

Design and Engineering of Intelligent
Information Systems

Execution Architecture with CPE and Deployment
Architecture with UIMA-AS

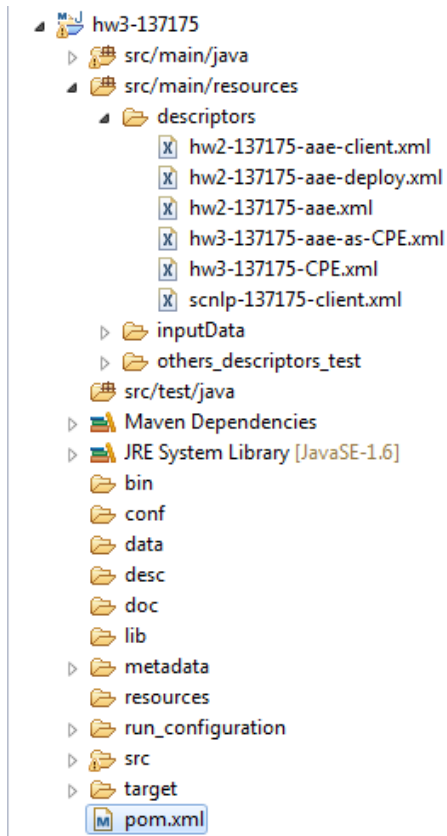
hw3-137175

Tania Patiño

28-03-2014

Organization for this project is the next one:

```
hw3-ID
|- pom.xml
\-- src
    \-- main
        |- java
        |   \-- **/*.java      /* Java classes generated by JCasGen
        |                       and your UIMA annotators          */
        \-- resources
            |- hw2-ID-aae.xml      /* your aggregate analysis engine of homework 2*/
            |- hw2-ID-aae-client.xml /* clinet descriptor of your AAE */
            |- hw2-ID-aae-deploy.xml /* deployment descriptor of your AAE */
            |- hw3-ID-aae-as-CPE.xml /* CPE descriptor to test your AAE service */
            |- hw3-ID-CPE.xml      /* CPE descriptor of your homework 2 pipeline */
            |- scnlp-ID-client.xml /* clinet descriptor for the remote UIMA-AS service
            | \-- **/*.*          /* analysis engine and other resources */
            \-- docs
                \-- hw3-ID-report.pdf /* your report for design */
```



On the folder others_descriptions_test, there are some files that were tested when the experimentation of the environment UIMA-Eclipse and UIMA-AS was done.

In this task, was created a Collection Processing Engine (CPE) and also the task was to understand how to use it. Requirements were, run the pipeline with a CPE instead of the UIMA Document Analyzer.

Task 1.1 Learning CPE

Resources CPE

Basic concepts and usage about CPE from *Chapter 2. Collection Processing Engine Developer's Guide* (http://uima.apache.org/d/uimaj-2.4.0/tutorials_and_users_guides.html#ugr.tug.cpe).

Manual for CPE GUI (<http://uima.apache.org/d/uimaj-2.4.0/tools.html#ugr.tools.cpe>).

The acronym CPE, corresponds to a *Collection Processing Engine*, which one illustrates the management of the data flow that goes between different types of components that make up a CPE. Each component form part of this data flow. There are many components, for example:

1. **CAS Initializer:** is a component to populate a CAS from a document. This one is focus on the population of a document, how a document could be filled by data and what kind of data.

2. **CAS Consumer:** there could be more than one component like this one, the main idea about this, is to consume the enriched CAS that was generated by the sequence of Analysis Engines. Types of CAS Consumers are search engine like Google, index (this one is related to how to index the data in a database or what?) or a database like Oracle.

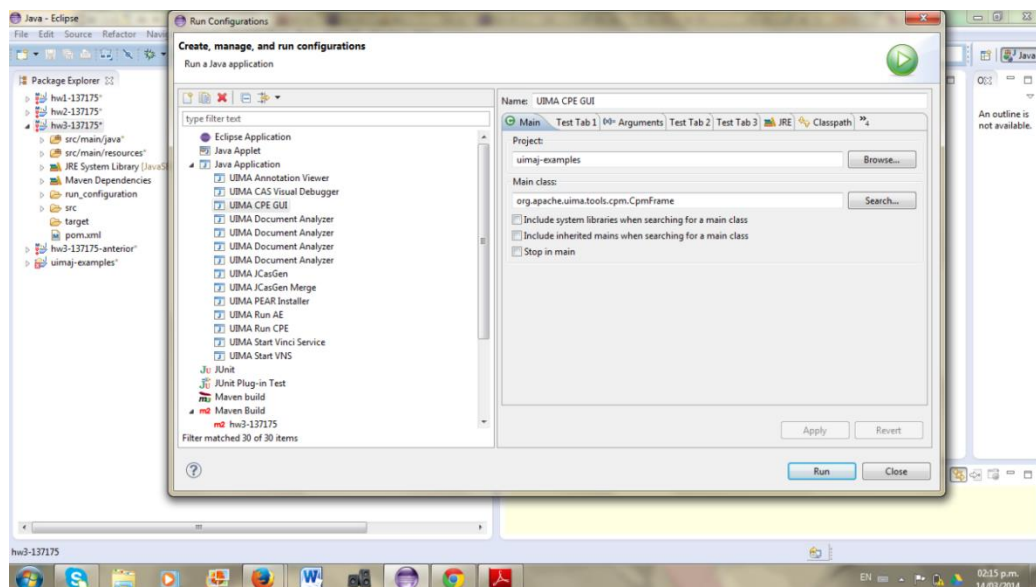
3. **Collection Reader:** interfaces to a collection of documents that should be analyzed. Maybe this is related to XML files and the analysis about every structure of the data like tags <building></building>, labels and keywords, etc.

4. **Analysis Engine:** this one takes a CAS analyses takes its contents and produce enriched CAS.

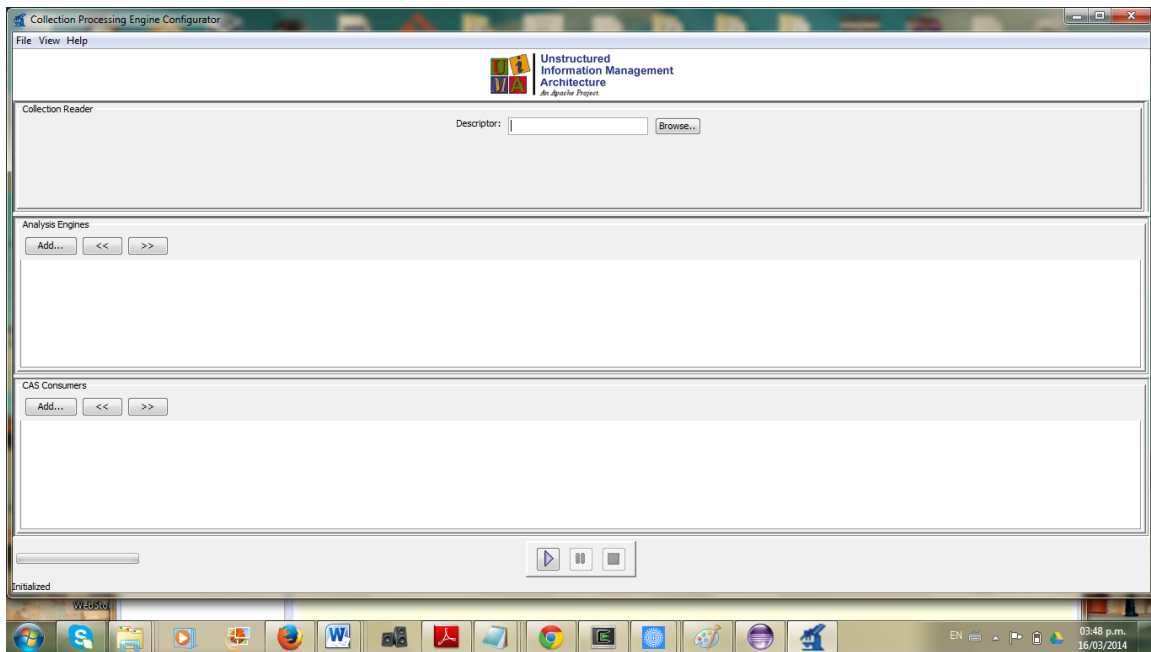
Secondly, following the next link related to how to make a CPE step by step, and also there is a explanation about exploration and applications of this tool: <http://uima.apache.org/d/uimaj-2.4.0/tools.html#ugr.tools.cde>

In the next description, there is an explanation related to the CPE Configuration steps and CAS viewer.

1. Open Eclipse-IDE then go to Menu option ->
2. Run -> Run Configurations -> Select UIMA CPE GUI, then click on Run

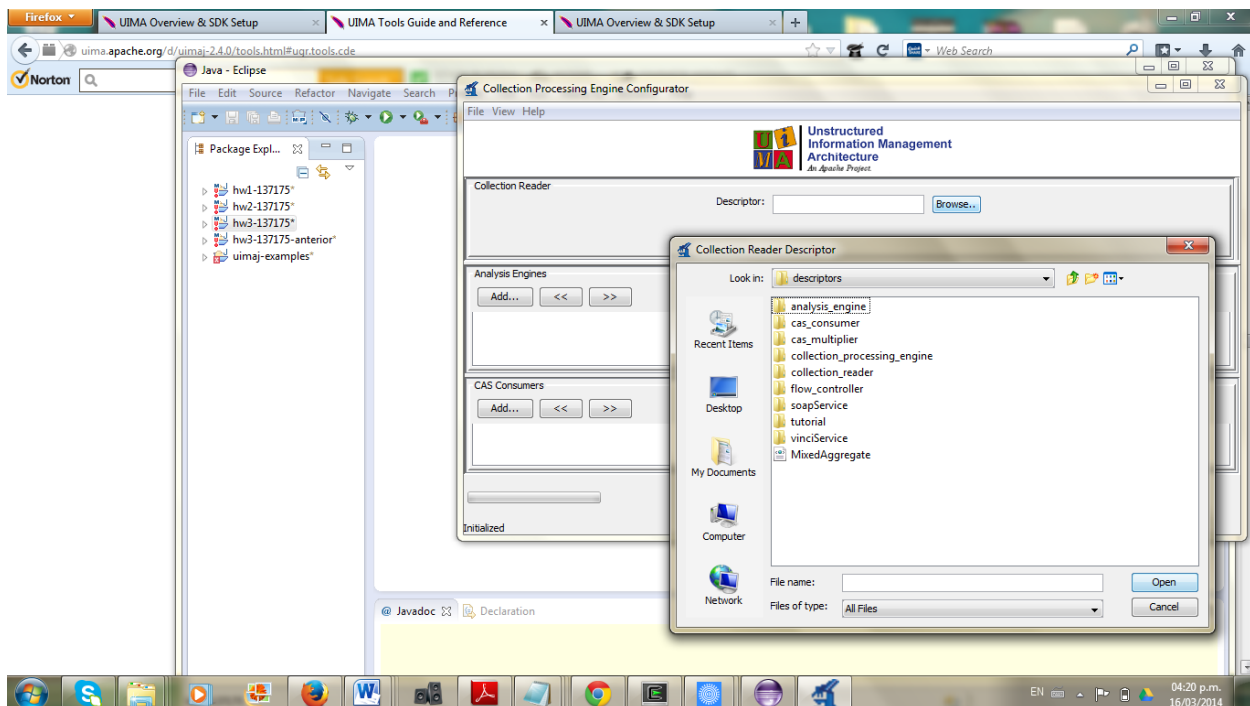


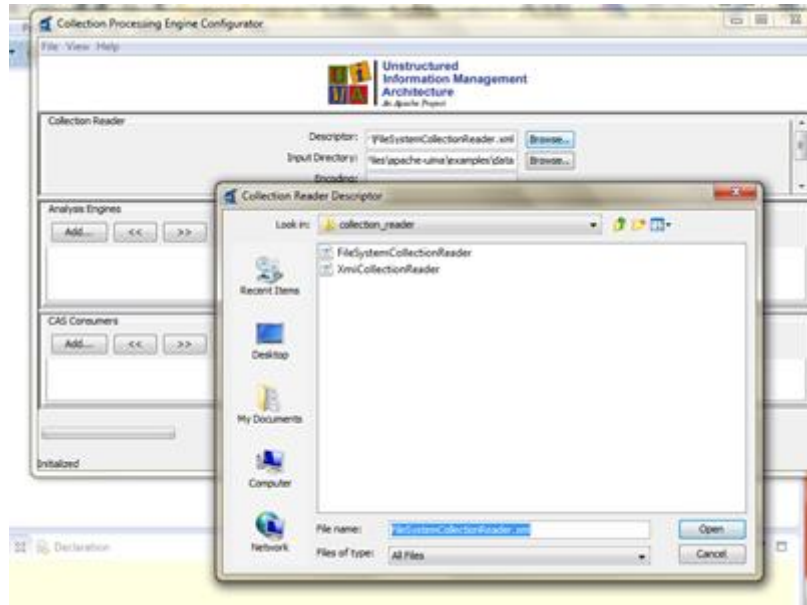
Then the next window related to Collection Processing Engine Configurator is open, where a descriptor have to be select the descriptor and more options are activated.



Then making some experiments with the files that come by default on the descriptors folder. Finally make a selection following this path:

Selecting Collection Reader Descriptor-> collection_reader->FileSystemCollectionReader.xml





The collection reader needs to be general enough to establish a connection to the file source and open a stream to read the content. Creating a `org.apache.uima.tools.components.FileSystemCollectionReader` directly, and only need to write a Collection Reader descriptor.

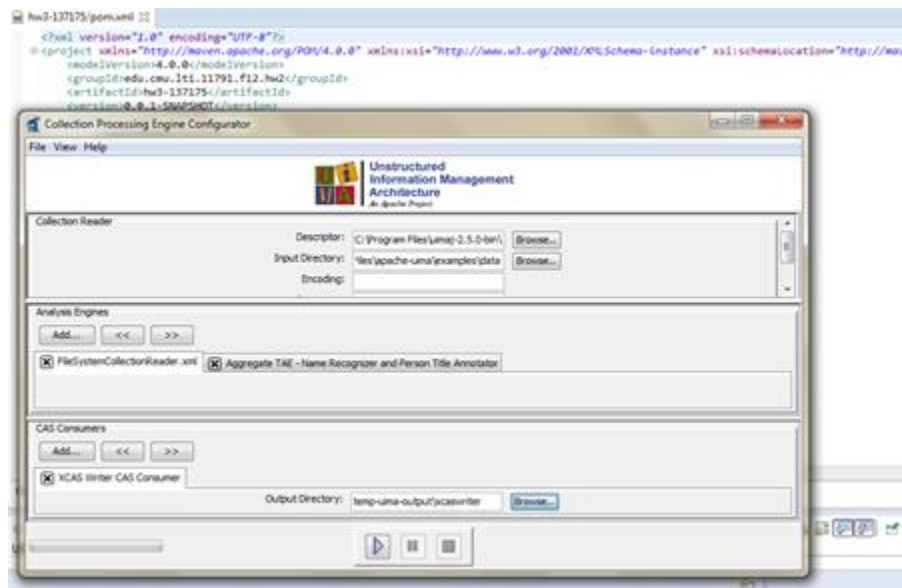
Descriptor: `C:\Program Files\ujm-2.5.0-bin\apache-uima\examples\descriptors\collection_reader\FileSystemCollectionReader.xml`

Input Directory: `C:\Program Files\apache-uima\examples\data`

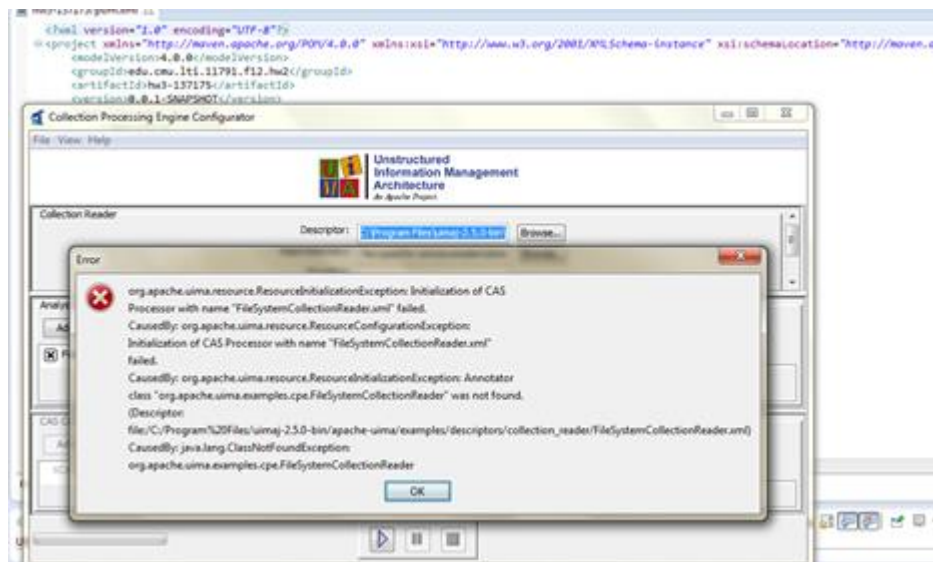
Analysis Engines:

Aggregate TAE- Name Recognizer and Person Title Annotator

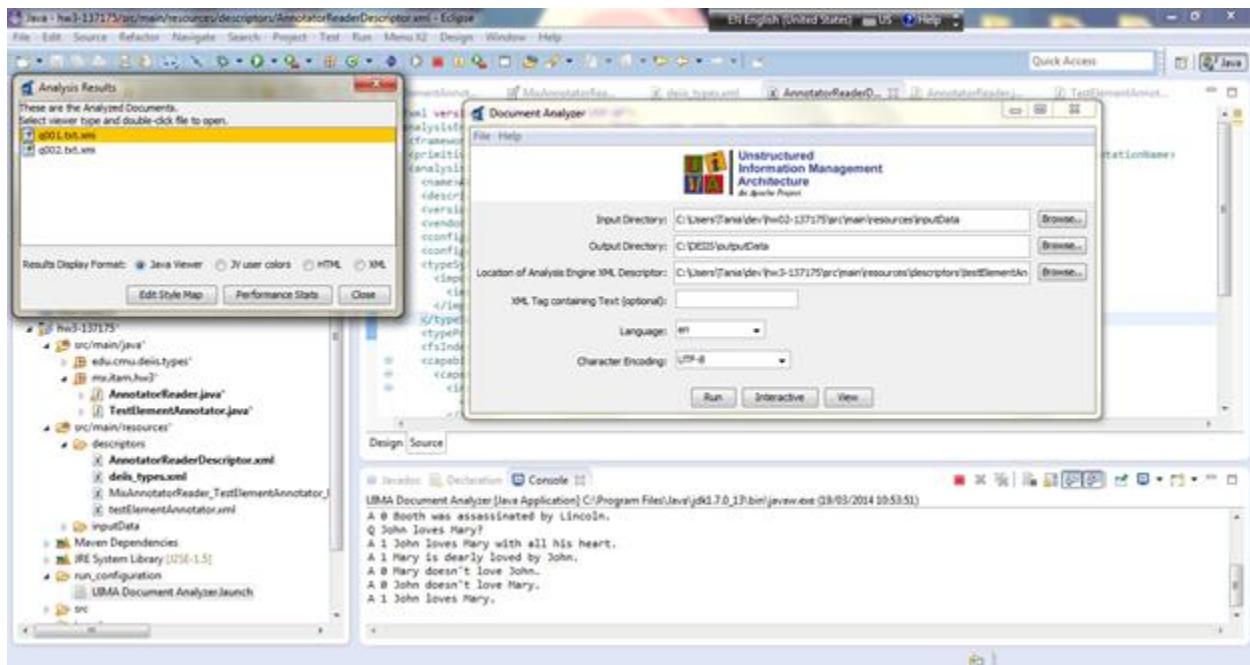
Cas Consumer: Xcas Writer Cas Consumer



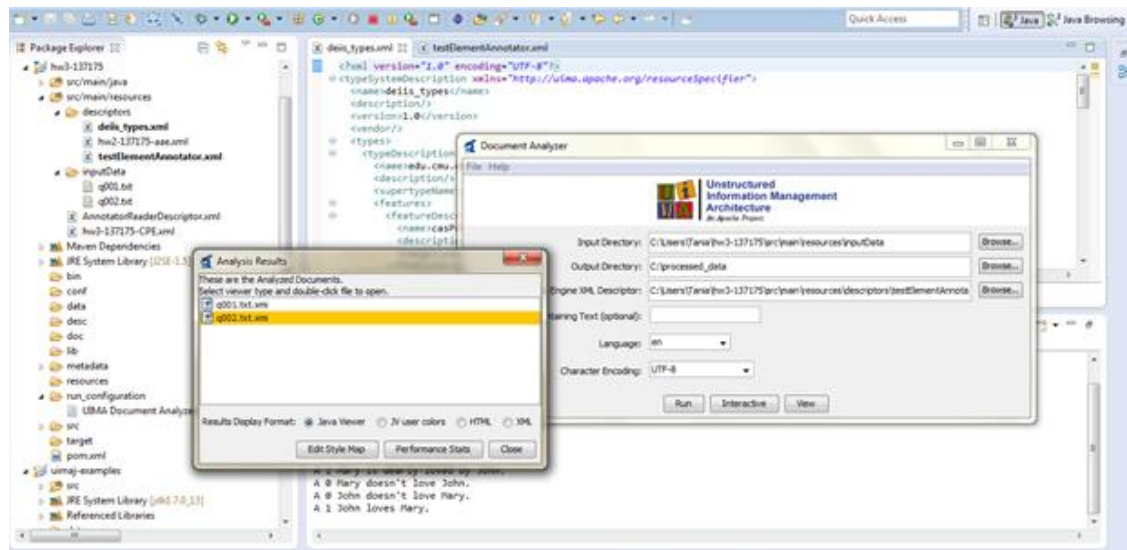
Following the first steps, appears many issues, some were created because the wrong configuration paths or the wrong calling of a file for a descriptor option.



Document Analyzer with UIMA making test and experiments that ran all right:



With the Document Analyzer with UIMA i made tests and experiments, that ran all right: testElementAnnotator.xml example was running fine and after some analysis of some files that came by default like q001.txt.xml and q002.txt.xml the entire test works.



Task 1.2 Creating and Running your CPE (25 pts)

- (10 pts) The collection reader needs to be general enough to establish a connection to the file source and open a stream to read the content.
- Task, consider a folder of files located on the file system, where apply `org.apache.uima.tools.components.FileSystemCollectionReader` directly, and write a Collection Reader descriptor appropried.

Some lines about the Collection Reader Descriptor that was created (CollectionReaderDescriptor.xml):

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Task 1
Section 1. Consider the folder of files located on the file system, wich means you can use:
org.apache.uima.tools.components.FileSystemCollectionReader.
Then write a Collection Reader Descriptor to fit the needs -->
<collectionReaderDescription xmlns="http://uima.apache.org/resourceSpecifier">
  <frameworkImplementation>org.apache.uima.java</frameworkImplementation>
  <implementationName>org.apache.uima.examples.cpe.FileSystemCollectionReader</implementationName>
  <processingResourceMetaData>
    <name>FileSystemCollectionReaderDescriptor</name>
    <description>FileSystem Collection Reader Descriptor that uses or not CAS Initializer.</description>
    <version>1.0</version>
    <vendor>Apache Enterprise Foundation</vendor>
    <configurationParameters searchStrategy="none">
      <configurationParameter>
        <name>InputDirectory</name>
        <description>this directory contains inputs files</description>
        <type>String</type>
        <multiValued>false</multiValued>
        <mandatory>true</mandatory>
      </configurationParameter>
      <configurationParameter>
        <name>Encoding</name>
        <description>Encoding parameter. The idea is that CAS is responsible to deal with character
encoding issues.</description>
        <type>String</type>
        <multiValued>false</multiValued>
        <mandatory>false</mandatory>
      </configurationParameter>
    </configurationParameters>
  </processingResourceMetaData>
</collectionReaderDescription>
```



```

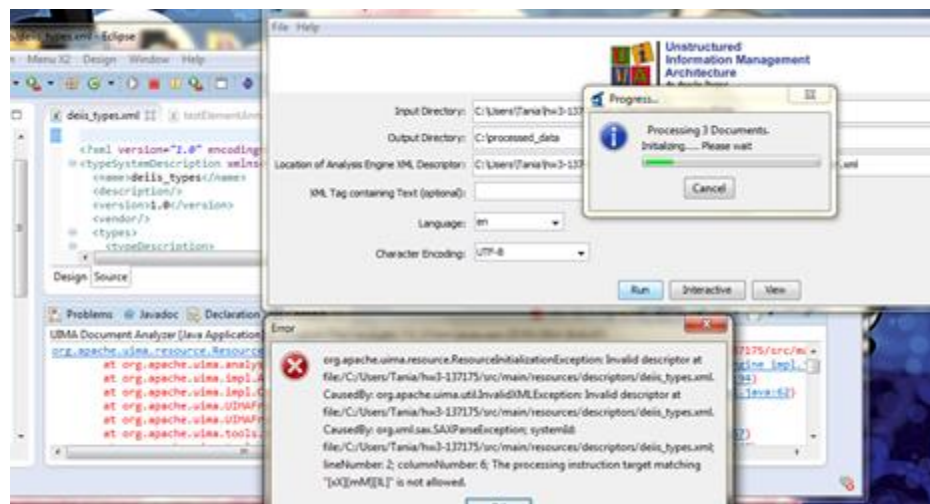
<configurationParameter>
  <name>InputDir</name>
  <description>C:\Users\Tania\hw3-137175\src\main\resources\inputData</description>

```

Note: the Java class called (FileSystemCollectionReader.java), was implemented and refactored and it is related to FileSystemCollectionReader which analyze the file CollectionReaderDescriptor.xml

Node	Content
?? xml	version="1.0" encoding="UTF-8"
!--	Task 1 Section 1. Consider the folder of files located on the file system, wich means y...
collectionReaderDescription	
xmlns	http://uima.apache.org/resourceSpecifier
frameworkImplementation	org.apache.uima.java
implementationName	org.apache.uima.examples.cpe.FileSystemCollectionReader
processingResourceMetaData	
resourceManagerConfiguration	

Error message that appears, this one occurred because there was a white space at the beginning of the xml code.



Error message: org.apache.uima.ResourceInitializationException: Invalid description at file c:/users/Tania/hw3-137175/src/main/resources/descriptors/deis_types.xml
Caused by: org.xml.sax.SAXParseException.

Delete White space and runs all right!!

3. (10 pts) You are required to create a Cas Consumer based on the Evaluator component of homework 2, and include it in your CPE pipeline.

```
<?xml version="1.0" encoding="UTF-8"?>
<casConsumerDescription xmlns="http://uima.apache.org/resourceSpecifier">
  <frameworkImplementation>org.apache.uima.java</frameworkImplementation>

  <implementationName>edu.cmu.deiis.types.FileSystemCollectionReader</implementationName>
  <processingResourceMetaData>
    <name>casConsumerDescriptor</name>
    <description>This is related to print answerScores</description>
    <version>1.0</version>
    <vendor>Apache Enterprise Foundation</vendor>
    <configurationParameters>
      <configurationParameter>
        <name>Outputdirectory</name>
        <description>This is the directory where XMI files should be located</description>
        <type>String</type>
```

Node	Content
?? xml	version="1.0" encoding="UTF-8"
casConsumerDescription	
xmlns	http://uima.apache.org/resourceSpecifier
frameworkImplementation	org.apache.uima.java
implementationName	edu.cmu.deiis.types.FileSystemCollectionReader
processingResourceMetaData	
name	casConsumerDescriptor
description	This is related to print answerScores
version	1.0
vendor	Apache Enterprise Foundation
configurationParameters	
configurationParameter	
configurationParameterSettings	
typeSystemDescription	
capabilities	
operationalProperties	
resourceManagerConfiguration	

4. (5 pts) Please name your CPE descriptor as hw3-ID-CPE.xml and put it under src/main/resources/, so that we could easily find the entry point of your pipeline.

Node	Content
?? xml	version="1.0" encoding="UTF-8"
cpeDescription	
xmlns	http://uima.apache.org/resourceSpecifier
collectionReader	
casProcessors	
casProcessor	
cpeConfig	Configuration CPE

```
<?xml version="1.0" encoding="UTF-8"?>
<cpeDescription xmlns="http://uima.apache.org/resourceSpecifier">
  <collectionReader>
    <collectionIterator>
      <descriptor>
        <import location="../../../../../../Program Files/uimaj-2.5.0-bin/apache-
uima/examples/descriptors/collection_reader/FileSystemCollectionReader.xml"/>
      </descriptor>
    </collectionIterator>
    <configurationParameterSettings>
      <nameValuePair>
```



```

        <name>InputDirectory</name>
        <value>
            <string>C:\</string>
        </value>
    </nameValuePair>
</configurationParameterSettings>
</collectionIterator>
</collectionReader>

<casProcessors casPoolSize="3" processingUnitThreadCount="1"></casProcessors>

    <casProcessor name="hw2-137175-aae" deployment="integrated">
        <descriptor>
            <import location="/hw3-137175/src/main/resources/hw2-137175-aae.xml"/>
        </descriptor>
        <deploymentParameters/>
    </casProcessor>
</casProcessors>

```

Task 2

Deployment Architecture with UIMA-AS

Integrate a remote UIMA-AS service (Stanford CoreNLP) into the CPE pipeline, and deploy the aggregate analysis engine in homework 2 as an UIMA-AS service.

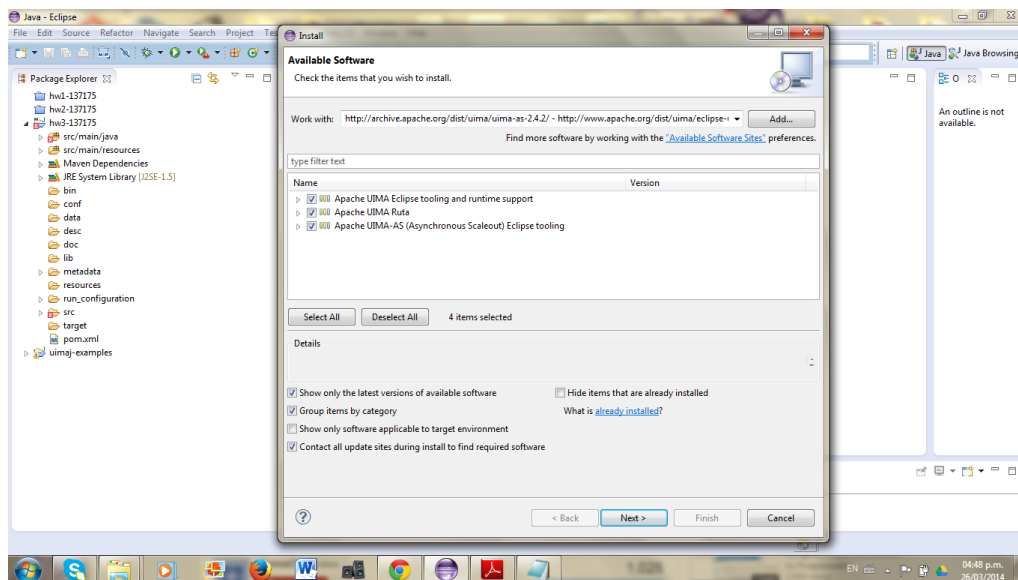
Task 2.1 Learning UIMA-AS,

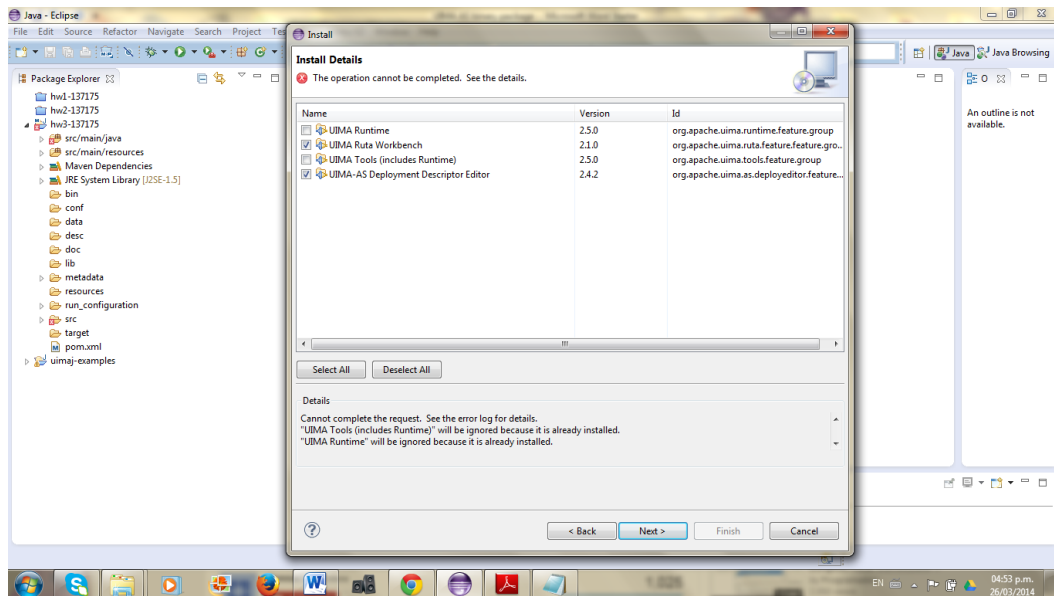
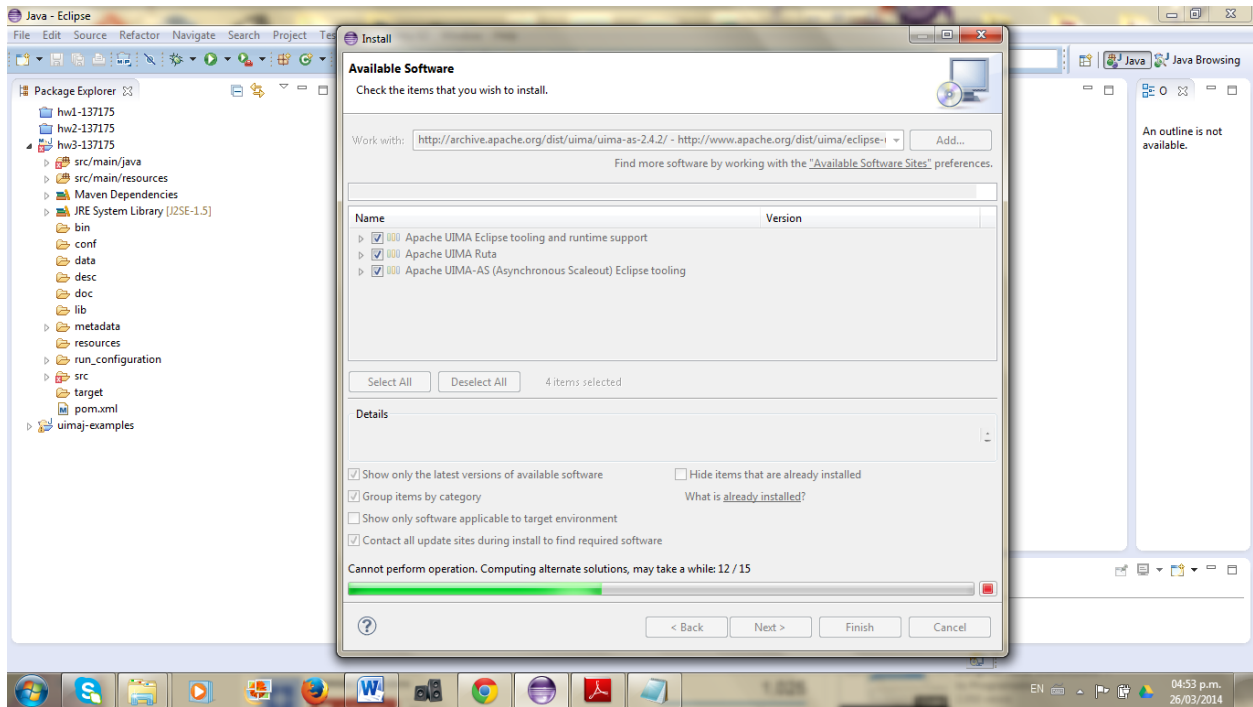
I read about the information that came on each link, the first one describes the concepts about UIMA-AS the other two makes a description about UIMA-AS and how to apply it for making some experiments with data.

Task 2.2 Creating an UIMA-AS client (25 pts)

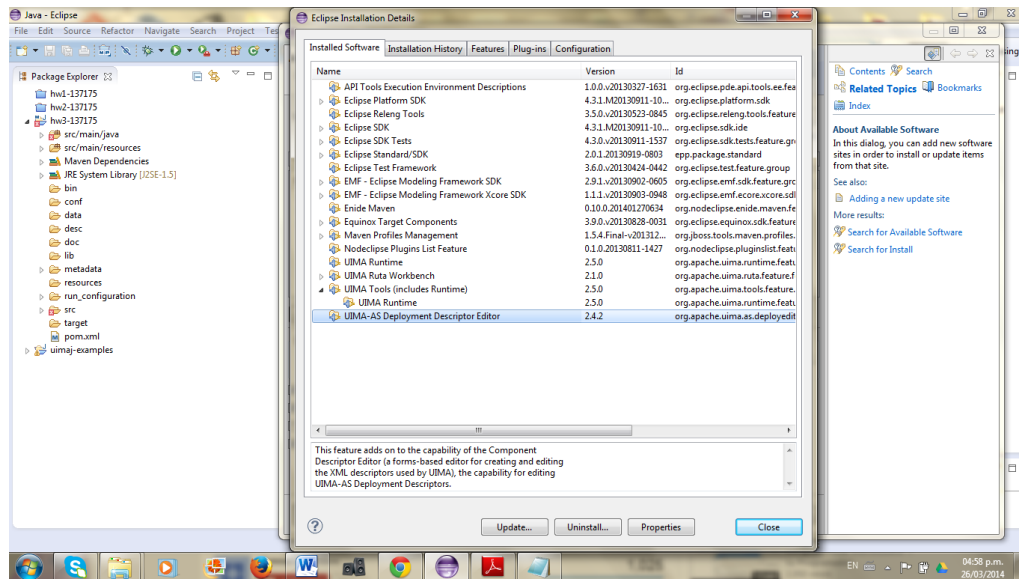
Checking that UIMA-AS was installed on Eclipse graphically by the Available Software Tool:

1. UIMA AS binary package installation.



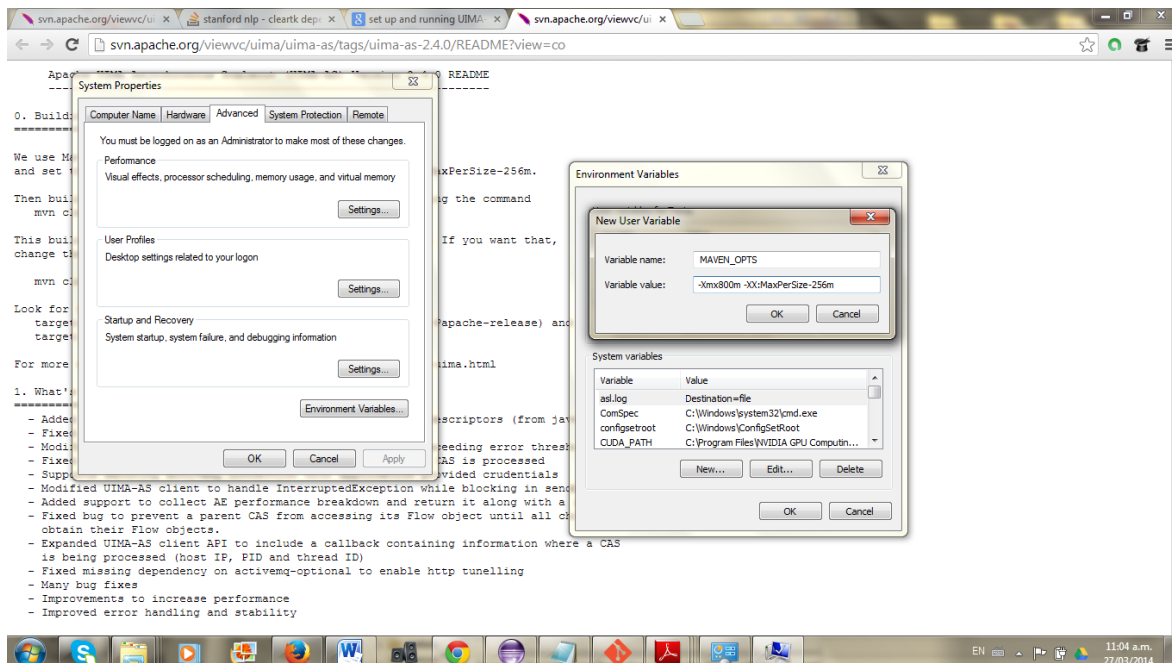


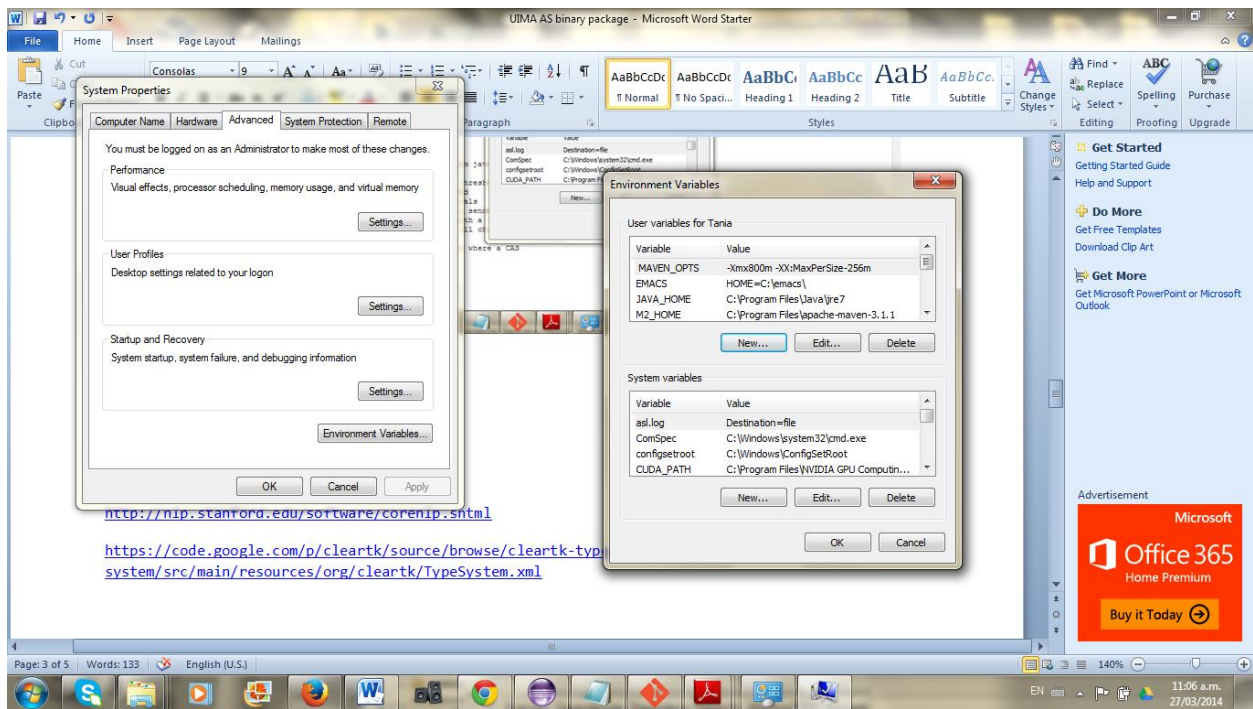
Finally, i realise that UIMA AS was installed in another moment, so is ready, UIMA AS version is 2.4.2



To configure UIMA-AS was the next one that is explained step by step in the link: <http://svn.apache.org/viewvc/uima/uima-as/tags/uima-as-2.4.0/README?view=co>, following the steps that are explained on README about UIMA-as.

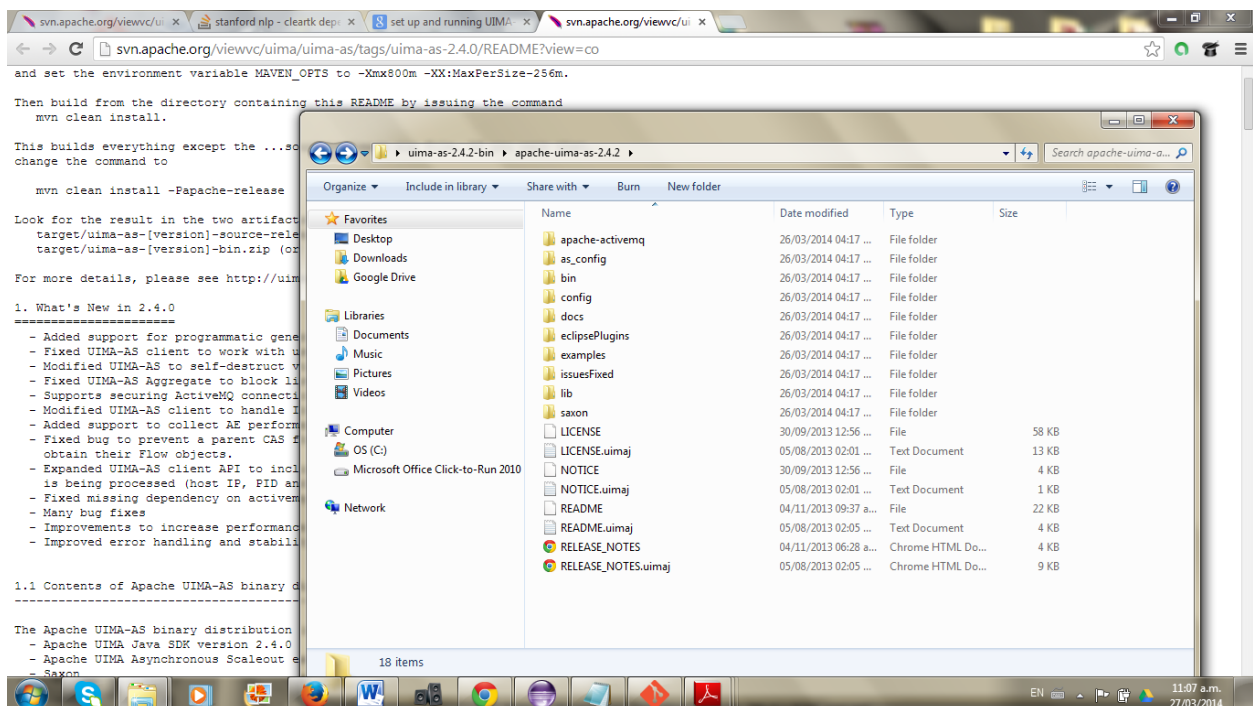
1. Set the environment variable MAVEN_OPTS to -Xmx800m -XX:MaxPerSize=256m.





Then build from the directory containing this README by using the command for MAVEN environment.

➤ `mvn clean install`



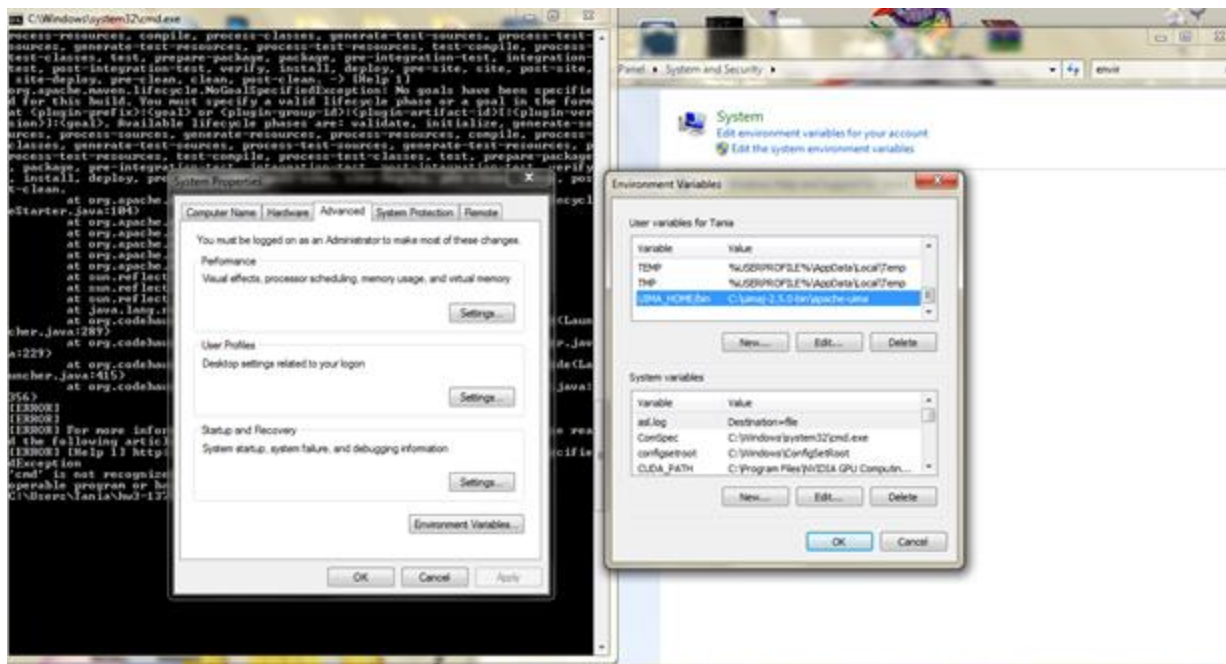
Many problems appears. issue: cmd is not recognized as an internal or external command. searching on internet then one option is to reinstall Java, i did that and magically all workjust fine!!

```

C:\Windows\system32\cmd.exe
C:\Users\Tania\hw3-137175>run clean install
[INFO] Scanning for projects...
[WARNING] Some problems were encountered while building the effective model for
edu.cmu.1c1:13791:12:hw3:hw3-137175:jar:0.0.1-SNAPSHOT
[WARNING] 'build.plugins.plugin.version' for org.apache.maven.plugins:maven-comp
iler-plugin is missing. @ line 27, column 12
[WARNING] It is highly recommended to fix these problems because they threaten t
he stability of your build.
[WARNING] For this reason, future Maven versions might no longer support buildin
g such malformed projects.
[WARNING]
[INFO] Building hw3-137175 0.0.1-SNAPSHOT
[INFO]
[INFO] --- maven-clean-plugin:2.5:clean (default-clean) @ hw3-137175 ---
Deleting C:\Users\Tania\hw3-137175\target
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @ hw3-137175
[INFO] Using platform encoding (Cp1252 actually) to copy filtered resources,
i.e. build is platform dependent!
[INFO] Copying 11 resources
[INFO] --- maven-compiler-plugin:2.5.1:compile (default-compile) @ hw3-137175 ---
[WARNING] File encoding has not been set, using platform encoding Cp1252, i.e. b
uild is platform dependent!
[INFO] Compiling 21 source files to C:\Users\Tania\hw3-137175\target\classes
[INFO] BUILD FAILURE
[INFO] Total time: 2.316s
[INFO] Finished at: Thu Mar 27 21:28:35 PDT 2014
[INFO] Final Memory: 15M/219M
[INFO]
[ERROR] Failed to execute goal org.apache.maven.plugins:maven-compiler-plugin:2.
5.1:compile (default-compile) on project hw3-137175: Fatal error compiling: too
l.jar not found: C:\Program Files\Java\jre7\..lib\tools.jar -> [Help 1]
[ERROR]
[ERROR] To see the full stack trace of the errors, re-run Maven with the -e swit
ch.
[ERROR] Re-run Maven using the -X switch to enable full debug logging.
[ERROR]
[ERROR] For more information about the errors and possible solutions, please rea
d the following articles:
[ERROR] [Help 1] http://wiki.apache.org/confluence/display/MAVEN/MojoExecutionE
xception
'cmd' is not recognized as an internal or external command,
operable program or batch file.
C:\Users\Tania\hw3-137175>run --version
Apache Maven 3.1.1 (0728685237757fbbf44136ace040295f723d9a; 2013-09-17 08:22:2
2-0700)

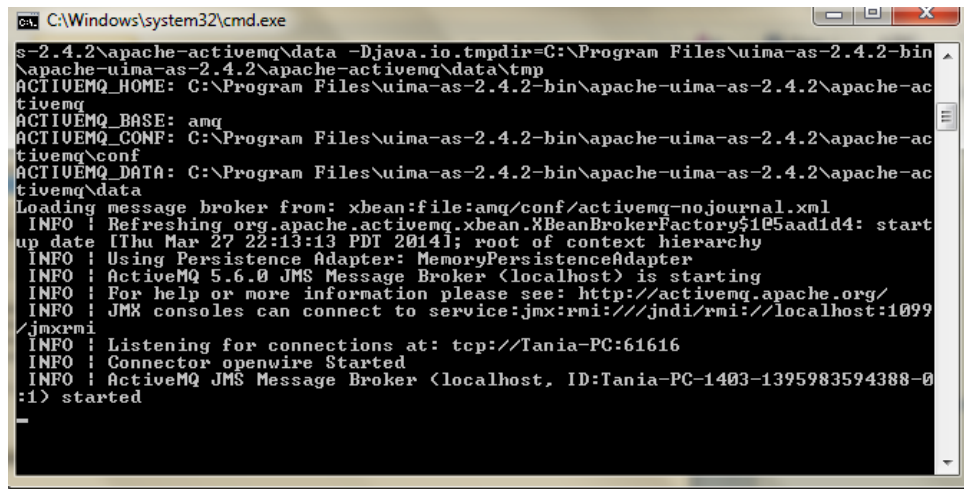
```

UIMA_HOME variables



In the next path is located the UIMA-AS, C:\Users\Tania\apache-uima-as-2.4.2\bin

Running the file \$startBroker, then appears this execution.



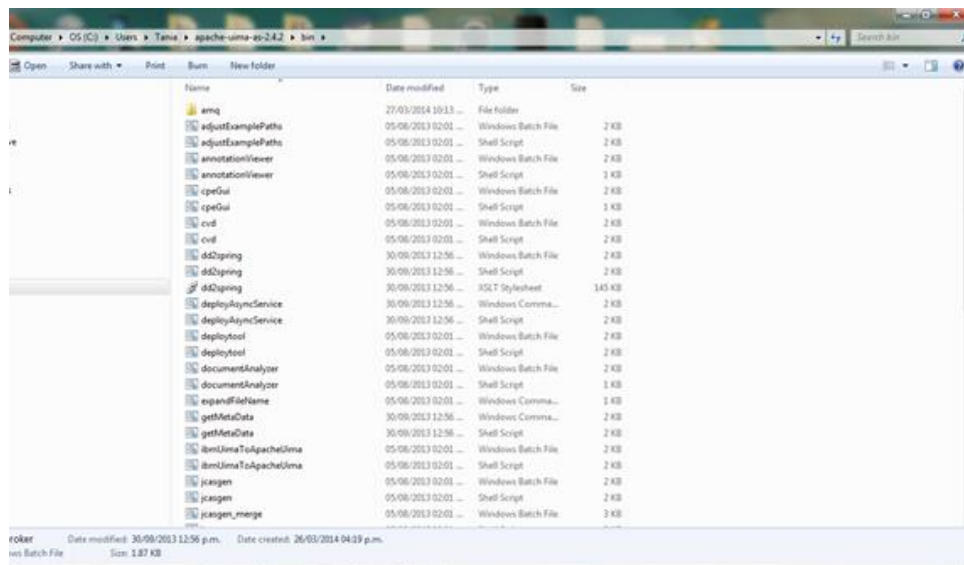
```
C:\Windows\system32\cmd.exe
s-2.4.2\apache-activemq\data -Djava.io.tmpdir=C:\Program Files\uima-as-2.4.2-bin
\apache-uima-as-2.4.2\apache-activemq\data\tmp
ACTIVEMQ_HOME: C:\Program Files\uima-as-2.4.2-bin\apache-uima-as-2.4.2\apache-ac
tivemq
ACTIVEMQ_BASE: amq
ACTIVEMQ_CONF: C:\Program Files\uima-as-2.4.2-bin\apache-uima-as-2.4.2\apache-ac
tivemq\conf
ACTIVEMQ_DATA: C:\Program Files\uima-as-2.4.2-bin\apache-uima-as-2.4.2\apache-ac
tivemq\data
Loading message broker from: xbean:file:amq/conf/activemq-nojournal.xml
INFO : Refreshing org.apache.activemq.xbean.XBeanBrokerFactory$1@5aad1d4: start
up date [Thu Mar 27 22:13:13 PDT 2014]; root of context hierarchy
INFO : Using Persistence Adapter: MemoryPersistenceAdapter
INFO : ActiveMQ 5.6.0 JMS Message Broker <localhost> is starting
INFO : For help or more information please see: http://activemq.apache.org/
INFO : JMX consoles can connect to service:jmx:rmi:///jndi/rmi://localhost:1099
/jmxrmi
INFO : Listening for connections at: tcp://Tania-PC:61616
INFO : Connector openwire Started
INFO : ActiveMQ JMS Message Broker <localhost, ID:Tania-PC-1403-1395983594388-0
:1> started
```

Listening for conections at: tcp://Tania-PC:61616

Connector openwire Started

ActiveMQ JMS Message Broker <localhost, ID:Tania-PC-1403-1395983594388-0:1> started

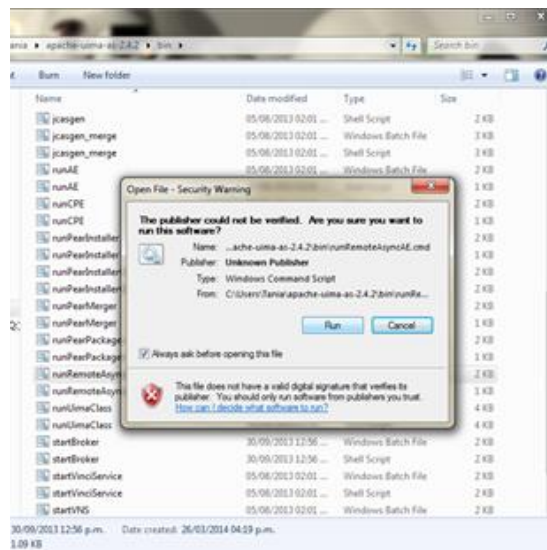
Then the amq folder is created after \$ startBroker file was executed.




```

C:\Windows\system32\cmd.exe
1 file(s) copied.
1 file(s) copied.
C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf\activemq-nojournal.xml
C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf\activemq-nojournal.xml
C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf\activemq-nojournal.xml
C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf\activemq-nojournal.xml
C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf\activemq-nojournal.xml
5 file(s) copied.
C:\Users\Tania\apache-uima-as-2.4.2\bin\call "C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\bin\activemq.bat" "%bean:file:amq/conf/activemq-nojournal.xml"
Java Runtime: Oracle Corporation 1.7.0_17 C:\Program Files\Java\jre7
Heap sizes: current=1004928k free=994639k max=1004928k
JVM args: -Dcom.sun.management.jmxremote -Xmx1G -Xms1G -Djava.util.logging.config.file=logging.properties -Dactivemq.classpath=C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf\amq\conf\C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf\activemq-nojournal.xml -Dactivemq.home=C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf\activemq-nojournal.xml -Dactivemq.data=C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\data -Djava.io.tmpdir=C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\data\temp
ACTIVEMQ_HOME: C:\Program Files\apache-uima-as-2.4.2\bin\apache-activemq
ACTIVEMQ_BASE: amq
ACTIVEMQ_CONF: C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\conf
ACTIVEMQ_DATA: C:\Program Files\apache-uima-as-2.4.2\bin\apache-uima-as-2.4.2\apache-activemq\data
Listening message broker from bean:file:amq/conf/activemq-nojournal.xml
INFO: Refreshing org.apache.activemq.xbean.XBeanBrokerFactory$1@5ad1d1: start up data [Thu Mar 27 22:13:13 PDT 2014]: root of context hierarchy
INFO: Using Persistence Adapter: MemoryPersistenceAdapter
INFO: ActiveMQ 5.6.0 JMS Message Broker (localhost) is starting
INFO: For help or more information please see: http://activemq.apache.org/
INFO: JMX console can connect to service:jmx:rmi:///jndi/rmi://localhost:1099/
INFO: Listening for connections at: tcp://Tania-PC:61616
INFO: Connector openwire Started
INFO: ActiveMQ JMS Message Broker (localhost, ID:Tania-PC-1403-1395983594388-0) started
  
```

The next step was: run the runRemoteAsynCAE.

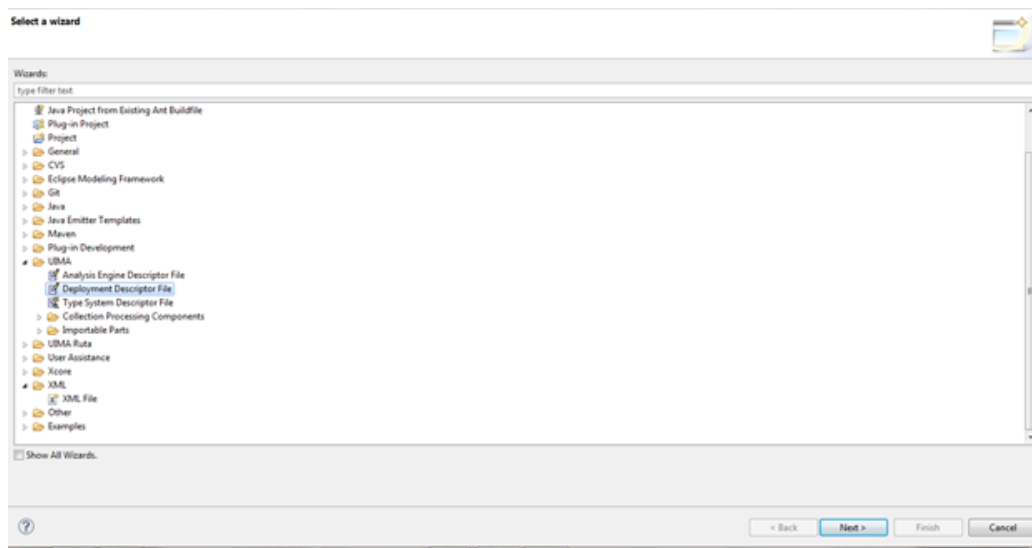


Task 2.2 Creating an UIMA-AS client

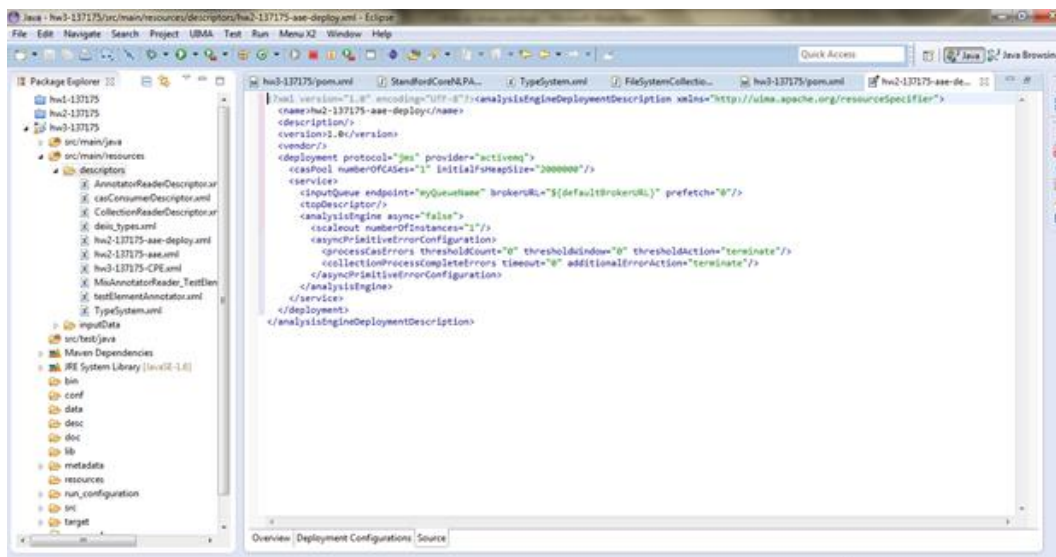
Create a UIMA-AS client descriptor (scnlp-137175-client.xml) for a remote UIMA-AS service (Stanford Core NLP), and integrate your client with your CPE pipeline.

UIMA-AS service provided for this homework is the Stanford CoreNLP. Annotator from Clear TK toolkit. This annotator reads the DocumentText form JCas and do tokenization, sentence splitting, POS tagging, lemmatization, NER, syntactic parsing, and coreference resolution.

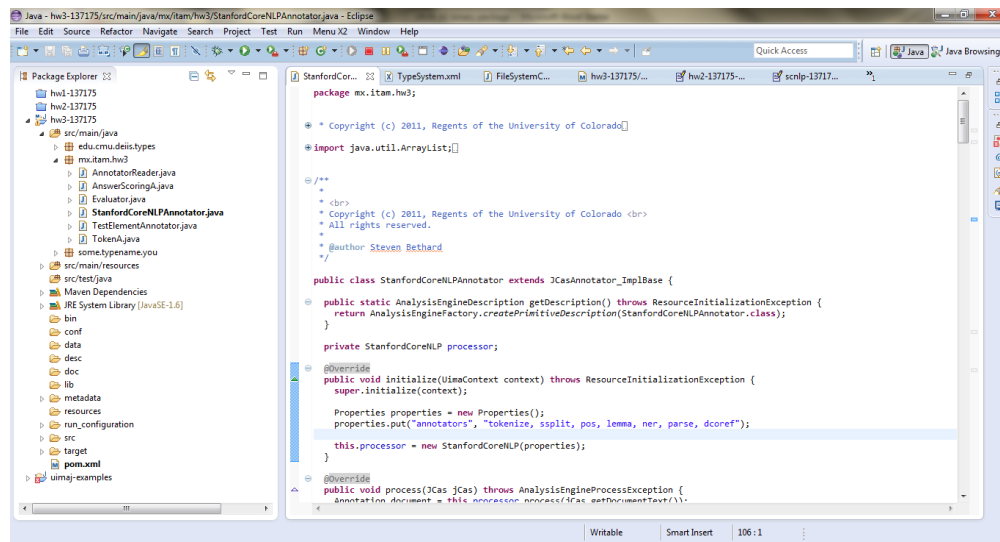
Creating the Deployment Descriptor File



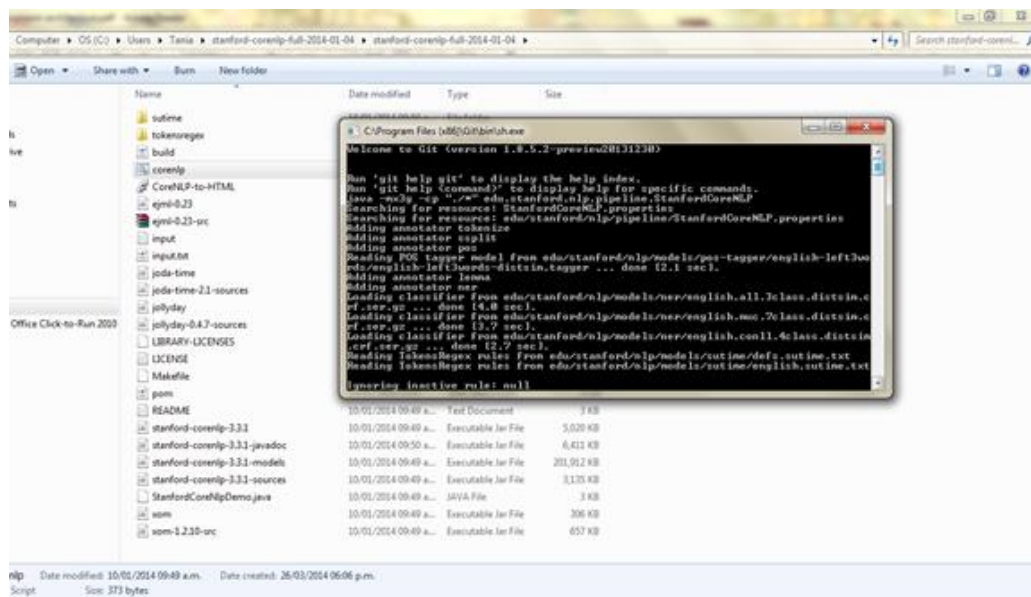
Creating the next files hw2-137175-aae-deploy.xml and scnlp-137175-client.xml



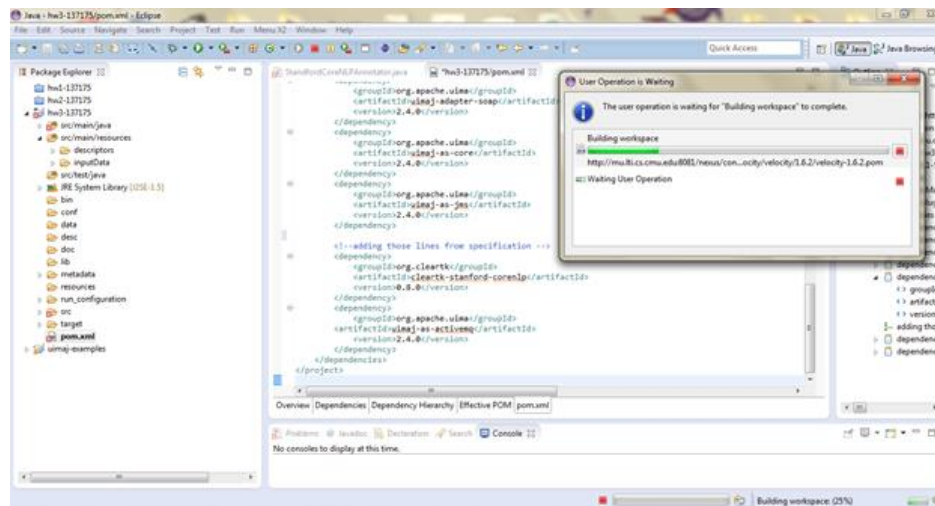
This one is the StanfordCoreNLPAnnotator, (<http://nlp.stanford.edu/software/corenlp.shtml>)



Testing and Running coreNLP:



Adding those lines into pom.xml file:



IMPORTING DEPENDENCIES INTO POM.XML

```
<dependency>
  <groupId>org.apache.uima</groupId>
  <artifactId>uimaj-as-jms</artifactId>
  <version>2.4.0</version>
</dependency>

<!--adding those lines from specification -->
  <dependency>
    <groupId>org.cleartk</groupId>
    <artifactId>cleartk-standford-corenlp</artifactId>
    <version>0.8.0</version>
  </dependency>

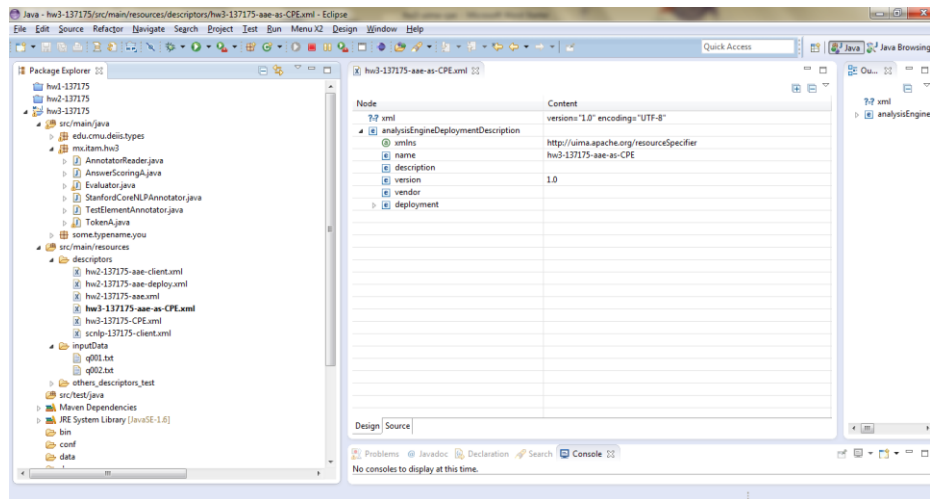
  <dependency>
    <groupId>org.apache.uima</groupId>
    <artifactId>uimaj-as-activemq</artifactId>
    <version>2.4.0</version>
  </dependency>
</dependencies>
</project>
```

Exploring the tool ClearTK and checking the TypeSystem.xml reference that is here:

<https://code.google.com/p/cleartk/source/browse/cleartk-typesystem/src/main/resources/org/cleartk/TypeSystem.xml>

Then the file \$runRemoteAsyncAE throws some messages. Also the installation about ClearTK is in progress.

Right now the project about the hw3-137175 UIMA and UIMA-AS has the next structure and organization:



Note: now I am in the integration the Name Entity annotations from *StanfordCoreNLP.java* into the answer scoring component but it is not running yet.