existentes al conjunto.

Inhabilitación de conjuntos de reglas

1. Pulse con el botón derecho del ratón en el nombre de conjuntos de reglas en el árbol.
2. En los menús contextuales, elija **Inhabilitar**. El icono del conjunto de reglas se vuelve gris y todas las reglas contenidas dentro de ese conjunto de reglas también se inhabilitan e ignoran durante el proceso.

Supresión de conjuntos de reglas

1. Pulse con el botón derecho del ratón en el nombre de conjuntos de reglas en el árbol.
2. En los menús contextuales, elija **Suprimir**. El conjunto de reglas y todas las reglas que contiene se suprimen de los recursos.

Elementos soportados para reglas y macros

Se aceptan los siguientes argumentos para los parámetros de valor en macros y reglas de análisis de enlace de texto:

Macros

Puede utilizar una macro directamente en una regla de análisis de enlace de texto o dentro de otra macro. Si está especificando el nombre de la macro a mano o desde la vista de origen (en lugar de seleccionar el nombre de la macro desde un menú contextual), asegúrese de poner un prefijo de carácter de signo de dólar (\$) al nombre, como por ejemplo \$mTopic. El nombre de la macro distingue entre mayúsculas y

minúsculas. Puede elegir entre cualquier macro definida en el separador Reglas de enlace de texto actual cuando selecciona macros a través de los menús contextuales.

Tipos

Puede utilizar un tipo directamente en una macro o regla de análisis de enlace de texto. Si está especificando el nombre de tipo a mano o en la vista de origen (en lugar de seleccionar el tipo desde un menú contextual), asegúrese de poner un prefijo de carácter de signo de dólar (\$) al nombre de tipo, como por ejemplo \$Person. El nombre de tipo distingue entre mayúsculas y minúsculas. Si utiliza los menús contextuales, puede elegir entre cualquier tipo del conjunto actual de recursos que se está utilizando.

Si hace referencia a un tipo no reconocido, recibirá un mensaje de aviso y la regla tendrá un icono de aviso en el Árbol de macros y reglas hasta que lo corrija.

Cadenas literales

Para incluir información que nunca se extrajo, puede definir una cadena literal que el buscador de extracción buscará. Todas las palabras o frases extraídas se han asignado a un tipo y por esta razón no pueden utilizarse en cadenas literales. Si utiliza una palabra que se ha extraído, se pasará por alto, incluso si su tipo es <Unknown>.

Una cadena literal puede estar compuesta por una o más palabras. Se aplican las reglas siguientes cuando se define una lista de cadenas literales:

- Especifique la lista de cadenas entre paréntesis como (su). Si hay una selección de cadenas literales, cada cadena debe ir separada por el operador OR, como por ejemplo (a|an|the) o (his|hers|its).
- Utilice palabras simples o compuestas.
- Separe cada palabra de la lista por el carácter |, que equivale al booleano OR.
- Escriba las formas singular y plural si desea que coincidan ambas. Las declinaciones no se generan automáticamente.
- Utilice solamente letras minúsculas.
- Para reutilizar cadenas literales, defínalas como una macro y después utilice dicha macro en las otras macros y reglas de análisis de enlace de texto.
- Si una cadena contiene puntos (puntos y aparte) o guiones, deberá incluirlos. Por ejemplo, para hacer coincidir a . k . a en el texto, escriba los puntos junto a las letras a . k . a como cadena literal.

Operador de exclusión

Utilice ! como operador de exclusión para detener cualquier expresión de la negación de ocupar una ranura determinada. Solo puede añadir un operador de exclusión a mano mediante la edición de celdas en línea (efectúe una doble pulsación en la celda de la tabla Valor de regla o la tabla Valor de macro) o en la vista de origen. Por ejemplo, si añade \$mTopic @{0,2} !(\$Positive) \$Budget a la regla de análisis de enlaces de texto, está buscando texto que contenga (1) un término asignado a cualquiera de los tipos de la macro mTopic, (2) un espacio entre palabras de cero a dos palabras de longitud, (3) ninguna instancia de un término asignado al tipo <Positive> y (4) un término asignado al tipo <Budget>. Es posible que esto capture "*los coches tienen un precio elevado*", pero ignoraría "*la tienda ofrece descuentos increíbles*".

Para utilizar este operador, debe especificar el signo de exclamación y los paréntesis manualmente en la celda del elemento efectuando una doble pulsación en la celda.

Espacios de palabras (<Any Token>)

Un espacio de palabras, también denominado <Any Token>, define un rango numérico de señales que pueden estar presentes entre dos elementos. Los espacios entre palabras son muy útiles cuando se hacen coincidir frases muy similares que solo difieren ligeramente por la presencia de elementos determinantes adicionales, frases preposicionales, adjetivos, etc.

Tabla 47. Ejemplo de los elementos en una tabla Valor de regla sin un espacio entre palabras	
#	Elemento
1	 Unknown
2	 mBeHave
3	 Positive

Nota: en la vista de origen, este valor se define como: \$Unknown \$mBeHave \$Positive

Este valor coincidirá con oraciones como "el *personal del hotel* fue agradable", donde *personal del hotel* pertenece al tipo <Unknown>, *fue* está en la macro mBeHave y *agradable* es <Positive>. Sin embargo, no coincidirá con "el *personal del hotel* fue muy agradable".

Tabla 48. Ejemplo de los elementos de una tabla Valor de regla con un espacio de palabras <Any Token>	
#	Elemento
1	 Unknown
2	 mBeHave
3	
4	 Positive

Nota: en la vista de origen, este valor se define como: \$Unknown \$mBeHave @{0,1} \$Positive

Si añade un espacio entre palabras al valor de regla, coincidirá con "el *personal del hotel* fue agradable" y "el *personal del hotel* fue muy agradable".

En la vista de origen o con la edición en línea, la sintaxis para un espacio entre palabras es @{#, #}, donde @ indica un espacio entre palabras y {#, #} define el mínimo y máximo de palabras permitidas entre el elemento anterior y el siguiente. Por ejemplo, @{1, 3} significa que se puede realizar una coincidencia entre los dos elementos definidos si hay al menos una palabra presente pero no más de tres palabras que aparezcan entre esos dos elementos. @{0, 3} significa que puede haber una coincidencia entre los dos elementos definidos si hay 0, 1, 2 o 3 palabras presentes, pero no más de tres.

Visualizar y trabajar en la modalidad de origen

Por cada regla y macro el editor TLA crea el código de origen subyacente que utiliza el Extractor para hacer coincidir y producir la salida TLA. Si prefiere trabajar con el código mismo, puede ver el código fuente y editarlo directamente pulsando el botón "Ver origen" en la parte superior del Editor. La vista Origen mostrará y resaltará la regla o macro actualmente seleccionada. Sin embargo, recomendamos utilizar los paneles de edición para reducir las probabilidades de que se produzcan errores.

Cuando haya terminado de ver o editar el origen, pulse **Salir de origen**. Si crea una sintaxis no válida para una regla, se le solicitará que lo solucione antes de salir de la vista de origen.

Importante: Si edita en la vista de origen, recomendamos que edite las reglas y macros una a la vez. Después de editar una macro, valide los resultados mediante la extracción. Si está conforme con el

resultado, recomendamos que guarde la plantilla antes de realizar otro cambio. Si no está conforme con el resultado o si se han producido errores, restituya los recursos guardados.

Macros en la vista Origen

```
[macro]
name = macro_name
value = ([type_name|macro_name|literal_string|word_gap])
```

Tabla 49. Entradas de macro

[macro]	Cada macro debe comenzar con la línea [macro] marcada para indicar el comienzo de una macro.
name	El nombre de la definición de macros. Todos los nombres deben ser exclusivos.
value	Una combinación de uno o más tipos, series literales, espacios entre palabras o macros. Consulte el tema “Elementos soportados para reglas y macros” en la página 224 para obtener más información. Cuando se combinan argumentos, debe utilizar paréntesis () para agrupar los argumentos y el carácter para indicar un valor booleano OR.

Además de las directrices y sintaxis que de la sección Macros, la vista de origen posee unas directrices adicionales que no son necesarias cuando se trabaja en la vista de editor. Las macros deben respetar lo siguiente cuando se trabaja en modalidad de origen:

- Cada macro debe comenzar con la línea [macro] marcada para indicar el comienzo de una macro.
- Para desactivar un elemento, coloque un indicador de comentario (#) delante de cada línea.

Ejemplo. En este ejemplo se define una macro denominada mTopic. El valor de mTopic es la presencia de un término que coincide con *uno* de los siguientes tipos: <Product>, <Person>, <Location>, <Organization>, <Budget> o <Unknown>.

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

Reglas de la vista Origen

```
[pattern(ID)]
name = pattern_name
value = [$type_name|macro_name|word_gaps|literal_strings]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]
```

Tabla 50. Entradas de regla

[pattern (<ID>)]	Indica el inicio de una regla de análisis de enlace de texto y proporciona un uso identificador numérico exclusivo para determinar el orden de procesamiento.
name	Proporciona un nombre exclusivo para esta regla de análisis de enlace de texto.
value	Proporciona la sintaxis y los argumentos que han de coincidir con el texto. Para obtener más información, consulte el tema “Elementos soportados para reglas y macros” en la página 224 .

Tabla 50. Entradas de regla (continuación)

output	El formato de la salida para los patrones coincidentes resultantes que se han descubierto en el texto. Los resultados no siempre coinciden con la posición original exacta de los elementos en el texto origen. Además, es posible tener varias líneas de salida para una regla de análisis de enlaces de texto mediante la colocación de cada salida en una línea separada. Sintaxis de los resultados: • Salida separada con el código de tabulador \t, tal como \$1\t#1\t\$3\t#3 • \$ y un número llama al término que coincide con el argumento definido en el parámetro de valor en esa posición. Así pues, \$1 indica el término que coincide con el primer argumento definido para el valor. • # y un número llama al nombre de tipo del elemento en esa posición. Si un elemento es una lista de cadenas literales, se asignará el tipo <Unknown>. • Un valor Null\tNull no creará una salida.
--------	---

Además de las directrices y la sintaxis tratada en la sección Reglas, la vista de origen posee algunas directrices adicionales que no son necesarias cuando se trabaja en la vista de editor. Las reglas también deben respetar lo siguiente cuando de trabaja en modalidad de origen:

- Siempre que se definen dos o más elementos, deben estar entre paréntesis tanto si son opcionales como si no (por ejemplo, (\$Negative|\$Positive) o (\$mCoord|\$SEP)?). \$SEP representa una coma.
- El primer elemento en una regla de análisis de enlaces de texto no puede ser opcional. Por ejemplo, no puede empezar con value = \$mTopic? o value = @{0,1}.
- Es posible asociar una cantidad (o recuento de instancias) a un elemento. Esto resulta útil cuando se escribe solo una regla que comprenda todos los casos en lugar de escribir una regla individual para cada caso. Por ejemplo, puede utilizar la serie literal (\$SEP|and) si está intentando emparejar , (coma) o and. Si amplía esto mediante la adición de una cantidad de modo que la serie literal se convierta en (\$SEP|and){1,2}, podrá ahora coincidir cualquiera de las siguientes instancias: ", " "and" ", and".
- No se admiten espacios entre el nombre de la macro y los caracteres \$ y ? en la regla de análisis de enlaces de texto value.
- Los espacios no están soportados en la regla de análisis de enlaces de texto output.
- Para desactivar un elemento, coloque un indicador de comentario (#) delante de cada línea.

Ejemplo. Supongamos que los recursos contienen la siguiente regla de análisis de enlaces de texto TLA y que ha habilitado la extracción de los resultados de TLA:

```
## Jean Doe was the former HR director of IBM in France
[pattern(201)]
name= 1_201
value = $Person ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Function
(of|with|for|in|to|at) @{0,1} $Organization @{0,2} $Location
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

Al realizar la extracción, el motor de extracción leerá cada frase e intentará la coincidencia de la secuencia siguiente:

Tabla 51. Ejemplo de secuencia de extracción

Posición	Descripción de los argumentos
1	El nombre de una persona (\$Person),

Tabla 51. Ejemplo de secuencia de extracción (continuación)

Posición	Descripción de los argumentos
2	Uno o dos de los siguientes: coma (\$SEP), determinador (\$mDet), verbo auxiliar (\$mSupport), las series “then” o “as”,
3	0 o 1 palabra (@{0,1})
4	Una función (\$Function)
5	Una de las siguientes series: “of”, “with”, “for”, “in”, “to”, o “at”,
6	0 o 1 palabra (@{0,1})
7	El nombre de una organización (\$Organization),
8	0, 1 o 2 palabras (@{0,2})
9	El nombre de una ubicación (\$Location),

Este ejemplo de regla de análisis de enlaces de texto hará coincidir las oraciones o frases como:

Jean Doe, la directora de recursos humanos de IBM en Francia

Jean Doe es la ex directora de recursos humanos de IBM en Francia

IBM nombró a Jean Doe como la directora de recursos humanos de IBM en Francia

Este ejemplo de regla de análisis de enlaces de texto produciría la siguiente salida:

```
jean doe <Person> hr director <Function> ibm <Organization> france <Location>
```

Donde:

- `jean doe` es el término correspondiente a \$1 (el primer elemento de la regla de análisis de enlaces de texto) y `<Person>` es el tipo para `jean doe` (#1),
- `hr director` es el término correspondiente a \$4 (el 4º elemento de la regla de análisis de enlaces de texto) y `<Function>` es el tipo para `hr director` (#4),
- `ibm` es el término correspondiente a \$7 (el 7º elemento de la regla de análisis de enlaces de texto) y `<Organization>` es el tipo para `ibm`. (#7),
- `france` es el término correspondiente a \$9 (el 9º elemento de la regla de análisis de enlaces de texto) y `<Location>` es el tipo para `france` (#9)

Conjuntos de reglas en la vista Origen

```
[set(<ID>)]
```

Donde `[set (<ID>)]` indica el inicio de un conjunto de reglas y proporciona un uso de identificador numérico único para determinar el orden de procesamiento de los conjuntos.

Ejemplo. La siguiente oración contiene información acerca de individuos, su función en la compañía y también las actividades de fusión/adquisición de dicha compañía.

Org1 Inc has entered into a definitive merger agreement with Org2 Ltd, said John Doe, CEO of Org2 Ltd.

Podría escribir una regla con varias salidas para manejar todas las salidas posibles, como:

```
## Org1 Inc entered into a definitive merger agreement with Org2 Ltd, said John Doe, CEO of Org2 Ltd.
```

```
[pattern(020)]
name=020
```

```

value = $Organization @{0,4} $ActionNouns @{0,6} $mOrg @{1,2}
$Person @{0,2} $Function @{0,1} $Organization
output = $1\t#1\t$3\t#3\t$5\t#5
output = $7\t#7\t$9\t#9\t$11\t#11

```

que generaría los dos patrones de resultados siguientes:

- org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd<Organization>
- john doe <Person> + ceo <Function> + org2 ltd<Organization>

Importante: Tenga en cuenta que otras operaciones lingüísticas se realizan durante la extracción del patrón TLA. En este caso, *merger* se agrupa bajo *merges with* durante la fase de agrupación de sinónimos del proceso de extracción. Y puesto que *merges with* pertenece al tipo <ActiveVerb>, este nombre de tipo es lo que aparece en la salida de patrón TLA final. Entonces, cuando la salida lea *t\$3\t#3*, significa que el patrón mostrará el concepto final del tercer elemento y el tipo final del tercer elemento después que se hayan aplicado todos los procesos lingüísticos (sinónimos y otras agrupaciones).

En lugar de escribir reglas complejas como las anteriores, puede ser más fácil gestionar y trabajar con dos reglas. El primero está especializado en descubrir las fusiones/adquisiciones entre las compañías:

```

[set(1)]
### Org1 Inc has entered into a definitive merger agreement with Org2 Ltd
[pattern(44)]
name=firm + action + firm_0044
value=$mOrg @{0,20} $ActionNouns @{0,6} $mOrg
output(1)=$1\t#1\t$3\t#3\t$5\t#5

```

que generaría *org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd <Organization>*

El segundo está especializado en la persona/función/compañía:

```

[set(2)]
### said John Doe, CEO of Org2 Ltd
[pattern(52)]
name=individual + role + firm_0007
value=$Person @{0,3} $mFunction (at|of)? ($mOrg|$Media|$Unknown)
output(1)=$1\t#1\t$3\tFunction\t$5\t#5

```

que generaría *john doe <Person> + ceo <Function> + org2 ltd <Organization>*

Avisos

Esta información se desarrolló para productos y servicios ofertados en todo el mundo.

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. El representante local de IBM le puede informar sobre los productos y servicios que están actualmente disponibles en su localidad. Las referencias a productos, programas o servicios de IBM no pretenden establecer ni implicar que sólo puedan utilizarse dichos productos, programas o servicios de IBM. En su lugar, se puede utilizar cualquier producto, programa o servicio equivalente que no infrinja ninguno de los derechos de propiedad intelectual de IBM. No obstante, es responsabilidad del usuario evaluar y verificar el funcionamiento de cualquier producto, programa o servicio que no sea de IBM.

Es posible que IBM tenga patentes o solicitudes de patente pendientes que traten el tema descrito en este documento. La entrega de este documento no le otorga ninguna licencia sobre dichas patentes. Puede enviar preguntas acerca de las licencias, por escrito, a:

*IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.*

Para consultas de licencia relativas a la información de doble byte (DBCS), póngase en contacto con el departamento de propiedad intelectual de IBM de su país, o envíe las consultas, por escrito, a:

*Intellectual Property Licensing
Derecho de propiedad intelectual y legal
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan*

INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL" SIN GARANTÍAS DE NINGUNA CLASE, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUYENDO, PERO SIN LIMITARSE A, LAS GARANTÍAS IMPLÍCITAS DE NO VULNERACIÓN, COMERCIALIZACIÓN O ADECUACIÓN A UN FIN DETERMINADO. Algunas jurisdicciones no permiten la renuncia a las garantías explícitas o implícitas en determinadas transacciones; por lo tanto, es posible que esta declaración no sea aplicable en su caso.

Esta información podría contener imprecisiones técnicas o errores tipográficos. Periódicamente se efectúan cambios en la información aquí contenida; estos cambios se incorporarán en nuevas ediciones de la publicación. IBM se reserva el derecho a realizar en cualquier momento y sin notificación previa, mejoras o modificaciones en los productos y programas que se describen en el presente manual.

Las referencias en esta información a sitios web no IBM se proporcionan solo por comodidad y en modo alguno constituyen un aval de dichos sitios web. El material de esos sitios web no forma parte del material de este producto de IBM y el uso de dichos sitios web es responsabilidad del usuario.

IBM puede utilizar o distribuir cualquier información que proporcione en la forma que crea adecuada sin incurrir por ello en ninguna obligación con el cliente.

Los titulares de licencias de este programa que deseen obtener información sobre el mismo con el fin de permitir: (i) el intercambio de información entre programas creados independientemente y otros programas (incluido éste) y (ii) el uso mutuo de la información intercambiada, deben ponerse en contacto con:

*IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119*

Dicha información puede estar disponible, sujeta a los términos y condiciones correspondientes, incluyendo, en algunos casos, el pago de una tarifa.

IBM proporciona el programa bajo licencia descrito en esta información y todo el material bajo licencia disponible para él de acuerdo con los términos del Contrato del cliente de IBM, el Acuerdo internacional de licencia de programas de IBM o cualquier contrato equivalente existente entre las partes.

Los datos de rendimiento y los ejemplos de cliente se presentan únicamente con fines ilustrativos. Los resultados reales del rendimiento pueden variar en función de determinadas configuraciones y condiciones de funcionamiento.

La información acerca de productos que no son de IBM se ha obtenido de los proveedores de dichos productos, sus publicaciones anunciadas u otras fuentes disponibles públicamente. IBM no ha comprobado estos productos y no puede confirmar la precisión de su rendimiento, compatibilidad ni contemplar ninguna otra reclamación relacionada con los productos que no son de IBM. Las preguntas acerca de las funciones de productos que no sean de IBM deben dirigirse a los proveedores de dichos productos.

Las declaraciones sobre la intención y el rumbo futuro IBM están sujetas a cambio o retirada sin previo aviso y representan únicamente metas y objetivos.

Esta información contiene ejemplos de datos e informes utilizados en operaciones comerciales diarias. Para ilustrarlos de la forma más completa posible, los ejemplos incluyen nombres de personas, empresas, marcas y productos. Todos estos nombres son ficticios y cualquier similitud con nombres de personas o empresas es mera coincidencia.

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas comerciales o marcas registradas de International Business Machines Corp., registradas en muchas jurisdicciones en todo el mundo. Otros nombres de productos y servicios pueden ser marcas registradas de IBM o de otras empresas. Existe una lista actual de las marcas registradas de IBM en la web bajo "Copyright and trademark information" en www.ibm.com/legal/copytrade.shtml.

Intel, el logotipo de Intel, Intel Inside, el logotipo de Intel Inside, Intel Centrino, el logotipo de Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium y Pentium son marcas registradas de Intel Corporation o de sus filiales en los Estados Unidos de América y en otros países.

Linux es una marca registrada de Linus Torvalds en los Estados Unidos o en otros países.

Microsoft, Windows, Windows NT, y el logotipo de Windows son marcas comerciales de Microsoft Corporation en Estados Unidos, otros países o ambos.

UNIX es una marca registrada de The Open Group en Estados Unidos y en otros países.

Java y todas las marcas registradas y logotipos basados en Java son marcas registradas de Oracle o sus filiales.

Otros nombres de productos y servicios pueden ser marcas registradas de IBM u otras compañías.

Índice

Caracteres Especiales

*.lib [183](#)

& |() operadores de reglas [129](#)

área de trabajo interactiva [23](#), [24](#), [26](#), [67](#), [77](#)

índice para mapas de concepto [88](#)

A

abreviaciones [206](#), [210](#)

activación de entidades no lingüísticas [205](#)

actualizar

bibliotecas [184](#), [185](#)

nodos de modelado [77](#)

plantilla y recursos del nodo [174](#)

plantillas [166](#), [173](#)

adición

bibliotecas públicas [180](#)

conceptos en categorías [138](#)

descriptores [100](#)

elementos opcionales [196](#)

sinónimos [90](#), [195](#)

sonidos [75](#), [76](#)

términos en diccionarios de tipo [190](#)

términos para la lista de exclusión [197](#)

tipos [91](#)

ajuste de columna [75](#)

almacenamiento

área de trabajo interactiva [77](#)

Canales de información [13](#)

plantillas [173](#)

recursos [176](#)

recursos como plantillas [166](#)

resultados de extracción de datos y sesión [24](#)

almacenamiento en antememoria

Canales de información [13](#)

resultados de extracción de datos y sesión [24](#)

aminoácidos (entidad no lingüística) [202](#)

ampliación de categorías [115](#)

análisis de enlaces de texto (TLA)

argumentos [224](#)

aviso en el árbol [215](#)

cuándo editar [212](#)

dónde comenzar [212](#)

editar macros y reglas [211](#)

editor de reglas [211](#)

en nodos de modelado de minería de textos [24](#)

especificar qué biblioteca [211](#), [215](#)

explorar patrones [149](#)

filtrado de patrones [151](#)

gráfico de malla [162](#), [163](#)

inhabilitación y supresión de reglas [222](#)

macros [216](#)

modalidad de origen [226](#)

navegación por reglas y macros [215](#)

nodo TLA [45](#)

orden de proceso de regla [223](#)

análisis de enlaces de texto (TLA) (*continuación*)

panel Datos [153](#)

Panel Visualización [162](#), [163](#)

proceso de varios pasos [224](#)

Reglas de reasignación de tipo [155](#)

simulación de resultados [213](#), [214](#)

TRR [155](#)

ver gráficos [162](#), [163](#)

análisis de texto [2](#)

AND, operador de regla [129](#)

anotaciones

para categorías [103](#)

antienlaces [110](#)

apertura de plantillas [173](#)

aplanamiento de categorías [140](#)

archivos .doc/.docx/.docm para minería de textos [11](#)

archivos .htm/.html para minería de textos [11](#)

archivos .pdf para minería de textos [11](#)

archivos .ppt/.pptx/.pptmfiles para minería de textos [11](#)

archivos .rtf para minería de textos [11](#)

archivos .shtml para minería de textos [11](#)

archivos .txt/.text para minería de textos [11](#)

Archivos .xls/.xlsx de Microsoft Excel

exportación de categorías predefinidas [134](#)

importación de categorías predefinidas [130](#)

archivos .xls/.xlsx/.xlsm para minería de textos [11](#)

archivos .xml para minería de textos [11](#)

arrastrar y soltar [119](#)

asterisco (*)

diccionario de exclusión [197](#)

sinónimos [195](#)

atajos de teclado [77](#), [78](#)

B

biblioteca Budget [188](#)

Biblioteca Opinions [188](#)

bibliotecas

actualizar [185](#)

adición [180](#)

advertencia de sincronización de bibliotecas [184](#)

biblioteca Budget [188](#)

Biblioteca Opinions [188](#)

bibliotecas locales [184](#)

bibliotecas predeterminadas enviadas [179](#)

bibliotecas públicas [184](#)

cambio del nombre [182](#)

Core library [188](#)

creación [180](#)

denominación [182](#)

desactivación [182](#)

diccionarios [179](#)

enlaces [180](#)

exportación [183](#)

importar [183](#)

publicar [185](#)

reparto y publicación [184](#)

bibliotecas (*continuación*)

Sincronización [184](#)

suprimir [183](#)

ver [182](#)

bibliotecas predeterminadas [179](#)

bibliotecas suministradas (valor predeterminado) [179](#)

botón mostrar [96](#)

botón recuento [96](#)

buscar y reemplazar (recursos avanzados) [200](#)

búsqueda de términos y tipos [181](#)

C

cadena literales [224](#)

calcular valores de similitud de enlaces [145](#)

cambio

plantillas [166](#), [173](#)

cambio del nombre

bibliotecas [182](#)

categorías [118](#)

diccionarios de tipo [193](#)

plantillas de recursos [175](#)

Campo de ID [45](#)

campos de documentos [53](#)

carga de plantillas de recursos [26](#), [45](#), [174](#)

categorías

ampliar [110](#), [115](#)

anotaciones [103](#)

añadir a [138](#)

aplanamiento [140](#)

cambio del nombre [118](#)

creación [98](#), [114](#), [119](#)

creación de categorías vacías nuevas [118](#)

creación manual [118](#)

descriptores [99](#), [100](#), [102](#)

desplazamiento [139](#)

editar [138](#), [139](#)

estrategias [98](#)

etiquetas [103](#)

fusión [140](#)

generación [106](#), [108](#), [110](#), [115](#)

nombres [103](#)

nuggets de modelo de categoría de minería de textos [26](#)

paquetes de análisis de texto [135–137](#)

Propiedades [103](#)

puntuación [96](#)

refinamiento de los resultados [138](#)

relevancia [105](#)

suprimir [141](#)

categorías léxicas [207](#), [209](#)

categorías nuevas [118](#)

categorías predefinidas

formato compacto [132](#)

formato con sangrado [133](#)

formato de lista plana [132](#)

categorización

derivación de raíz de conceptos [108](#), [110](#)

inclusión de conceptos [110](#), [112](#)

manualmente [118](#)

Métodos [98](#)

redes semánticas [108](#), [110](#), [112](#)

reglas de co-ocurrencia [110](#), [113](#)

técnicas de frecuencia [114](#)

técnicas lingüísticas [106](#), [115](#)

categorización (*continuación*)

uso de técnicas de agrupación [108](#)

utilización de técnicas [110](#)

cerrar la sesión [77](#)

clústeres

acerca de [143](#)

descriptores [147](#)

exploración [146](#)

generación [144](#)

gráfico clúster web [161](#), [162](#)

gráfico concepto web [161](#)

valores de similitud de enlaces [145](#)

coincidencia de texto [103](#)

color de fuente [189](#)

colores

diccionario de exclusión [197](#)

establecimiento de opciones de color [75](#)

para tipos y términos [189](#)

sinónimos [195](#)

colores personalizados [75](#)

columna de documentos [96](#)

combinación de categorías [140](#)

Compartimiento de bibliotecas

actualizar [185](#)

adición de bibliotecas públicas [180](#)

publicar [185](#)

conceptos

adición a las categorías [99](#), [102](#), [138](#)

adición a tipos [91](#)

como campos o registros para puntuación [33](#), [40](#)

creación de tipos [89](#)

en categorías [99](#), [102](#)

en clústeres [147](#)

exclusión de la extracción [92](#)

extraer [81](#)

filtrado [85](#)

forzado en la extracción [93](#)

los mejores descriptores [100](#)

mapas de conceptos [86](#)

conceptos de correlación [86](#)

configuración [74–76](#)

copia de seguridad de recursos [176](#)

Core library [188](#)

correo electrónico (entidad no lingüística) [202](#)

creación

bibliotecas [180](#)

categorías [26](#), [98](#), [106](#), [119](#)

categorías con reglas [120](#)

diccionarios de tipo [189](#)

elementos opcionales [196](#)

entradas del diccionario de exclusión [197](#)

Nodos de modelado y los nuggets de modelo de categoría [76](#)

plantilla a partir de los recursos [166](#)

plantillas [173](#)

reglas de categoría [119](#), [120](#), [129](#)

sinónimos [89](#), [90](#), [195](#)

tipos [91](#)

creación de plantillas a partir de recursos [166](#)

cree un índice de mapa de concepto [88](#)

D

Datos

Datos (*continuación*)

- agrupación en clústeres [143](#)
- análisis de enlace de texto [149](#)
- categorización [95](#), [106](#), [118](#)
- extracción de patrones de enlace de texto [149](#)
- extraer [81](#), [82](#), [150](#)
- filtrar resultados [85](#), [151](#)
- generación de categorías [108](#), [110](#), [115](#)
- panel Datos [103](#), [153](#)
- reestructuración [48](#)
- refinamiento de los resultados [89](#)
- definiciones [99](#), [102](#)
- definiciones forzadas [206](#), [209](#)
- delimitador global [75](#)
- delimiter [75](#)
- denominación
 - bibliotecas [182](#)
 - categorías [103](#)
 - diccionarios de tipo [193](#)
- desactivación
 - bibliotecas [182](#)
 - diccionarios de exclusión [197](#)
 - diccionarios de sinónimos [201](#)
 - diccionarios de sustitución [196](#)
 - diccionarios de tipo [193](#)
 - entidades no lingüísticas [205](#)
- desactivación de entidades no lingüísticas [205](#)
- descriptores
 - categorías [99](#), [102](#)
 - clústeres [147](#)
 - edición en categorías [139](#)
 - selección de los mejores [100](#)
- desplazamiento
 - categorías [139](#)
 - diccionarios de tipo [193](#)
- diccionario de exclusión [179](#), [197](#)
- diccionario de sustitución [179](#), [194](#)–[196](#)
- diccionario de tipo
 - adición de términos [190](#)
 - cambio del nombre [193](#)
 - creación de tipos [189](#)
 - desactivación [193](#)
 - desplazamiento [193](#)
 - elementos opcionales [187](#)
 - forzado de términos [192](#)
 - sinónimos [187](#)
 - suprimir [193](#)
 - tipos incorporados [188](#)
- diccionario de tipo Budget [188](#)
- diccionario de tipo Location [188](#)
- diccionario de tipo negativo [188](#)
- diccionario de tipo Organization [188](#)
- diccionario de tipo Person [188](#)
- diccionario de tipo Positive [188](#)
- diccionario de tipo Product [188](#)
- diccionario de tipo Uncertain [188](#)
- diccionario de tipo Unknown [188](#)
- diccionarios
 - excluidos [179](#), [187](#), [197](#)
 - sustituciones [179](#), [187](#), [194](#)
 - tipos [179](#), [187](#)
- dígitos (entidad no lingüística) [202](#)
- direcciones (entidad no lingüística) [202](#)
- direcciones IP (entidad no lingüística) [202](#)

- documentos
 - listado [53](#)

E

- editar
 - categorías [138](#), [139](#)
 - refinado de los resultados de la extracción [89](#)
 - reglas de categoría [130](#)
- Editor de plantillas
 - actualización de los recursos en el nodo [174](#)
 - apertura de plantillas [173](#)
 - bibliotecas de recursos [179](#)
 - cambio de nombre de plantillas [175](#)
 - cómo salir del editor [176](#)
 - eliminación de plantillas [175](#)
 - guardado de plantillas [173](#)
 - importación y exportación [175](#)
- editor de recursos
 - actualización de plantillas [166](#)
 - cambio de recursos [166](#)
 - creación de plantillas [166](#)
- elementos opcionales
 - adición [196](#)
 - definición [194](#)
 - destino [196](#)
 - eliminación de entradas [196](#)
- enlaces en clústeres [143](#)
- enlaces externos [143](#)
- enlaces internos [143](#)
- entidades no lingüísticas
 - aminoácidos [202](#)
 - dígitos [202](#)
 - direcciones de correo electrónico [202](#)
 - direcciones de HTTP/URL [202](#)
 - Direcciones IP [202](#)
 - Direcciones TCP/IP [202](#)
 - expresiones regulares, RegExp.ini [203](#)
 - fechas [202](#)
 - formato de fecha [205](#)
 - habilitar y inhabilitar [205](#)
 - horas [202](#)
 - monedas [202](#)
 - normalización, NonLingNorm.ini [205](#)
 - números de la seguridad social [202](#)
 - números de teléfono [202](#)
 - pesos y medidas [202](#)
 - porcentajes [202](#)
 - proteínas [202](#)
- errores ortográficos [201](#)
- espacios entre palabras [224](#)
- estructuración en componentes [110](#)
- estructuración en componentes de los términos [110](#)
- etiqueta
 - reutilizar canales de información web [13](#)
- etiquetas para las categorías [103](#)
- excepciones de agrupación difusa [199](#), [201](#)
- excepciones de enlace [110](#)
- exclusión
 - conceptos de la extracción [92](#)
 - de enlaces de categoría [110](#)
 - de exclusión difusa [201](#)
 - desactivación de bibliotecas [182](#)
 - desactivación de diccionarios [193](#), [196](#)

exclusión (*continuación*)
 desactivación de entradas de exclusión [197](#)
exportación
 bibliotecas públicas [183](#)
 categorías predefinidas [134](#)
 plantillas [175](#)
extraer
 forzado de palabras [93](#)
 patrones a partir de datos [45](#)
 patrones TLA [150](#)
 refinamiento de los resultados [89](#)
 resultados de la extracción [81](#)
 unitérminos [5](#)

F

fechas (entidad no lingüística) [202](#), [205](#)
filtrado de bibliotecas [182](#)
filtrar resultados [85](#), [151](#)
forma plural de las palabras [189](#)
formas declinadas [110](#), [187](#), [189](#), [190](#)
formato compacto [132](#)
formato con sangrado [133](#)
formato de fecha
 entidades no lingüísticas [205](#)
formato de lista plana [132](#)
formatos HTML para canales de información web [13](#), [14](#)
formatos RSS para canales de información web [13](#), [14](#)
frecuencia [114](#)
frecuencia de tipo [114](#)
fusión de categorías [140](#)

G

generación
 categorías [2](#), [6](#), [106](#), [108](#), [110](#), [112–115](#), [118](#)
 clústeres [144](#)
generación de categorías
 excepciones de enlace de clasificación [110](#)
 técnica de derivación de raíz de conceptos [115](#)
 técnica de inclusión de conceptos [115](#)
 técnica de redes semánticas [115](#)
 técnica de reglas de coocurrencia [115](#)
generación de nodos y nuggets de modelo [76](#)
generador de expresiones [78](#)
generar formas declinadas [187](#), [189](#), [190](#)
gestionar
 bibliotecas locales [182](#)
 bibliotecas públicas [183](#)
 categorías [138](#)
gráfico concepto web [161](#)
gráfico de barras de categorías [160](#)
gráfico web concepto TLA [162](#), [163](#)
gráfico/tabla de malla de categorías [160](#)
gráficos
 editar [163](#)
 gráfico clúster web [161](#), [162](#)
 gráfico concepto web [161](#)
 gráfico web concepto TLA [162](#), [163](#)
 mapas de conceptos [86](#)
 modo de exploración [163](#)
 tipo de gráfico web [162](#), [163](#)
gráficos de malla

gráficos de malla (*continuación*)
 gráfico clúster web [161](#), [162](#)
 gráfico concepto web [161](#)
 gráfico web concepto TLA [162](#), [163](#)
 tipo de gráfico web [162](#), [163](#)

H

horas (entidad no lingüística) [202](#)
HTTP/URL (no lingüísticos) [202](#)

I

idioma de destino [201](#)
importar
 bibliotecas públicas [183](#)
 categorías predefinidas [130](#)
 plantillas [175](#)
información de sesión [23](#), [24](#), [26](#)
iniciar área de trabajo interactiva [23](#)

L

lectores de pantallas [77](#), [78](#)
lengua
 configuración del idioma de destino para los recursos [201](#)
lista de extensiones en nodo lista de archivos [11](#)

M

macros
 mNonLingEntities [218](#)
 mTopic [218](#)
mapas de conceptos
 compilar índice [88](#)
marcos de código [130](#)
minería de textos [2](#)
mNonLingEntities [218](#)
modalidad partición [21](#)
modelos
 argumentos [224](#)
 editor de reglas de enlace de texto [211](#)
 proceso de varios pasos [224](#)
modo de edición [163](#)
modo de exploración [163](#)
monedas (entidad no lingüística) [202](#)
mostrar columnas en el panel categorías [96](#)
mostrar columnas en el panel Datos [153](#)
mTopic [218](#)

N

navegación de atajos de teclado [77](#)
nodo canal de información web
 ejemplo [16](#)
 entrada, pestaña [13](#)
 Etiqueta para el almacenamiento en antememoria y reutilización [13](#)
 pestaña contenido [16](#)
 pestaña registros [14](#)
 propiedades de los scripts [57](#)
nodo de análisis de enlaces de texto

- nodo de análisis de enlaces de texto (*continuación*)
 - almacenamiento en antememoria de TLA [49](#)
 - ejemplo [49](#)
 - pestaña de campos [45](#)
 - pestaña experto [47](#)
 - propiedades de los scripts [64](#)
 - reestructuración de datos [48](#)
 - salida [48](#)
- nodo de idioma
 - pestaña Configuración [17](#)
 - propiedades de los scripts [58](#)
- nodo de modelado de minería de textos
 - actualizar [77](#)
 - ejemplo [29](#)
 - generación de nodo nuevo [76](#)
 - pestaña de campos [21](#)
 - pestaña de modelos [23](#)
 - pestaña experto [27](#)
 - propiedades de script para TextMiningWorkbench [59](#)
- nodo lista de archivos
 - ejemplo [12](#)
 - lista de extensiones [11](#)
 - otras pestañas [12](#)
 - pestaña Configuración [11](#)
 - propiedades de los scripts [57](#)
- nodo Muestrear
 - al minar textos [29](#)
- nodo visor
 - ejemplo [53](#)
 - para minería de textos [53](#)
 - pestaña Configuración [53](#)
- nodos
 - análisis de enlace de texto [7](#), [45](#)
 - canal de información web [7](#), [13](#)
 - lengua [17](#)
 - lista de archivos [7](#), [11](#)
 - nodo de modelado de minería de textos [7](#), [20](#)
 - nugget de modelo de concepto [30](#)
 - nugget de modelo de minería de textos [7](#)
 - nuggets de modelo de categoría [38](#)
 - visor de minería de textos [7](#), [53](#)
- nodos de origen
 - canal de información web [7](#), [13](#)
 - lista de archivos [7](#), [11](#)
- nombre de categoría [96](#)
- normalización [205](#)
- NOT, operador de regla [129](#)
- nugget de modelo de minería de textos
 - propiedades de script para TMWBModelApplier [62](#)
- nuggets de modelo
 - generación desde el área de trabajo interactiva [76](#)
 - nuggets de modelo de categoría [19](#), [23](#), [26](#), [38](#), [39](#)
 - nuggets de modelo de concepto [19](#), [23](#), [26](#), [30](#), [31](#)
- nuggets de modelo de categoría
 - conceptos como campos o registros [40](#)
 - creación a través de entorno de trabajo [24](#)
 - creación a través de nodo [26](#)
 - ejemplo [42](#)
 - generación [76](#)
 - pestaña Configuración [40](#)
 - pestaña de campos [41](#)
 - pestaña de modelos [39](#)
 - pestaña resumen [41](#)
 - salida [39](#)

- nuggets de modelo de concepto
 - conceptos como campos o registros [33](#)
 - conceptos para puntuación [31](#)
 - creación a través de nodo [26](#)
 - ejemplo [35](#)
 - pestaña Configuración [33](#)
 - pestaña de campos [34](#)
 - pestaña de modelos [31](#)
 - pestaña resumen [35](#)
 - sinónimos [33](#)
- número de la seguridad social (entidad no lingüística) [202](#)
- número máximo de categorías para crear. [108](#)
- números de teléfono (entidad no lingüística) [202](#)

O

- opción de coincidencia [187](#), [189](#), [190](#)
- opciones
 - opciones de sesión [75](#)
 - opciones de sonido [76](#)
 - Opciones de visualización (colores) [75](#)
- opciones de sonido [76](#)
- operador de exclusión [224](#)
- Operadores booleanos [129](#)
- operadores de regla & | !() [129](#)
- OR, operador de regla [129](#)

P

- panel Categorías [96](#)
- panel Datos
 - botón mostrar [96](#)
 - Reglas de reasignación de tipo [155](#)
 - TRR [155](#)
 - vista Análisis para los enlaces de texto [153](#)
 - vistas de categorías y conceptos [103](#)
- panel Visualización
 - gráfico clúster web [161](#), [162](#)
 - gráfico concepto web [161](#)
 - gráfico web concepto TLA [162](#), [163](#)
 - tipo de gráfico web [162](#), [163](#)
 - Vista Análisis para los enlaces de texto [162](#), [163](#)
- paquetes de análisis de texto
 - cargando [136](#)
- paquetes de análisis de texto *.tap [135–137](#)
- pasar por alto conceptos [92](#)
- patrones de concepto [151](#)
- patrones de extracción [206](#), [207](#)
- patrones de tipo [151](#)
- pesos/medidas (no lingüísticos) [202](#)
- plantillas
 - actualización o acción de guardar como [166](#)
 - almacenamiento [173](#)
 - apertura de plantillas [173](#)
 - cambio de recursos [166](#)
 - cambio del nombre [175](#)
 - creación a partir de recursos [166](#)
 - importación y exportación [175](#)
 - realización de copia de seguridad [176](#)
 - recuadro de diálogo de carga de plantillas de recursos [26](#)
 - restauración [176](#)
 - suprimir [175](#)

- plantillas (*continuación*)
 - TLA [166](#)
- plantillas de recursos [5](#), [45](#), [73](#), [149](#), [165](#), [169](#)
- porcentajes (entidad no lingüística) [202](#)
- preferencias [74–76](#)
- procedimiento forzado
 - extracción de conceptos [93](#)
 - términos [192](#)
- proceso de varios pasos [224](#)
- Propiedades
 - categorías [103](#)
- propiedades de languageidentifier [58](#)
- propiedades de script de filelistnode [57](#)
- propiedades de script para TMWBModelApplier [62](#)
- propiedades de script TextMiningWorkbench [59](#)
- propiedades de webfeednode [57](#)
- propiedades textlinkanalysis [64](#)
- proteínas (entidad no lingüística) [202](#)
- publicar
 - adición de bibliotecas públicas [180](#)
 - bibliotecas [184](#)
- puntuación
 - conceptos [32](#)

R

- recursos
 - bibliotecas predeterminadas enviadas [179](#)
 - cambio de los recursos de una plantilla [166](#)
 - edición de recursos avanzados [199](#)
 - realización de copia de seguridad [176](#)
 - restauración [176](#)
- recursos avanzados
 - buscar y reemplazar en el editor [200](#)
- recursos lingüísticos
 - paquetes de análisis de texto [135–137](#)
 - plantillas [165](#)
 - plantillas de recursos [169](#)
- refinamiento de los resultados
 - adición de conceptos a tipos [91](#)
 - adición de sinónimos [90](#)
 - categorías [138](#)
 - creación de tipos [91](#)
 - exclusión de conceptos [92](#)
 - forzado de extracción de conceptos [93](#)
 - resultados de la extracción [89](#)
- registros [103](#), [153](#)
- reglas
 - creación [129](#)
 - editar [130](#)
 - Operadores booleanos [129](#)
 - sintaxis [120](#)
 - suprimir [130](#)
 - técnica de reglas de coocurrencia [113](#)
- reglas de categoría
 - de la coocurrencia de conceptos [110](#), [113](#), [115](#)
 - de sinónimos [108](#), [110](#), [115](#)
 - ejemplos [127](#)
 - reglas de co-ocurrencia [110](#), [115](#)
 - sintaxis [120](#)
- Reglas de reasignación de tipo [155](#)
- relevancia de las respuestas y categorías [105](#)
- restauración de recursos [176](#)
- resultados de las extracciones

- resultados de las extracciones (*continuación*)
 - filtrar resultados [85](#), [151](#)

S

- secciones de gestión de idiomas
 - abreviaciones [206](#), [210](#)
 - definiciones forzadas [206](#), [209](#)
 - patrones de extracción [206](#), [207](#)
- seleccionar conceptos para su puntuación [32](#)
- separadores [75](#)
- separadores de texto [75](#)
- signo de exclamación (!) [195](#)
- silenciado de sonidos [76](#)
- símbolo de dólar (\$) [195](#)
- símbolo de intercalación (^) [195](#)
- Símbolos ! ^ * \$ en sinónimos [195](#)
- simulación de resultados del análisis de enlace de texto
 - definición de datos [213](#)
- sin categorizar [96](#)
- sincronización de bibliotecas [184](#), [185](#)
- sinónimos
 - adición [90](#), [195](#)
 - colores [195](#)
 - definición [194](#)
 - eliminación de entradas [196](#)
 - en nuggets de modelo de concepto [33](#)
 - excepciones de agrupación difusa [201](#)
 - símbolos ! ^ * \$ [195](#)
 - términos de destino [195](#)
- suprimir
 - bibliotecas [183](#)
 - categorías [141](#)
 - desactivación de bibliotecas [182](#)
 - diccionarios de tipo [193](#)
 - elementos opcionales [196](#)
 - entradas excluidas [197](#)
 - plantillas de recursos [175](#)
 - reglas de categoría [130](#)
 - sinónimos [196](#)
- sustitución de recursos por una plantilla [166](#)

T

- tablas [78](#)
- teclas de métodos abreviados [77](#), [78](#)
- técnica de derivación de raíz de conceptos [108](#), [110](#), [115](#)
- técnica de inclusión de conceptos [110](#), [112](#), [115](#)
- técnica de redes semánticas [108](#), [110](#), [112](#), [115](#)
- técnica de reglas de coocurrencia [110](#), [113](#), [115](#)
- técnicas
 - arrastrar y soltar [119](#)
 - derivación de raíz de conceptos [108](#), [110](#), [115](#)
 - frecuencia [114](#)
 - inclusión de conceptos [110](#), [112](#), [115](#)
 - redes semánticas [108](#), [110](#), [112](#), [115](#)
 - reglas de co-ocurrencia [110](#), [113](#), [115](#)
- técnicas lingüísticas [2](#)
- términos
 - adición a tipos [190](#)
 - adición en el diccionario de exclusión [197](#)
 - búsqueda en el editor [181](#)
 - color [189](#)

- términos (*continuación*)
 - formas declinadas [187](#)
 - forzado de términos [192](#)
 - opciones de coincidencia [187](#)
- términos de destino [195](#)
- términos subyacente [33](#)
- tipo de gráfico web [162](#), [163](#)
- tipos
 - adición de conceptos [89](#)
 - búsqueda en el editor [181](#)
 - color predeterminado [75](#), [189](#)
 - creación [189](#)
 - diccionarios [179](#)
 - extraer [81](#)
 - filtrado [85](#), [151](#)
 - frecuencia de tipo [114](#)
 - tipos incorporados [188](#)
- títulos [53](#)
- TLA [166](#)
- todos los documentos [96](#)
- TRR [155](#)

U

- URL [13](#), [14](#)

V

- valor de enlace mínimo [108](#)
- valores de enlace [145](#)
- valores de similitud de enlaces [145](#)
- valores de visualización [75](#)
- ver
 - análisis de enlace de texto [162](#), [163](#)
 - bibliotecas [182](#)
 - clústeres [161](#)
 - documentos [53](#)
- vista de clústeres [70](#)
- vistas de categorías y conceptos
 - panel Categorías [96](#)
 - panel Datos [103](#)
- vistas en el entorno de trabajo interactivo
 - análisis de enlace de texto [71](#)
 - categorías y conceptos [67](#), [95](#)
 - clústeres [70](#)
 - editor de recursos [73](#)
- volver a utilizar
 - Canales de información [13](#)
 - resultados de extracción de datos y sesión [24](#)

W

- workbench [23](#), [24](#), [26](#)

