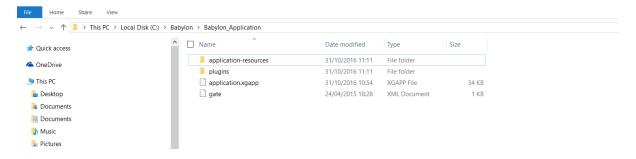## SYSTEM SETUP

- 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

##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

```
==LIST OF SYMPTOMS==
tiredness
dizziness
pain
slow heart beat
  chest pain, which may be worse during physical activity, such as climbing the stairs shortness of breath   tiring easily when doing physical activity feeling very dizzy suddenly when
dizzy
  light-headedness dizziness fainting  (temporary loss of consciousness)
fainting
  light-headedness dizziness fainting fatigue (extreme tiredness) chest pain slow heart beat (bradycardia)
Pain
shortness of breath
chest pain
Tiredness
  unusually pale and blotchy skin lethargy (a lack of energy) unwillingness to take part in exercise or physical activity dizziness fainting
light-headedness
lethargy
```

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

```
==============================
See more details below
==============================
==DETAILED LIST OF SYMPTOMS==
==============================
TITLE: They may also have additional symptoms such as
SYMPTOMS: chest pain, which may be worse during physical activity, such as climbing the stairs shortness of breath   tiring easily when doing physical activity feeling very dizzy sudde
==============================
TITLE: Symptoms of congenital third-degree heart block in older infants or young children include
SYMPTOMS: unusually pale and blotchy skin lethargy (a lack of energy) unwillingness to take part in exercise or physical activity dizziness fainting
==============================
TITLE: Symptoms of acquired third-degree heart block include
SYMPTOMS: light-headedness dizziness fainting fatigue (extreme tiredness) chest pain slow heart beat (bradycardia)
==============================
TITLE: However, some people may have symptoms such as
SYMPTOMS: light-headedness dizziness fainting  (temporary loss of consciousness)
==============================
```

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)

```
SIMPTOMS: light-headedness dizziness fainting   (temporary loss of consciousness)
=============================
==See More details below==
=============================
SYMPTOMS TYPE: light-headedness
FEATURES
majorType=Cognition
disease=Dizziness and giddiness
disease_class=Other symptoms and signs involving cognitive functions and awareness
minorType=Symptoms-Sign
=============================
=============================
SYMPTOMS TYPE: dizziness
FEATURES
majorType=Cognition
disease=Dizziness and giddiness
disease_class=Other symptoms and signs involving cognitive functions and awareness
minorType=Symptoms-Sign
=============================
=============================
SYMPTOMS TYPE: fainting
FEATURES
majorType=General_symptoms
disease=Syncope and collapse
disease_class=General symptoms and signs
minorType=Symptoms-Sign
=============================
=============================
SYMPTOMS TYPE: tiredness
FEATURES
majorType=General_symptoms
disease=Malaise and fatigue
disease_class=General symptoms and signs
minorType=Symptoms-Sign
=============================
=============================
SYMPTOMS TYPE: chest pain
FEATURES
majorType=Circulatory and Respiratory System
minorType=Symptoms-Sign
=============================
=============================
```

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

spread to the arms, neck, jaw, back or stomach.

Angina is often triggered by physical activity or stressful situations. Symptoms usually pass in less than 10 minutes, and can be relieved by resting or using a nitrate tablet or spray.

Read more about treating angina.

Heart attacks

If your arteries become completely blocked, it can cause a heart attack (myocardial infarction).

Heart attacks can permanently damage the heart muscle and, if not treated straight away, can be fatal.

Dial 999 for immediate medical assistence if you think you're having a heart attack.

Although symptoms can vary, the discomfort or pain of a heart attack is usually similar to that of angina, but it's often more severe.

During a heart attack, you may also experience the following symptoms:

- pain in other parts of the body - it can feel as if the pain is travelling from your chest to your arms, jaw, neck, back and abdomen
- lightheadedness
- sweating
- nausea
- breathlessness

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.