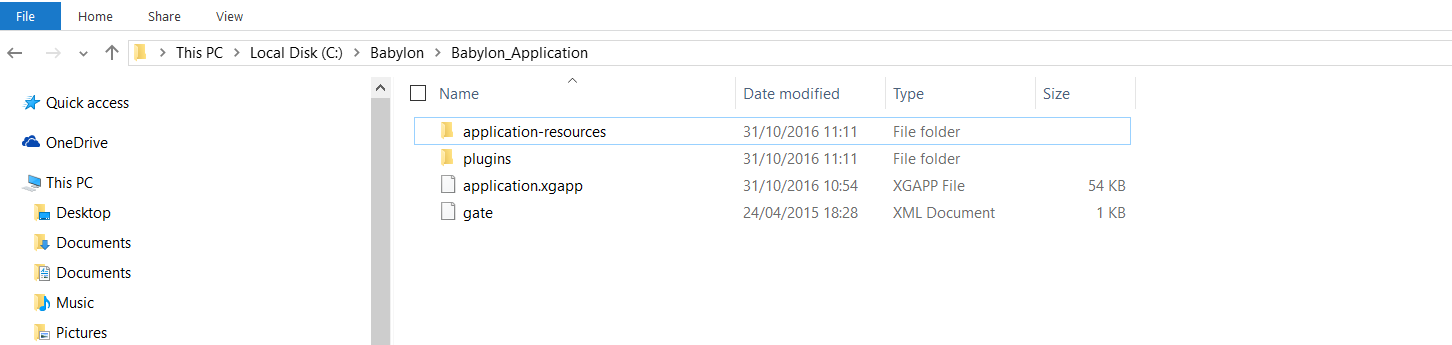
## **SYSTEM SETUP**

* Java 1.8
* GATE 8.1
* Dictionaries and knowledge base used – ICD-Version 2016 (http://apps.who.int/classifications/icd10/browse/2016/en#/R50-R69). R00-R69 consists of symptoms and diseases. The XML for this ICD.xml was converted to a set of Gazetteers available in the our GATE plugin as ICD\_Gazetteer.
* Build Application using maven. Source files can be found in Babylon\_source\_jars.jar
* Result report for each document can be found in ‘Babylon\_results’ directory

## **GATE SETUP**

The GATE processors must be located in the directory C:/Babylon/Babylon\_Application. This is because the Spring bean.xml file calls the GATE processor from ‘C:/Babylon/Babylon\_Application’. Alternatively, you can open the bean.xml file in the source code and then edit lines 17 and 21 so that the GATE processors point to a directory of your choosing.

a. Unzip Babylon\_Application into C:/Babylon/Babylon\_Application (See image below or link - http://prntscr.com/d18yc5



## **RUNNING AS A BATCH FROM CMD**

1. In ‘C:\Babylon’ create a directory called ‘batcher’

2. Copy or edit env.bat and set the following parameters

SET BABYLONDIR=C:\Babylon\batcher #directory from which the app would run

SET JDK\_PATH="C:\Program Files\Java\jdk" #path to your java jdk

#Build using Maven to generate class files. Copy class files to #%BABYLONDIR%\classes

SET CLASSPATH=%BABYLONDIR%\classes;%BABYLONDIR% #classpath to the necessary classes

3.Save env.bat in ‘C:\Babylon\batchProcesses

4. Copy or edit run.bat and set the following parameters

call env.bat

%JDK\_PATH%\bin\java -Xmx2448M -cp %CLASSPATH% com.util.DocProcessor [http://www.nhs.uk/Conditions/Obsessive-compulsive-disorder/Pages/Symptoms.aspx obsessive-compulsive-disorder.txt](http://www.nhs.uk/Conditions/Obsessive-compulsive-disorder/Pages/Symptoms.aspx%20obsessive-compulsive-disorder.txt)

##Note that to run the second line, You must enter the url followed by a space and then the name of a file where the report will be stored i.e [url] [file-name]

5. Run run.bat

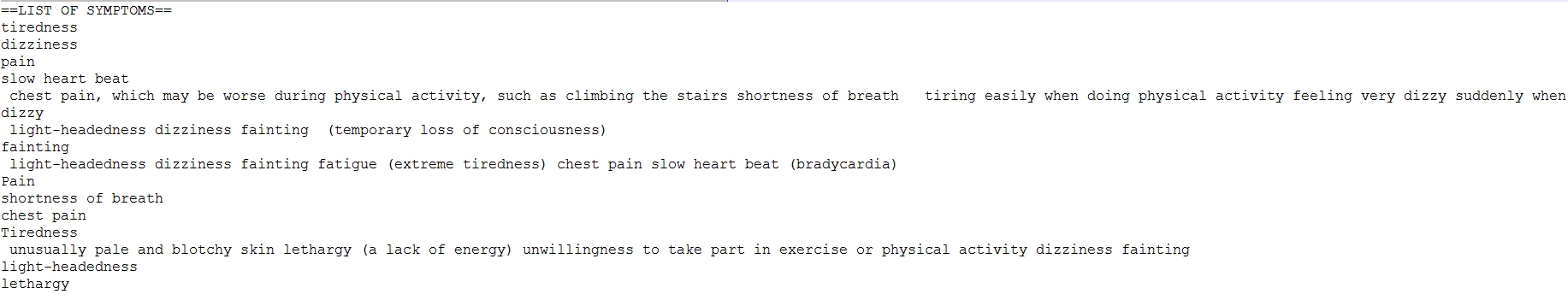
6. The result of the analysis would be stored in the file.

7. Please see the results folder for the result of the extractions for the 10 urls sent.

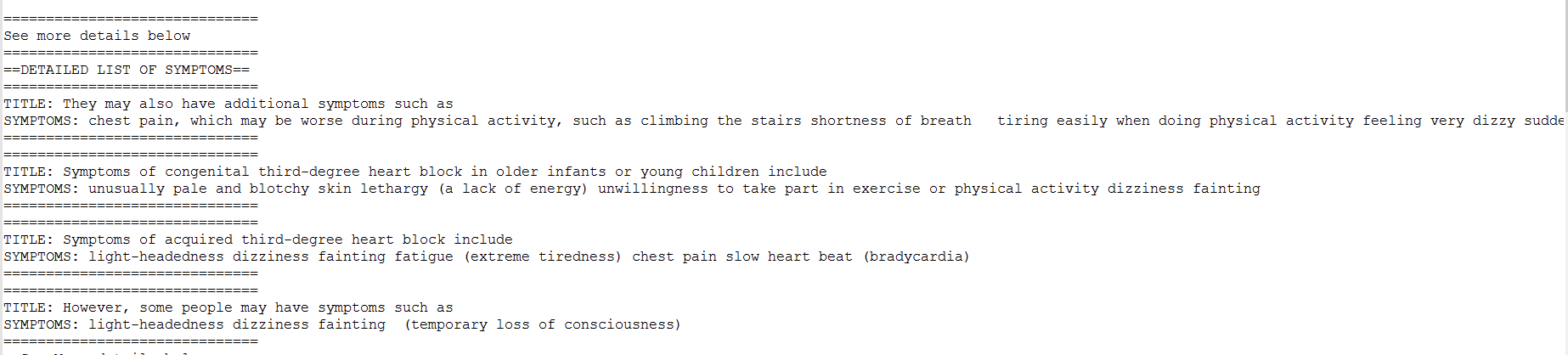
## **RESULTS OUTLINE**

There are 10 results files located in ‘Babylon\_Results’. The result.txt file is organized as follows.

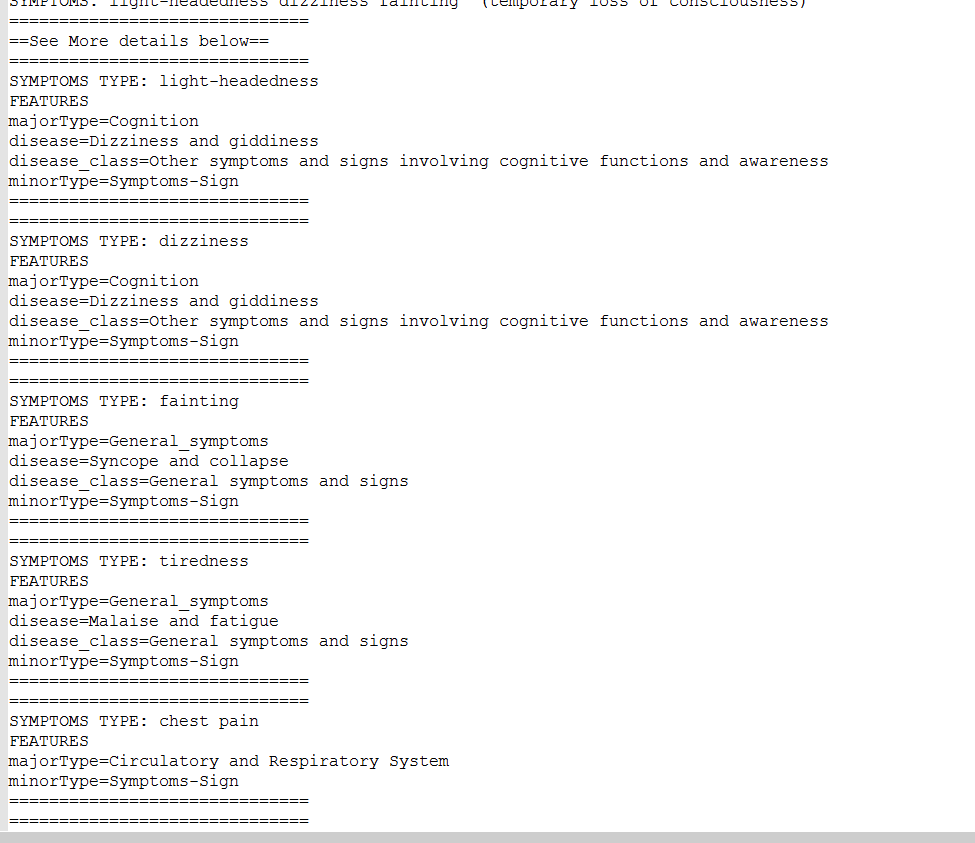
The first part shows a list of all symptoms identified. See image below or this url - http://prntscr.com/d1990z



The second part shows symptoms derived from lexical clues in the document (See image or this url) - http://prntscr.com/d199qb



The third part show symptoms identified using our knowledgebase and some of their associated features – eg the disease associated with the symptom, the part of the body, the mode of presentation and the bodily system affected. (See image or link - http://prntscr.com/d19av3 )



## BRIEF DESCRIPTION OF DESIGN

The design approach involved:

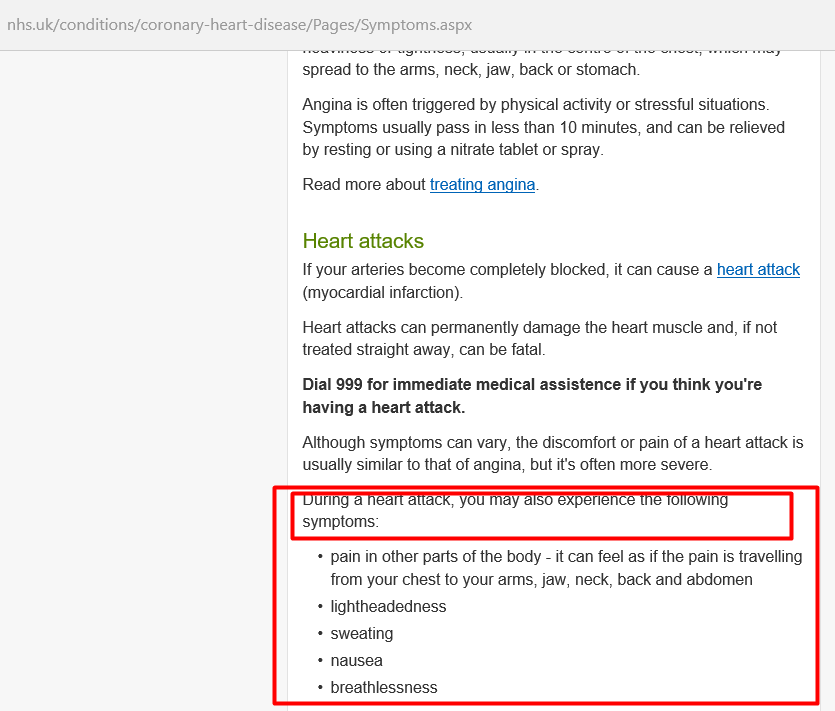
1.Resource Generation

a. Symptom entities were identified using a knowledge base. The preferred knowledge base was UMLS but because I was unable to get a license key as quickly as I’d hoped. I opted for ICD-10 Version 2016. The ICD xml was converted to a GATE Gazetteer. See the Gate Application folder’s ‘Babylon\_Application/application-resources/ICD\_GAZETTEER’. Each symptom entity was marked up with vital features such as the type of disease, body part etc

b. ICD-10 does is quite limited in its coverage of Symptoms and diseases and so when applied to documents , it produces low recall especially for symptoms expressed as sentences. To improve recall and identify these sentence based symptoms, the documents were examined to identify marked lexical and structural cues.

It was discovered that most symptom ‘sentences’ were always enumerated as bullet points and preceded by a sentence ending with a colon and containing the word ‘symptom or symptoms’. The symptoms were also enumerated as bullet points and marked up with ‘<li>’ tag. The intuition behind this is that for formal documents and medical websites it is quite common to enumerate vital and relevant facts and snippets as bullet points so that they are easily identifiable by readers who are not interested in reading through masses of irrelevant content.

A set of heuristic rules were written (These can be found in the GATE Application folder’s directory ‘HEURISTIC\_JAPE’) to identify the symptoms and their titles/enumerating sentence. An example of this can be seen in the image below



2. Identification of Symptom Entities

Before applying the algorithms and rules, the document would normally be pre-processed but because of time constraints, this step was omitted. The effect of its omission did not have a major consequence on the extraction.

The GATE application was applied using a Java class in the java project ‘Babaylon1’. The processing class is called ‘com.util.DocProcessor’. Users can enter a document url and a text file where a basic result report would be articulated.