Natural Language Processing Cell

Installation Guide (Windows)

Version 1.0

1. Table of Contents

2. Document Version History

Date	**Version**	**Description**	**Author(s)**
11/26/2007	1.0	Version 1.0	Sergey Goryachev

3. Introduction

Informatics for Integrating Biology and the Bedside (I2B2) is one of the sponsored initiatives of the NIH Roadmap National Centers for Biomedical Computing (http://www.bisti.nih.gov/ncbc/). One of the goals of I2B2 is to provide clinical investigators with the software tools necessary to parse and extract clinical information from unstructured medical records using natural language processing (NLP). This guide with the provided source code will help you to learn how to deploy an I2B2 cell.

4. Prerequisites

The following sections describe the required software to run NLP cell. Please keep in mind, that NLP cell was tested in isolation. There may be supporting jar version conflicts if you try to install it with the core cells. Be careful not to break any existing installations by overwriting existing jars with new ones!

4.1 Package Contents

The current NLP cell distribution archive contains four folders. The “core” folder contains HITEx core files. HITEx is used by the NLP cell to do the server-side natural language processing. The following files are included in the “core” folder:

1. **apidocs.zip** - HITEx API JavaDoc.
2. **examples.zip** - contains 5 sample de-identified unstructured medical records to test HITEx core.
3. **hitex.jar** - HITEx core Java jar file.
4. **hitex_manual.html** – contains a quick start guide for the HITEx core, as well as examples of HITEx usage.
5. **i2b2-LICENSE.txt** – the I2B2 license agreement. You must accept the terms and conditions of this license agreement in order to use the NLP cell.
6. **resources.zip** - contains all required server-side resources to run NLP pipelines.
7. **source.zip** - contains the HITEx core source files.
8. **supporting-jars.zip** - contains supporting Java jar files required by the NLP cell. These jar files also include Java source.
9. **umls_2004aa_200702061416.zip** - contains SQL backup of the modified version of UMLS 2004AA database used by the cell for principal diagnoses extraction. The SQL backup is in the MySQL Administrator backup format.

The “webservice” folder contains server-side NLP cell files:

1. **config.xml** – the NLP cell configuration file. It is used by the cell to build and initialize the standard processing pipelines.
2. **hitex-server-2.3.zip** – contains the source code of the NLP cell web service.
3. **hitexservice.aar** – the NLP cell in the deployable Axis2 archive format.
4. **hitexservice.properties** – the property file that points the NLP cell to the location of config.xml file described above. It should be placed in the Tomcat installation directory (see more about it in sections below).
5. **i2b2-LICENSE.txt** – the I2B2 license agreement. You must accept the terms and conditions of this license agreement in order to use the NLP cell.
6. **sample_request_1.1.xml** – sample web service XML request in the Version 1.1 I2B2 messaging format.

7. **sample_response_1.1.xml** – sample web service XML response in the Version 1.1 I2B2 messaging format.

The Eclipse NLP Client plug-in is used to communicate with the NLP cell. The "ws-client" folder contains the Eclipse NLP Client plug-in related files:

1. **edu.harvard.i2b2.eclipse.plugins.nlp_1.0.0.jar** – The deployable version of Eclipse NLP client plug-in.
2. **hitex-client-src-1.0.zip** – the source code of the plug-in.

The "screenshots" contains a few screenshots of the Eclipse NLP client plug-in. Finally, md5.txt file contains the md5 checksums of all distribution files. It can be used to validate the distribution integrity.

4.2 Required Software

4.2.1 Java JDK (5.0 or higher is required)

1. Download JDK from http://java.sun.com/javase/downloads.
2. Install the JDK into directory of your choice
3. Set %JAVA_HOME% environment variable to point to the installation directory (e.g., C:\jata\tools\jdk1.5.0_12)
4. Add %JAVA_HOME%\bin to your %Path% variable.

4.2.2 Apache Ant

1. Download Apache Ant version 1.6.5 (apache-ant-1.6.5-bin.zip) from http://archive.apache.org/dist/ant/binaries.
2. Unzip the contents of the archive into directory of your choice.
3. Set %ANT_HOME% directory to point to your Ant installation directory (e.g., C:\java\tools\apache-ant-1.6.5).
4. Add %ANT_HOME%\bin to your %Path% variable.

4.2.3 MySQL Database

1. Download MySQL Community Edition database server version 5.0 from http://dev.mysql.com/downloads. MySQL database will be used to install UMLS.
2. Install database to the location of your choice. The database does not need to reside on the same machine where the NLP cell is installed. It is assumed that Windows XP machine is used. Administrative account is required to install MySQL database on Windows. The database does not need to support transactions. MyISAM engine type is the only type that is required. It is recommended to make MyISAM the default engine type. It is also recommended to select the "Dedicated server" installation profile to

improve cell performance in finding principal diagnoses. Choose the database startup as windows service during the installation.

3. Verify that MySQL server can start and that firewall is not blocking access to it.

4.2.4 GATE Framework

1. Download version 3.1 of Gate NLP framework from http://www.gate.ac.uk/download. Choose Windows-specific installer without bundled JRE if you are on Windows. NLP Cell was tested to work with Gate 3.1 version for Windows. It was not tested with the newest Gate version 4.0.
2. Install Gate into **directory that does not contain any spaces**. The default installation directory in **Program Files should not be used**.
3. Set %GATE_HOME% environment variable to point to the Gate installation directory (e.g., C:\java\tools\gate31).
4. Verify that Gate can start: execute %GATE_HOME%\GATE-3.1.exe, or use the shortcut created by Gate installer.

4.2.5 UMLS

NLP Cell relies on the UMLS database for principal diagnoses extraction. UMLS requires users to obtain a usage license. Please obtain a UMLS license at http://umlsks.nlm.nih.gov.

1. Download and install MySQL GUI Tools from http://dev.mysql.com/downloads. MySQL GUI Tools contain MySQL Administrator that will be used to restore UMLS database from the backup file.
2. Start MySQL Administrator and connect to the database using the root account.
3. Create a new schema and call it umls_2004aa. This schema will hold a modified version of UMLS 2004AA database used by the cell for the principal diagnoses extraction.
4. Create a new user (e.g., umlsuser) and give this user read access (i.e., SELECT) to the umls_2004aa schema.
5. Test if umlsuser can in fact connect to the umls_2004aa database. Start MySQL Query Browser tool from GUI Tools suite and use the umlsuser's credentials to connect to umls_2004aa database.
6. If the previous test succeeded, restore the UMLS 200AA database table into the umls_2004aa. Extract the umls_2004aa_200702061416.sql file from the umls_2004aa_200702061416.zip file. Use MySQL Administrator to log in as root, go to 'Restore' menu, select the umls_2004aa_200702061416.sql file, choose umls_2004aa as the destination schema and press 'Restore' button. The restoration process may take a few minutes (or longer). When it is finished, umls_2004aa will contain all the required tables and the associated indexes.

4.2.6 Apache Tomcat

1. Download Apache Tomcat 5.5.25 from http://tomcat.apache.org/download-55.cgi. Select the Windows Service Installer distribution, because we want to install Tomcat as a Windows service.
2. Install Tomcat into a directory of your choice.
3. Set %CATALINA_HOME% environment variable to point to the installation directory (e.g., C:\java\tools\apache-tomcat-5.5.23)
4. By default, Tomcat uses port 8080 to listen to connections. If this port is already in use, select a different port number for HTTP connector in %CATALINA_HOME%\conf\server.xml.
5. Set Tomcat to run with 1GB of memory. Right click the Tomcat service icon on the task bar, choose Configure, go to Java tab and type 1024M in the "Maximum memory pool" field.
6. Download MySQL Connector/J version 5.1 from http://dev.mysql.com/downloads. Connector/J is used by Java programs to communicate with MySQL database. Copy mysql-connector-java-x.x.x-bin.jar file into %CATALINA_HOME%\common\lib folder.
7. Download Weka 3.4.4 from http://sourceforge.net/projects/weka. Follow the link Download -> Browse all files to select the correct version. Download the zip archive version - weka-3-4-4.zip. Warning: NLP cell relies on this particular version of Weka, because the classification modes were created using this version of Weka. Recent versions of Weka seem to be non-backwards compatible with these models. Copy weka.jar file into the %CATALINA_HOME%\common\lib folder.
8. Download JDOM v1.0 from http://www.jdom.org. Copy jdom.jar into the %CATALINA_HOME%\common\lib folder. Note that NLP cell was not tested with the most recent version of JDOM (1.1).
9. Copy the following supporting jar files into the %CATALINA_HOME%\common\lib folder:
    a. **gate.jar** (from %GATE_HOME%\bin folder)
    b. **jasper-compiler-jdt.jar** (from %GATE_HOME%\lib folder)
    c. **ontotext.jar** (from %GATE_HOME%\lib folder)
    d. **hitex.jar** (from the 'core' folder of this this distribution)
    e. **ngram.jar** (from the supporting-jars.zip included with this distribution)
    f. **umls.jar** (from the supporting-jars.zip included with this distribution)
10. Restart the Tomcat.

4.2.7 Apache Axis2

Warning: do not perform the following steps if Apache Axis2 is already installed on the system, for example, if the core cells were previously installed.

1. Download Apache Axis2 version 1.1 from http://ws.apache.org/axis2/download/1_1/download.cgi. Choose WAR (Web Archive) distribution.
2. Copy axis2.war file into the %CATALINA_HOME%/webapps folder.

3. Restart Tomcat. The Axis2 web application will be installed. You can verify the successful installation by pointing your browser to http://hostname:port/axis2 page and following the “Validate” link. On successful installation you should see the Axis2 happiness page.

5. Installation

5.1 NLP Web Service Installation

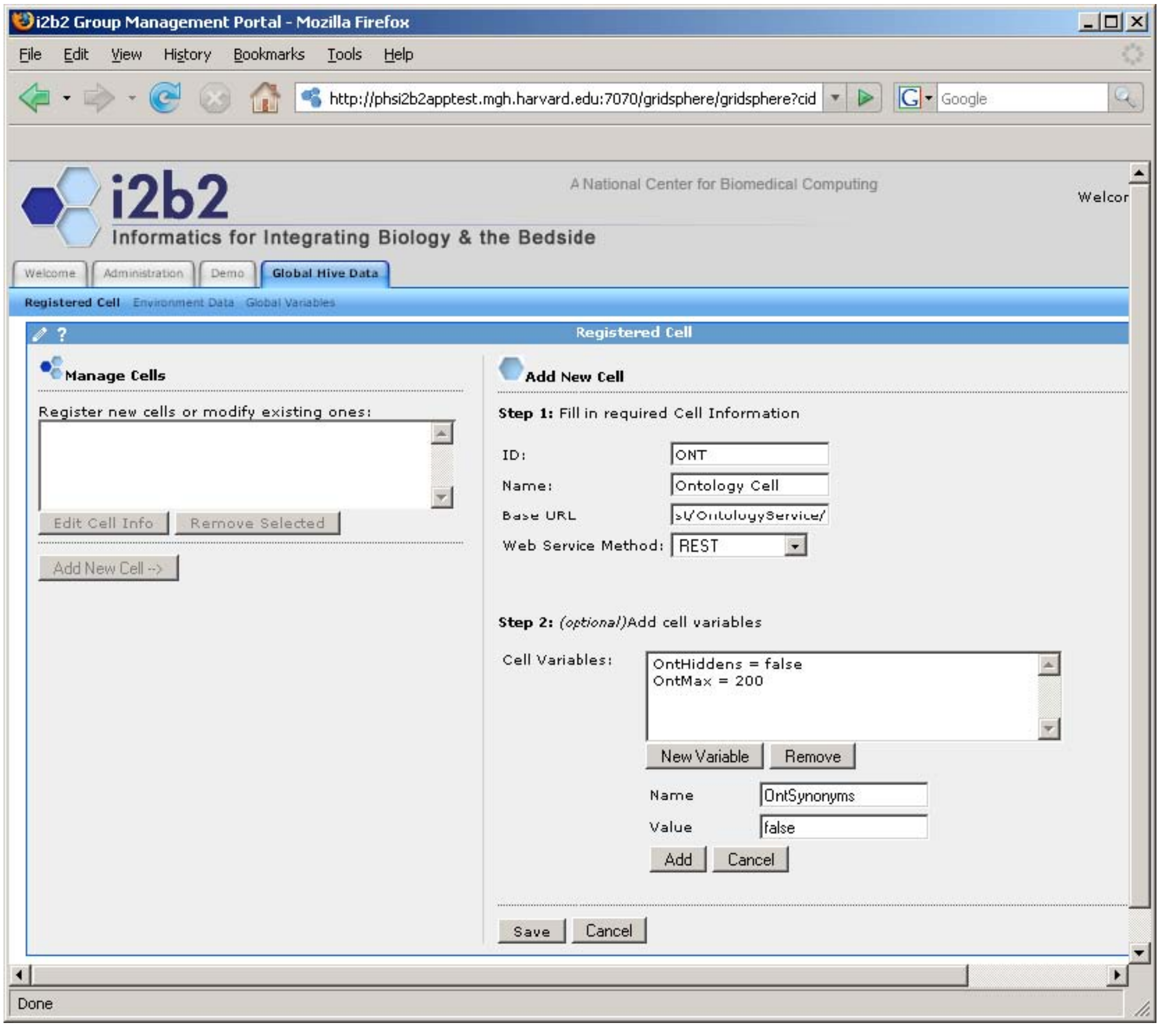
1. Shut down the Tomcat server.
2. Copy hitexservice.aar file into %CATALINA_HOME%\webapps\axis2\WEB-INF\service\ folder.
3. Extract the contents of /core/resources.zip into a location of your choice without spaces in the location name, preserving the directory structure.
4. Copy bundled config.xml into the /resources folder. This file contains configuration for the pre-assembled NLP pipelines.
5. Edit config.xml to adjust the configuration according to your environment. In particular, adjust all absolute URLs to point to valid locations in your environment, and database connection properties. Also, adjust the GATE_HOME variable.
6. Copy hitexservice.properties to %CATALINA_HOME% folder and modify hitex.config.url property to point to config.xml file inside the /resources folder.
7. Restart the Tomcat server.

5.2 NLP Cell Client Eclipse Plug-in Installation

1. Locate your I2B2 Workbench version 1.2 software's i2b2workbench directory, for example C:\i2b2workbenchv1.2\i2b2workbench.
2. Place the edu.harvard.i2b2.eclipse.plugins.nlp_1.0.0.jar file into the workbench's plug-ins directory, e.g., C:\i2b2workbenchv1.2\i2b2workbench\plugins.

5.3 Configure NLP Cell Information in Gridsphere

The detailed information on how to register a cell in Gridsphere can be found in the section 5 of the Project Management Cell Installation Guide. Please follow the instructions in this document, if NLP cell has not yet been configured. First, you'll need to register as a new user in the Gridsphere. Please refer to section 4 of Project Management Cell Installation Guide – "Creating Users". The user must have sufficient privileges to add a new cell. Next, select "Add new cell" option in the "Global Hive Data" section of the Gridsphere management page, and specify an ID, Name, Base URL and Web Service Method (e.g., REST) for the NLP cell. Below is a screenshot taken from the Project Management Cell Installation Guide that shows an addition of the Ontology Cell to Gridsphere. The addition of NLP cell is analogous.


i2b2 Group Management Portal - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://phsi2b2apptest.mgh.harvard.edu:7070/gridsphere/gridsphere?cid
Google
i2b2
Informatics for Integrating Biology & the Bedside
A National Center for Biomedical Computing
Welcor
Welcome
Administration
Demo
Global Hive Data
Registered Cell
Environment Data
Global Variables
Registered Cell
Manage Cells
Edit Cell Info
Remove Selected
Add New Cell -->
Add New Cell
New Variable
Remove
Add
Cancel
Save
Cancel
Done

Register new cells or modify existing ones:

Step 1: Fill in required Cell Information

ID: ONT

Name: Ontology Cell

Base URL: s/OntologyService/

Web Service Method: REST

Step 2: *(optional)*Add cell variables

Cell Variables:

```
OntHiddens = false
OntMax = 200
```

Name: OntSynonyms

Value: false

As a final step, enter the location of the Project Management Cell in the i2b2workbench.properties file inside the workbench directory:

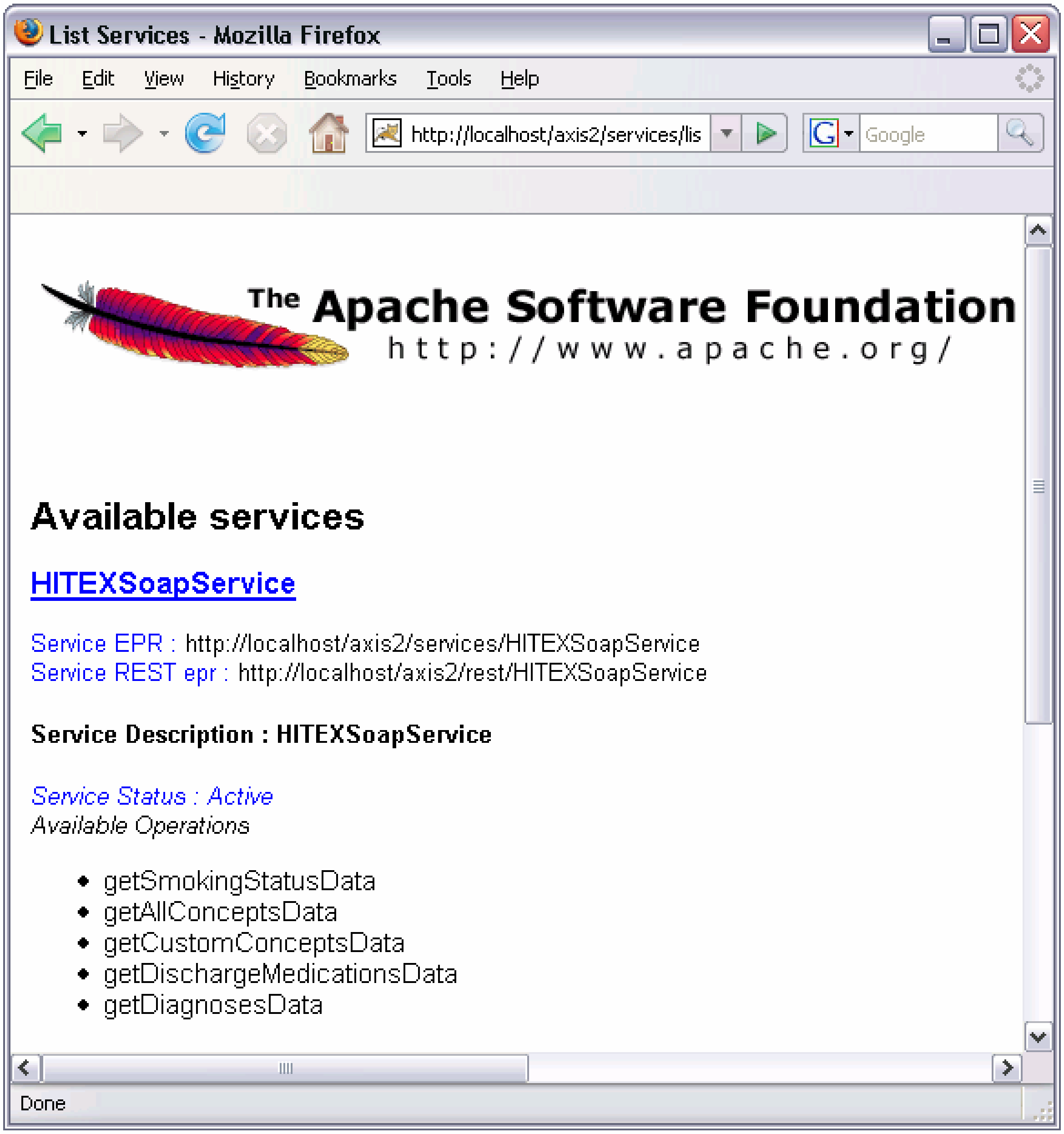
I2b2.2=Your Site Location,REST,http://tomcatHost:tomcatPort/axis2/rest/PMService/

For example:

i2b2.1=Harvard Demo,REST,http://webservices.i2b2.org/PM/rest/PMService/

6. Installation Verification

Go to http://hostname:port/axis2/services to verify that the NLP web service was installed. You should see a page similar to the following:


List Services - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
http://localhost/axis2/services/lis
Google
The Apache Software Foundation
http://www.apache.org/
Done

Available services

HITEXSoapService

Service EPR : http://localhost/axis2/services/HITEXSoapService
Service REST epr : http://localhost/axis2/rest/HITEXSoapService

Service Description : HITEXSoapService

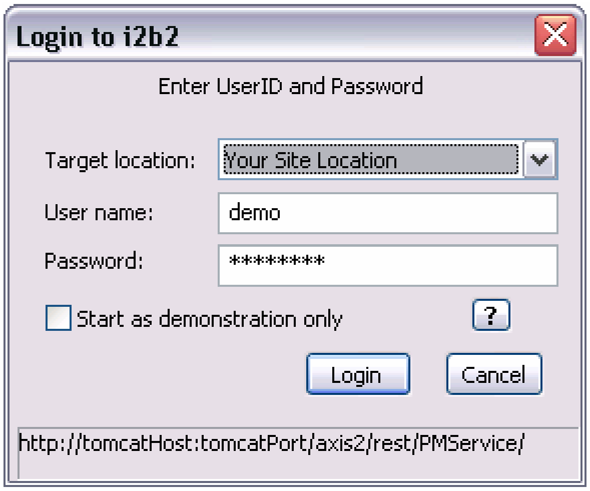
Service Status : Active
Available Operations

- getSmokingStatusData
- getAllConceptsData
- getCustomConceptsData
- getDischargeMedicationsData
- getDiagnosesData

Next, let's verify that the NLP cell is able to correctly process the requests. Launch the I2B2 Workbench (double-click on i2b2Workbench.exe). Login to I2B2:

- Select your target location (the location of Project Management Cell)

- Enter a valid username and password that you set up in the Gridsphere, e.g., demo/demouser.
- Click on Login button.



Target location: Your Site Location

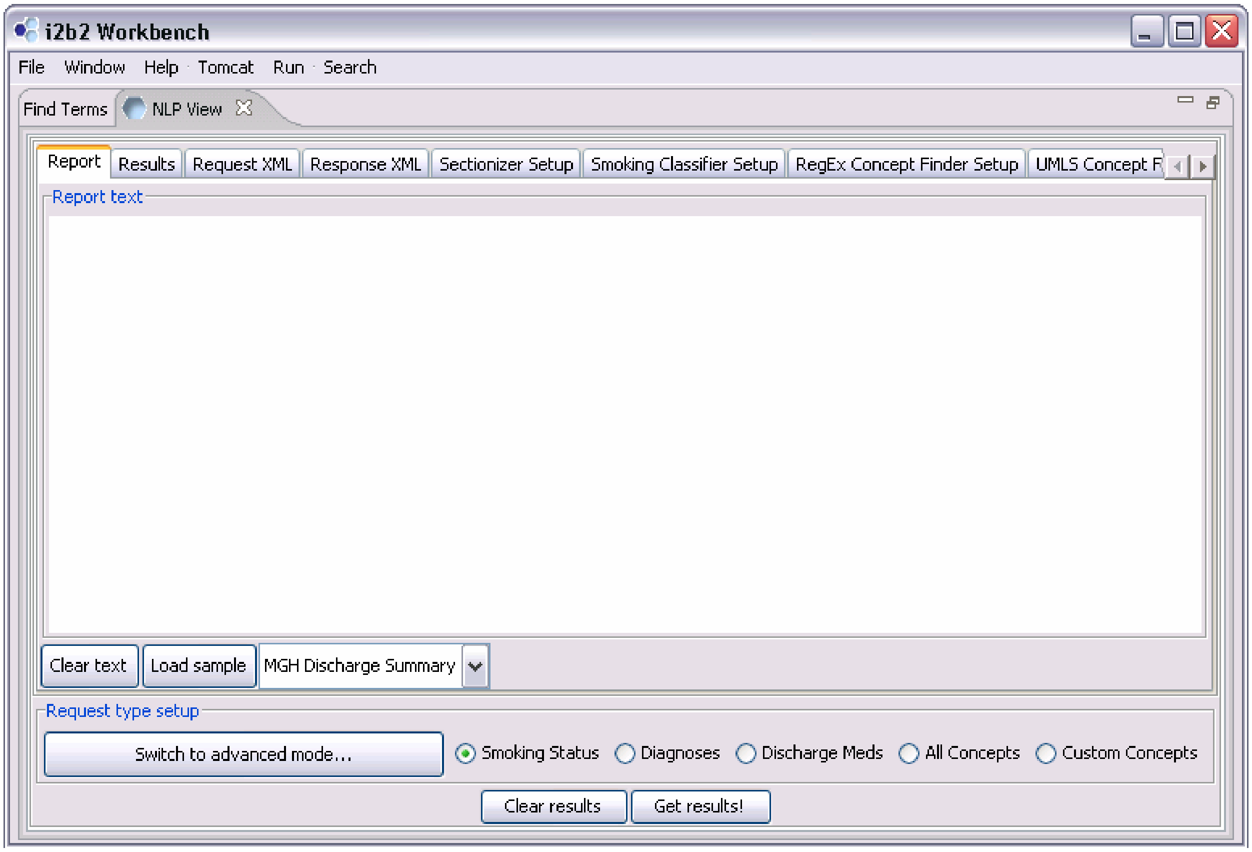
User name: demo

Password: ********

□ Start as demonstration only

http://tomcatHost:tomcatPort/axis2/rest/PMService/

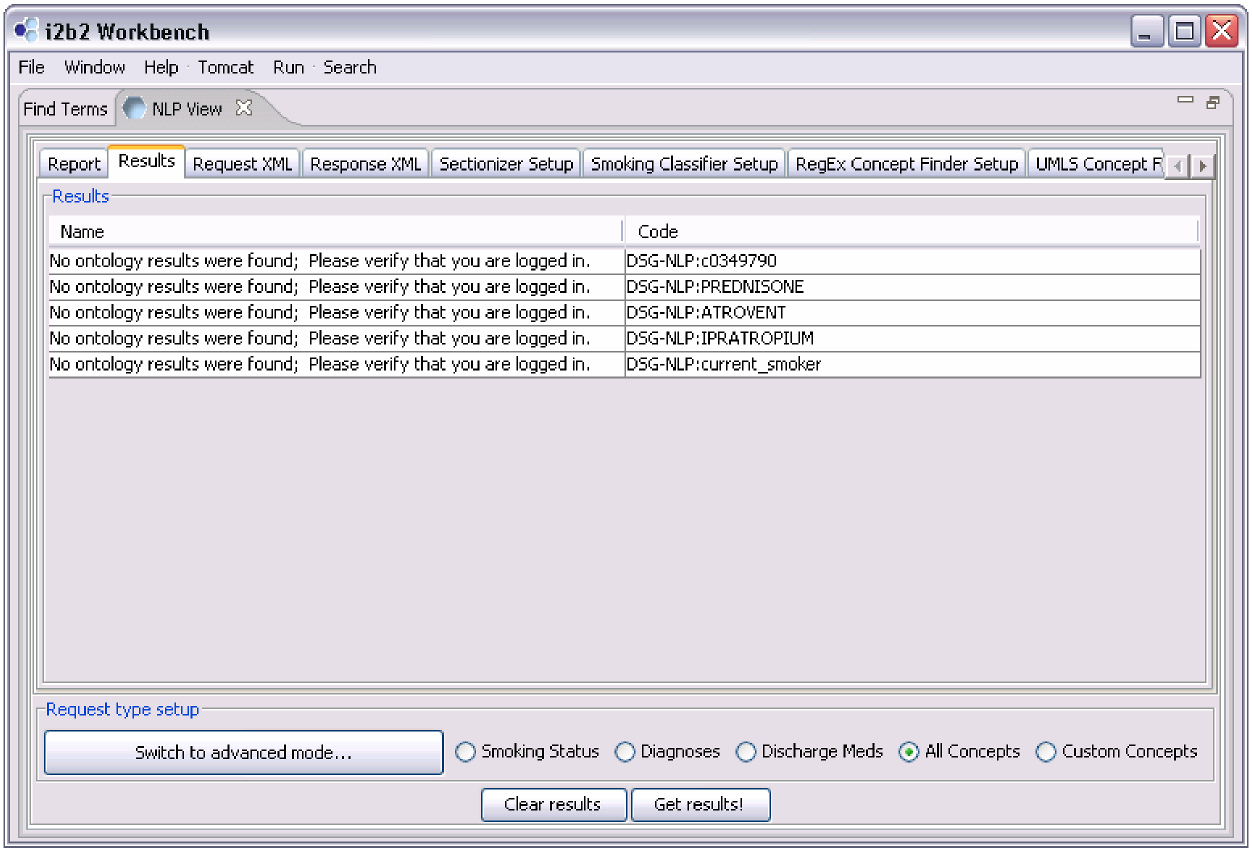
After the Workbench window opens, enable the NLP View. For that, Go to Window -> Show View -> Other… -> NLP Category and select “NLP View”. You’ll be presented with the following screen:


i2b2 Workbench
File Window Help Tomcat Run Search
Find Terms
NLP View
Report Results Request XML Response XML Sectionizer Setup Smoking Classifier Setup RegEx Concept Finder Setup UMLS Concept F
Report text
Clear text
Load sample
Request type setup
Switch to advanced mode…
Clear results
Get results!

MGH Discharge Summary

● Smoking Status ○ Diagnoses ○ Discharge Meds ○ All Concepts ○ Custom Concepts

Click on “Load Sample” button and select the “All Concepts” radio button. Next, click on “Get Results” button. In a few moments, the results of the web service call should appear in the “Results” tab:

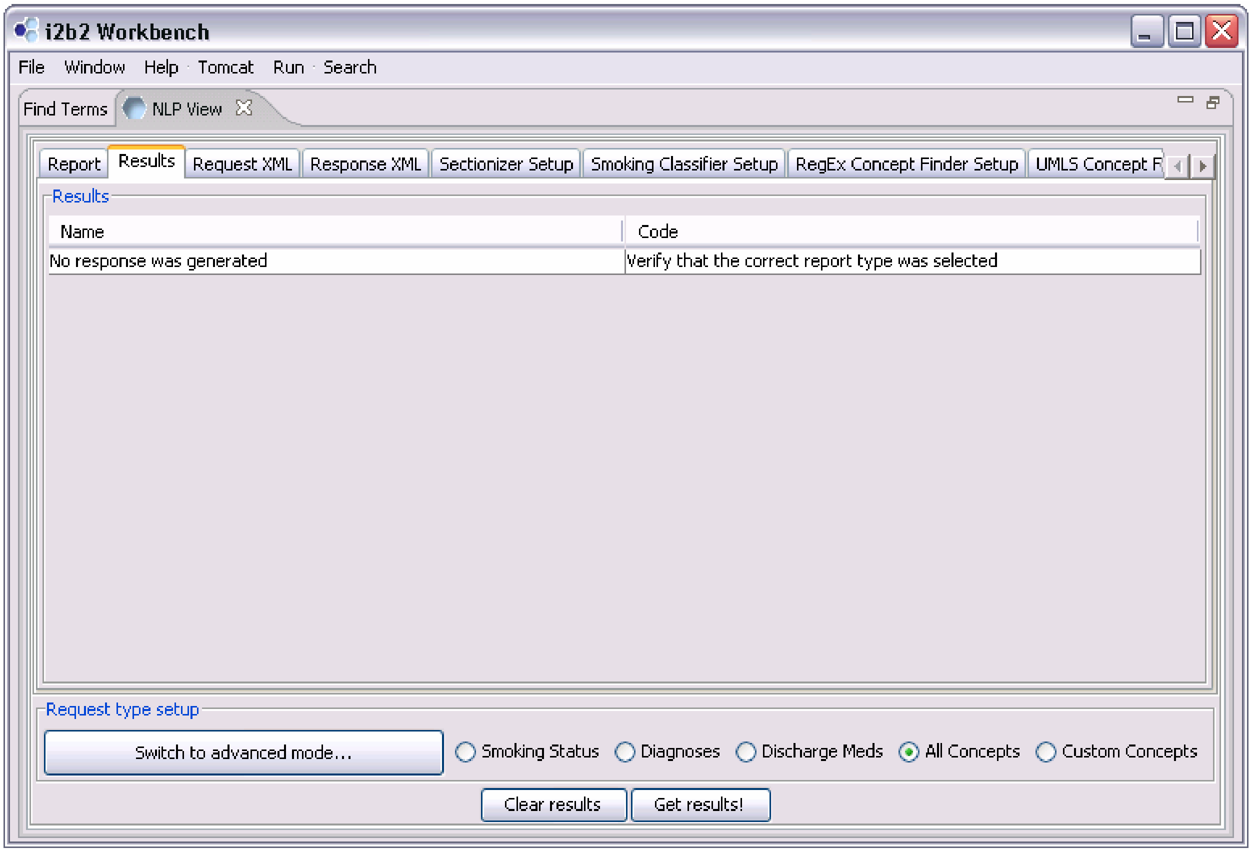


Results

Name	Code
No ontology results were found; Please verify that you are logged in.	DSG-NLP:c0349790
No ontology results were found; Please verify that you are logged in.	DSG-NLP:PREDNISONE
No ontology results were found; Please verify that you are logged in.	DSG-NLP:ATROVENT
No ontology results were found; Please verify that you are logged in.	DSG-NLP:IPRATROPIUM
No ontology results were found; Please verify that you are logged in.	DSG-NLP:current_smoker

○ Smoking Status ○ Diagnoses ○ Discharge Meds ● All Concepts ○ Custom Concepts

If error has occurred, no results will be returned. Instead, a warning message will be returned:



Results

Name	Code
No response was generated	Verify that the correct report type was selected

Request type setup

○ Smoking Status ○ Diagnoses ○ Discharge Meds ● All Concepts ○ Custom Concepts

There may be several reasons for this error:

1. Remote server may be unavailable

Server may be down. Restart the server.

2. NLP cell may be unavailable

Check the http://hostname:port/axis2/services to see if HITEXSOAPService is listed as active. Check Tomcat log files to see any error messages.

3. Database error may have occurred

Check Tomcat log files to see if there are any database-related errors. Common database problems include: database server is not running, database schema does not exist, database user is not given sufficient privileges to access the database schema, database user is not authenticated, database schema doesn't contain all required tables or may be corrupted, database cannot be accessed due to firewall problems or other communication link problems, Connector/J is not available in %CATALINA_HOME%\common\lib folder.

7. License

The I2B2 source code is licensed under the I2B2 Software License Software. This includes but not limited to all code in the edu.harvard.i2b2.* package namespace.