Question/Answering System for Diseases

Team 12

- Gupta, Arunit – 11
- Patel, Marmikkumar Navinchandra – 31
- McDuff, Luke Joseph – 21
- Ejjirothu, Manoj Prabhakar – 7

Table of Contents

1. Introduction

Question and Answering system for Diseases

The QA system is designed for users around the globe. Users can access the web page and ask their queries regarding health and diseases. The system is designed for a user to have the ability to ask a question on the webpage in a textbox and retrieve the answer for the same in an organized manner.

The system uses a corpus of unstructured information across the web and live data from twitter tweets for the knowledge base and processes the data to return results.

Input is provided by the user on the webpage and output is displayed on the same page by processing data from the knowledge base. It uses natural language processing to lemmatize the data, divide into Parts of Speech, parsing, and creating named entity relationships. The system uses machine learning principles and ontologies to process data and come up with a correct answer or closest match.

The system is useful for people to learn about a particular disease in a precise and organized format. If we have any problem or query we usually search on the internet. We come up with multiple links where we get scattered information and need to search across links to get the answer; this system has the ability to analyze and interpret the question, and get the answer from our knowledge base after applying processing techniques. The system then returns the answer in an organized and specific format with graphical and image representation which is useful for naïve user.

2. Project Goals and Objectives

2.1 Motivation

People always have medical queries and usually they use search engines to get answers for their question, but all information is scattered on the web, we have a huge corpus of information but not in a structured manner. This system is designed for the users where they can ask their queries and get response in a precise and structured manner.

2.2 Objective

To design a system which can take a question from a user and give an answer as output by applying natural language and machine learning principles and display the information in a structured format.

2.3 Expected Outcomes

A webpage accessible by users across the world having the ability to ask a question, and get a response on the webpage in a structured format.

3. Project Domain & Datasets

3.1 Project Domain

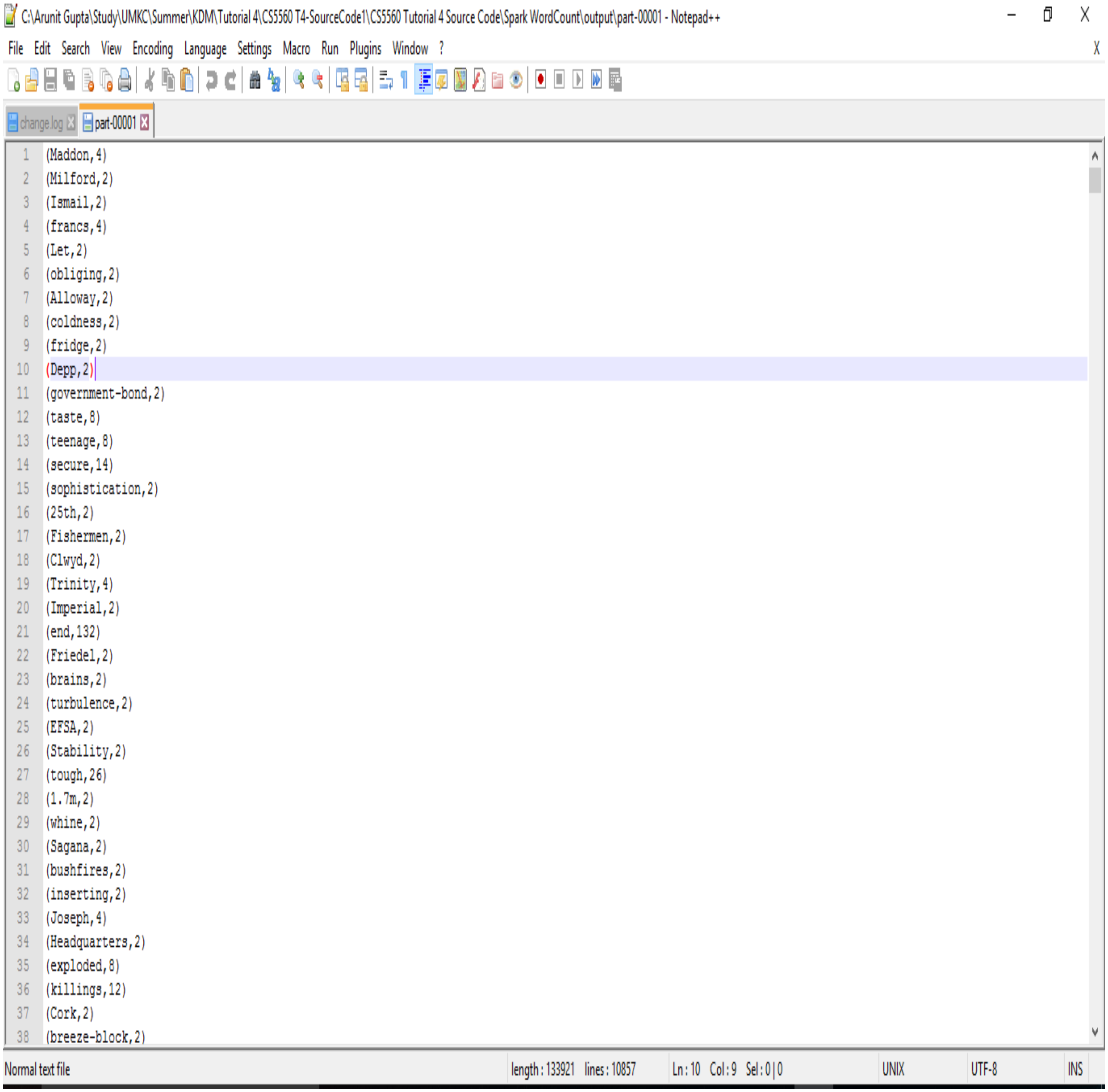
Healthcare - Consisting of Diseases, health issues, drug information, trending viruses, allergies etc.

3.2 Dataset Links

- https://dev.twitter.com/overview/api - Collected 200K tweets on health and Diseases.
- http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2
- http://www.statmt.org/lm-benchmark/1-billion-word-language-modeling-benchmark-r13output.tar.gz
- http://ebiquity.umbc.edu/redirect/to/resource/id/351/UMBC-webbase-corpus
- http://blog.wolframalpha.com/category/health-med/
- http://archive.ics.uci.edu/ml/datasets/Hepatitis

4. Features Implementation

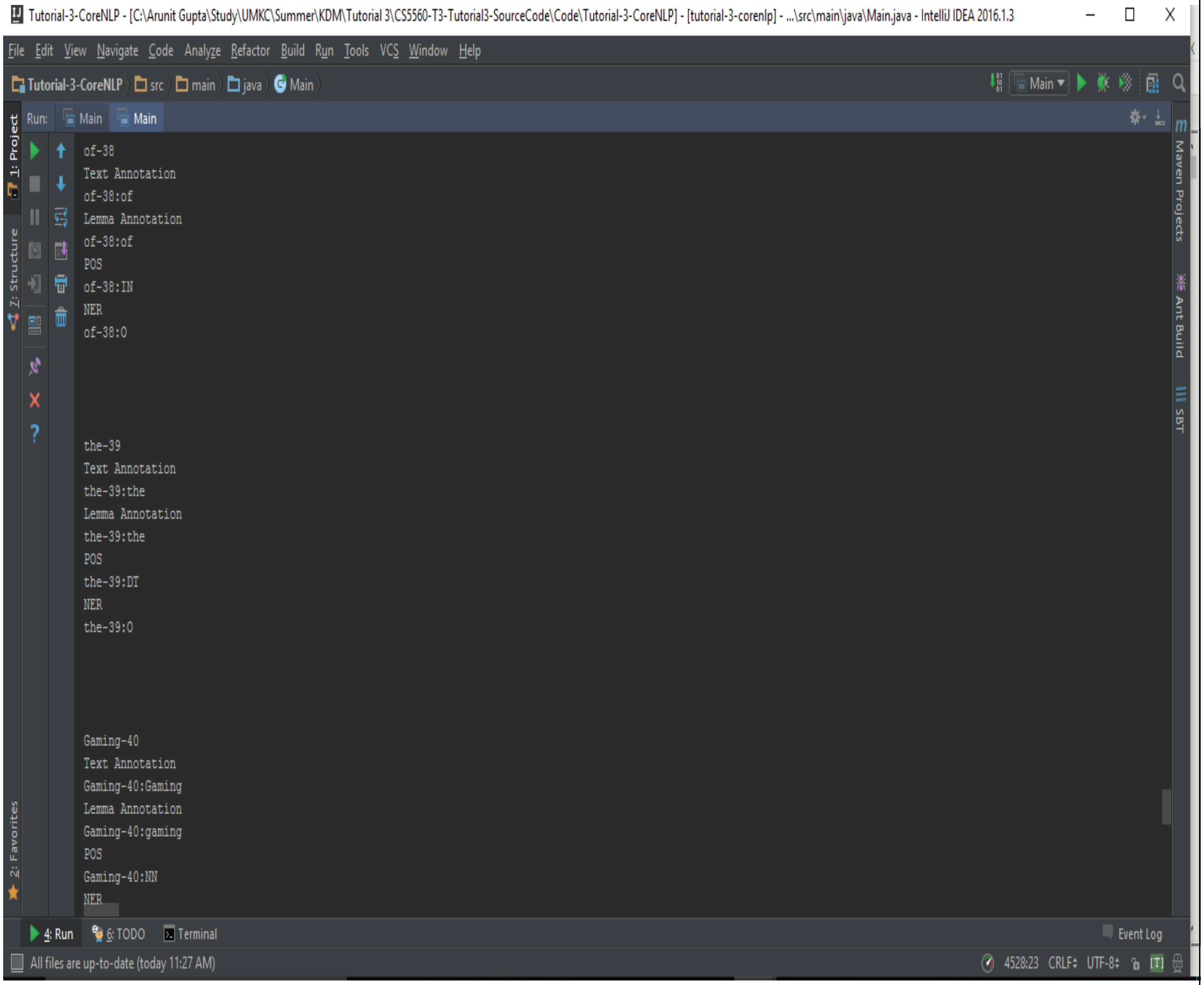
4.1 Word Count


C:\Arunit Gupta\Study\UMKC\Summer\KDM\Tutorial 4\CS5560 T4-SourceCode1\CS5560 Tutorial 4 Source Code\Spark WordCount\output\part-00001 - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
change.log
part-00001
Normal text file
length : 133921 lines : 10857
Ln : 10 Col : 9 Sel : 0 | 0
UNIX
UTF-8
INS

```
(Maddon,4)
(Milford,2)
(Ismail,2)
(francs,4)
(Let,2)
(obliging,2)
(Alloway,2)
(coldness,2)
(fridge,2)
(Depp,2)
(government-bond,2)
(taste,8)
(teenage,8)
(secure,14)
(sophistication,2)
(25th,2)
(Fishermen,2)
(Clwyd,2)
(Trinity,4)
(Imperial,2)
(end,132)
(Friedel,2)
(brains,2)
(turbulence,2)
(EFSA,2)
(Stability,2)
(tough,26)
(1.7m,2)
(whine,2)
(Sagana,2)
(bushfires,2)
(inserting,2)
(Joseph,4)
(Headquarters,2)
(exploded,8)
(killings,12)
(Cork,2)
(breeze-block,2)
```

4.2 NLP Processing

Lemmatization

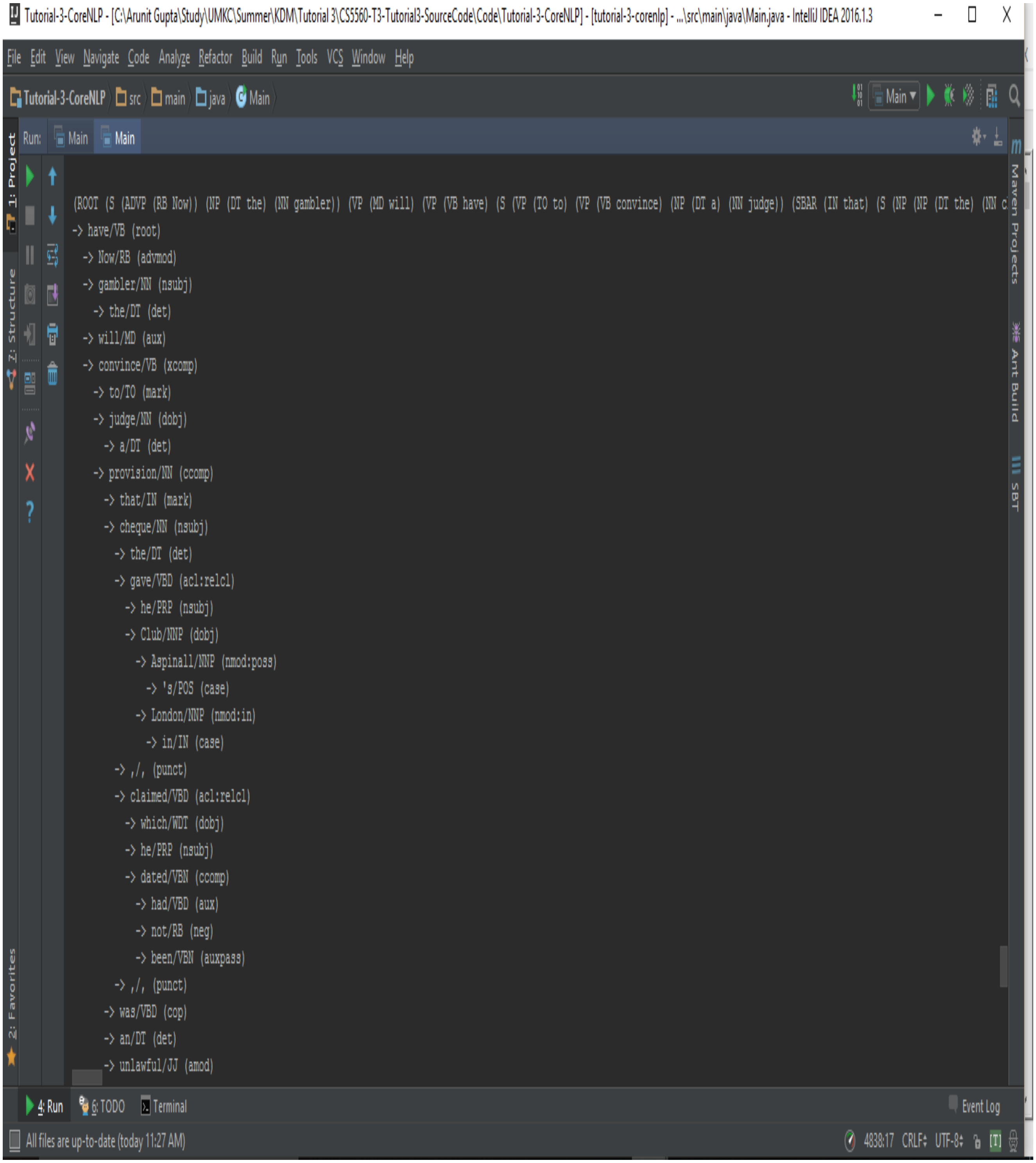


Tutorial-3-CoreNLP - [C:\Arunit Gupta\Study\UMKC\Summer\KDM\Tutorial 3\CS5560-T3-Tutorial3-SourceCode\Code\Tutorial-3-CoreNLP] - [tutorial-3-corenlp] - ...\src\main\java\Main.java - IntelliJ IDEA 2016.1.3

```
of-38
Text Annotation
of-38:of
Lemma Annotation
of-38:of
POS
of-38:IN
NER
of-38:O




the-39
Text Annotation
the-39:the
Lemma Annotation
the-39:the
POS
the-39:DT
NER
the-39:O




Gaming-40
Text Annotation
Gaming-40:Gaming
Lemma Annotation
Gaming-40:gaming
POS
Gaming-40:NN
NER
```

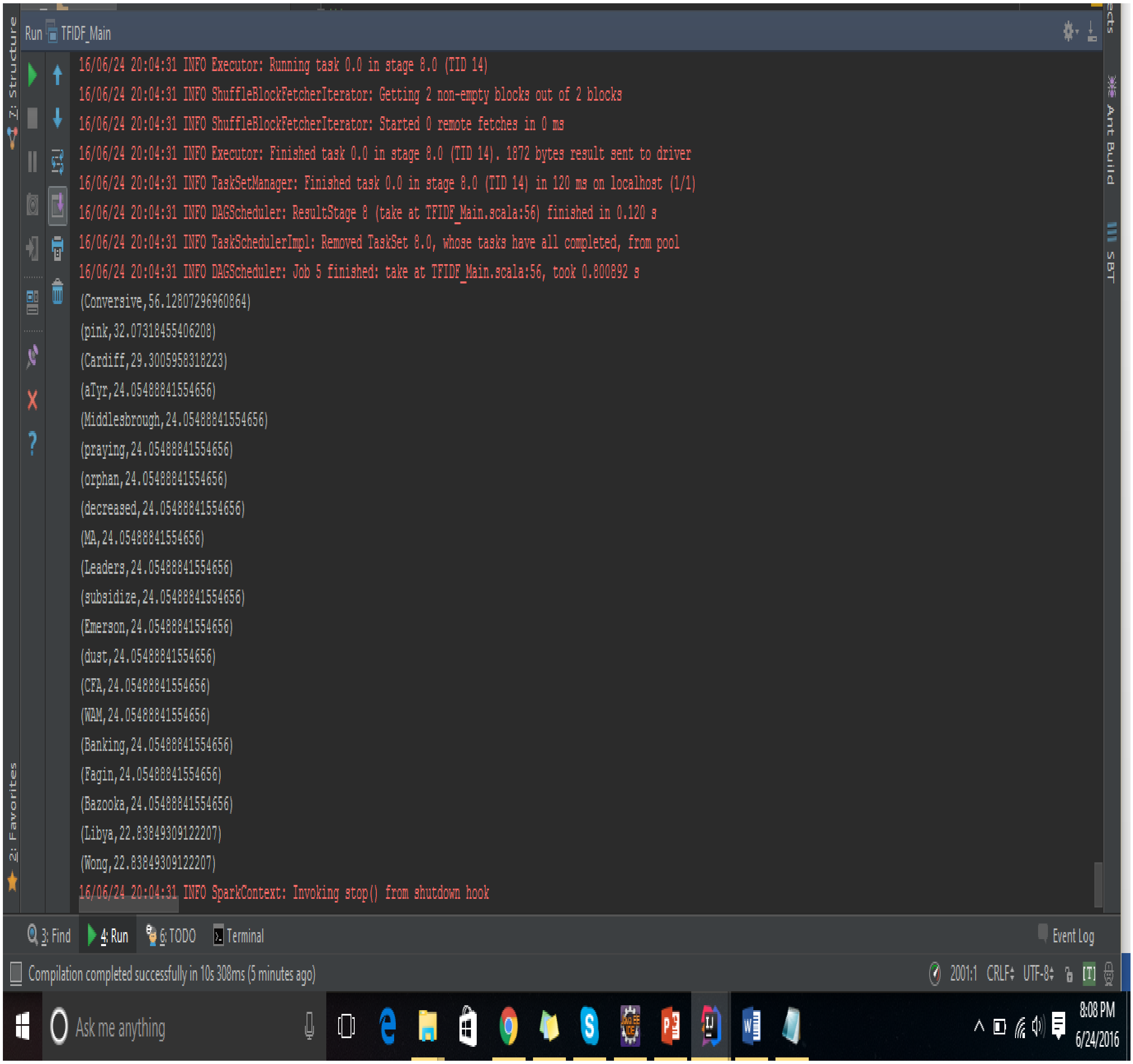
All files are up-to-date (today 11:27 AM)

POS


Tutorial-3-CoreNLP - [C:\Arunit Gupta\Study\UMKC\Summer\KDM\Tutorial 3\CS5560-T3-Tutorial3-SourceCode\Code\Tutorial-3-CoreNLP] - [tutorial-3-corenlp] - ...\src\main\java\Main.java - IntelliJ IDEA 2016.1.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Tutorial-3-CoreNLP src main java Main
Main
Run: Main Main
1: Project
Z: Structure
2: Favorites
Maven Projects
Ant Build
SBT
4: Run 6: TODO Terminal
Event Log
All files are up-to-date (today 11:27 AM)
4838:17 CRLF UTF-8

```
(ROOT (S (ADVP (RB Now)) (NP (DT the) (NN gambler)) (VP (MD will) (VP (VB have) (S (VP (TO to) (VP (VB convince) (NP (DT a) (NN judge)) (SBAR (IN that) (S (NP (NP (DT the) (NN c
-> have/VB (root)
  -> Now/RB (advmod)
  -> gambler/NN (nsubj)
    -> the/DT (det)
  -> will/MD (aux)
  -> convince/VB (xcomp)
    -> to/TO (mark)
    -> judge/NN (dobj)
      -> a/DT (det)
    -> provision/NN (ccomp)
      -> that/IN (mark)
      -> cheque/NN (nsubj)
        -> the/DT (det)
        -> gave/VBD (acl:relcl)
          -> he/PRP (nsubj)
          -> Club/NNP (dobj)
            -> Aspinall/NNP (nmod:poss)
              -> 's/POS (case)
            -> London/NNP (nmod:in)
              -> in/IN (case)
        -> ,/, (punct)
        -> claimed/VBD (acl:relcl)
          -> which/WDT (dobj)
          -> he/PRP (nsubj)
          -> dated/VBN (ccomp)
            -> had/VBD (aux)
            -> not/RB (neg)
            -> been/VBN (auxpass)
        -> ,/, (punct)
      -> was/VBD (cop)
      -> an/DT (det)
      -> unlawful/JJ (amod)
```

4.3 Information Retrieval/Extraction Technologies

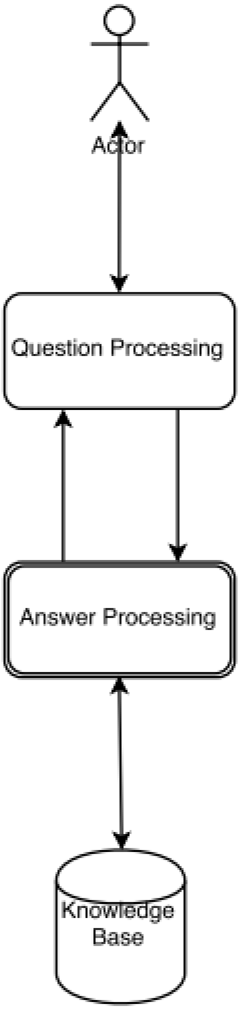


```
16/06/24 20:04:31 INFO Executor: Running task 0.0 in stage 8.0 (TID 14)
16/06/24 20:04:31 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
16/06/24 20:04:31 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/06/24 20:04:31 INFO Executor: Finished task 0.0 in stage 8.0 (TID 14). 1872 bytes result sent to driver
16/06/24 20:04:31 INFO TaskSetManager: Finished task 0.0 in stage 8.0 (TID 14) in 120 ms on localhost (1/1)
16/06/24 20:04:31 INFO DAGScheduler: ResultStage 8 (take at TFIDF_Main.scala:56) finished in 0.120 s
16/06/24 20:04:31 INFO TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed, from pool
16/06/24 20:04:31 INFO DAGScheduler: Job 5 finished: take at TFIDF_Main.scala:56, took 0.800892 s
(Conversive,56.12807296960864)
(pink,32.07318455406208)
(Cardiff,29.3005958318223)
(aTyr,24.05488841554656)
(Middlesbrough,24.05488841554656)
(praying,24.05488841554656)
(orphan,24.05488841554656)
(decreased,24.05488841554656)
(MA,24.05488841554656)
(Leaders,24.05488841554656)
(subsidize,24.05488841554656)
(Emerson,24.05488841554656)
(dust,24.05488841554656)
(CFA,24.05488841554656)
(WAM,24.05488841554656)
(Banking,24.05488841554656)
(Fagin,24.05488841554656)
(Bazooka,24.05488841554656)
(Libya,22.83849309122207)
(Wong,22.83849309122207)
16/06/24 20:04:31 INFO SparkContext: Invoking stop() from shutdown hook
```

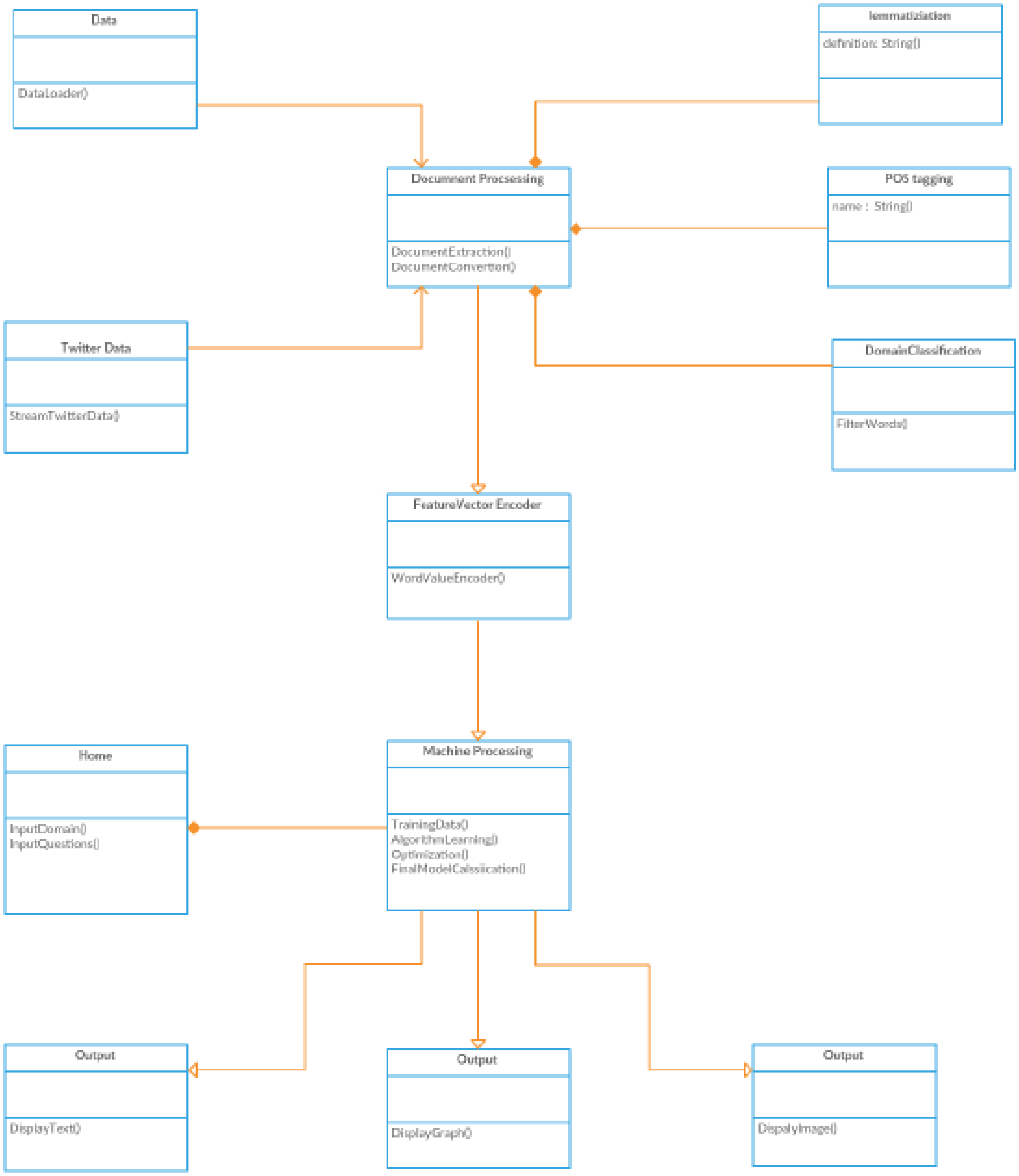
Compilation completed successfully in 10s 308ms (5 minutes ago)

5. Implementation Specification

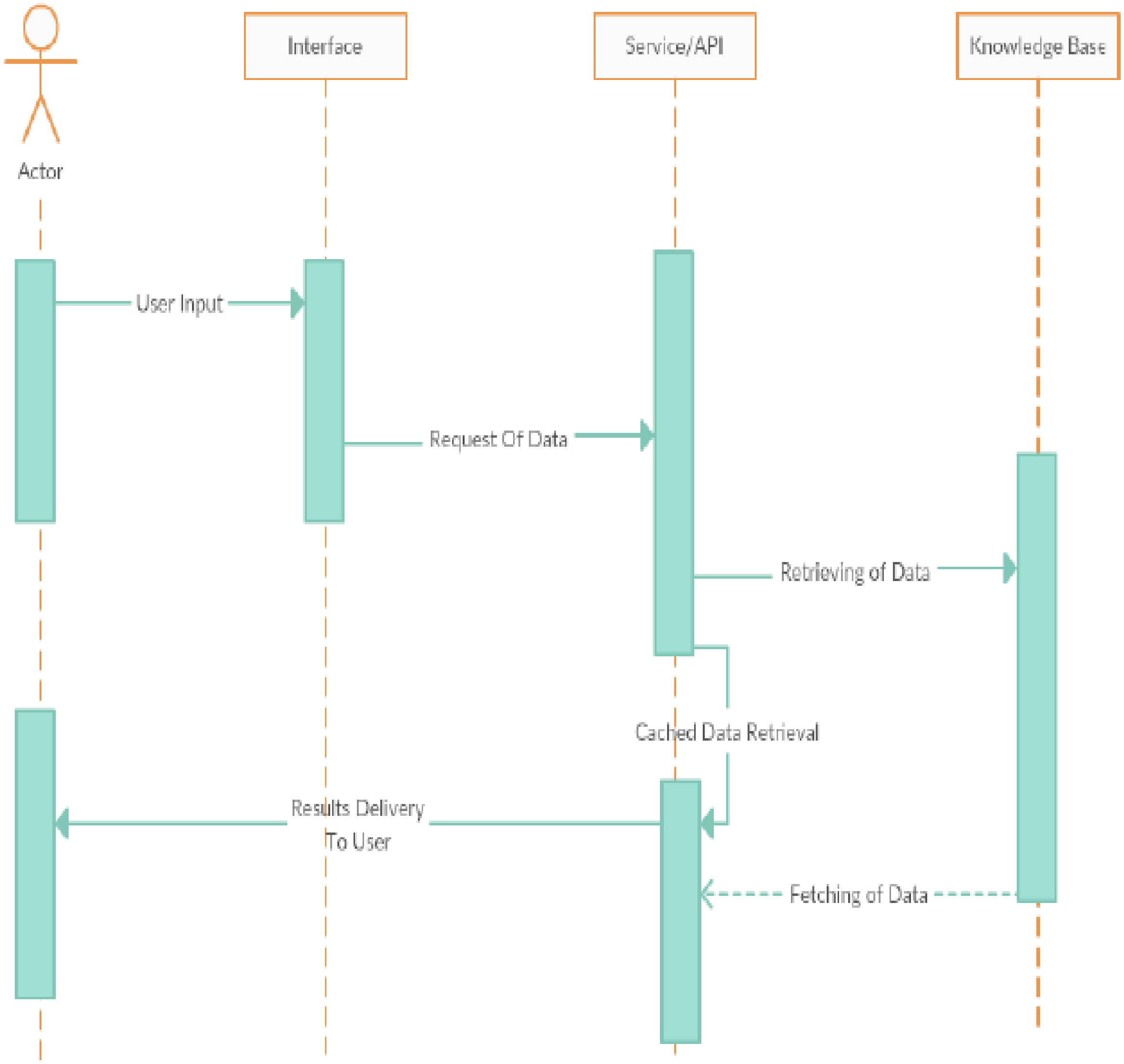
5.1 Software Architecture Diagram



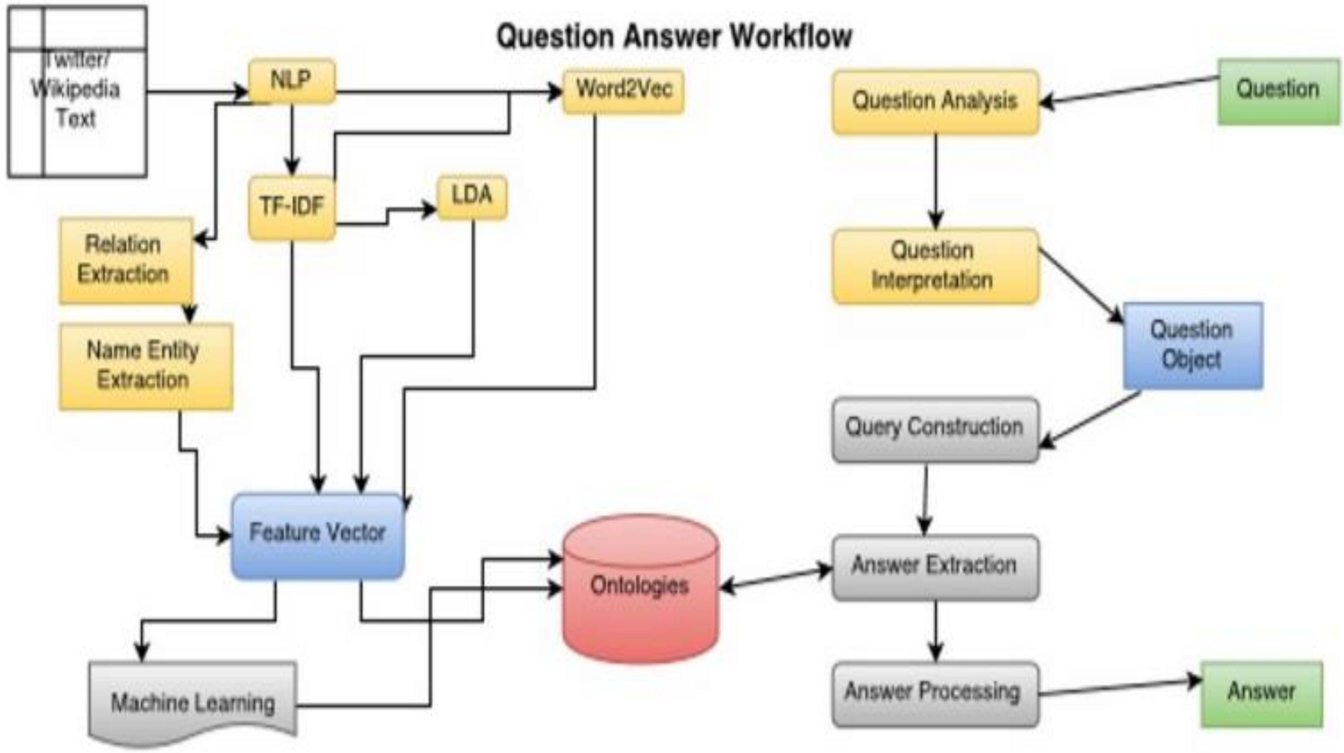
5.2 Class Diagram



5.3 Sequence Diagram



5.4 Workflow



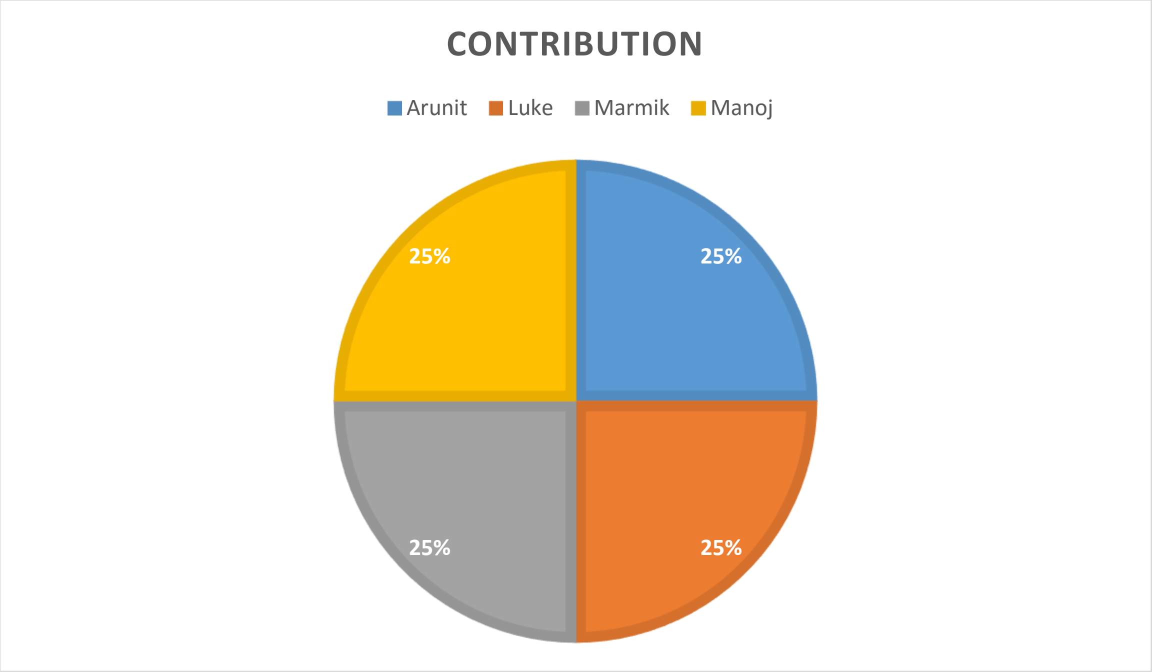
5.5 Existing services/features used

AWS, Alchemy API

6. Project Management

6.1 Contribution of each member

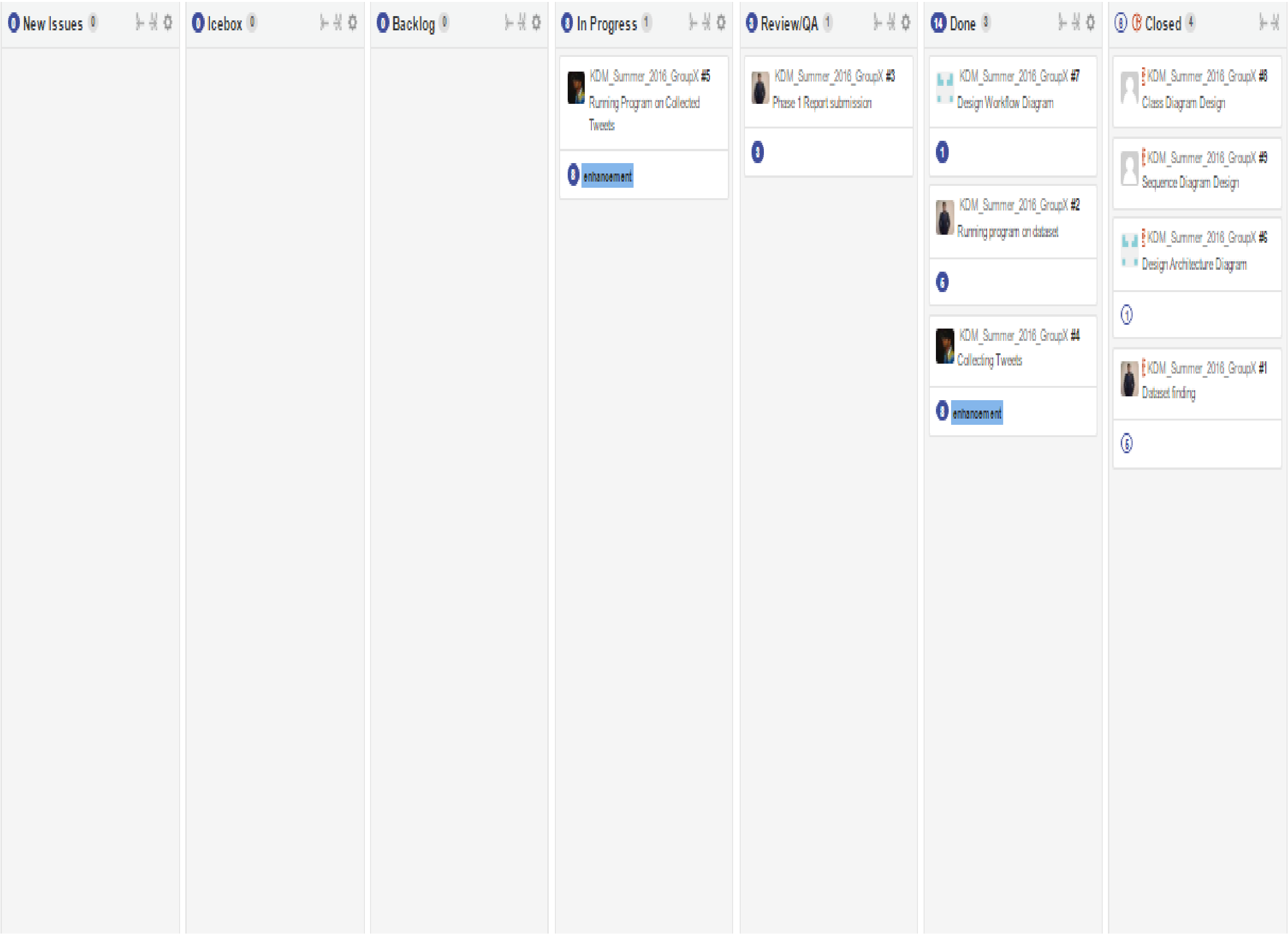
- **Gupta, Arunit** – Dataset Findings, Running dataset on Program, Report compilation
- **Patel, Marmikkumar Navinchandra** – collection of tweets, running program on tweets collected
- **McDuff, Luke Joseph** – Architecture Diagram design and Workflow design
- **Ejjirothu, Manoj Prabhakar** – Class Diagram and Sequence Diagram



6.2 Zen hub and GitHub URL/Statistics

GitHub link: https://github.com/ljm7b2/KDM_Summer_2016_GroupX

Zen hub



0 New Issues 0

0 Icebox 0

0 Backlog 0

8 In Progress 1

KDM_Summer_2016_GroupX #5
Running Program on Collected Tweets
8 enhancement

8 Review/QA 1

KDM_Summer_2016_GroupX #3
Phase 1 Report submission
8

14 Done 3

KDM_Summer_2016_GroupX #7
Design Workflow Diagram
1

KDM_Summer_2016_GroupX #2
Running program on dataset
5

KDM_Summer_2016_GroupX #4
Collecting Tweets
8 enhancement

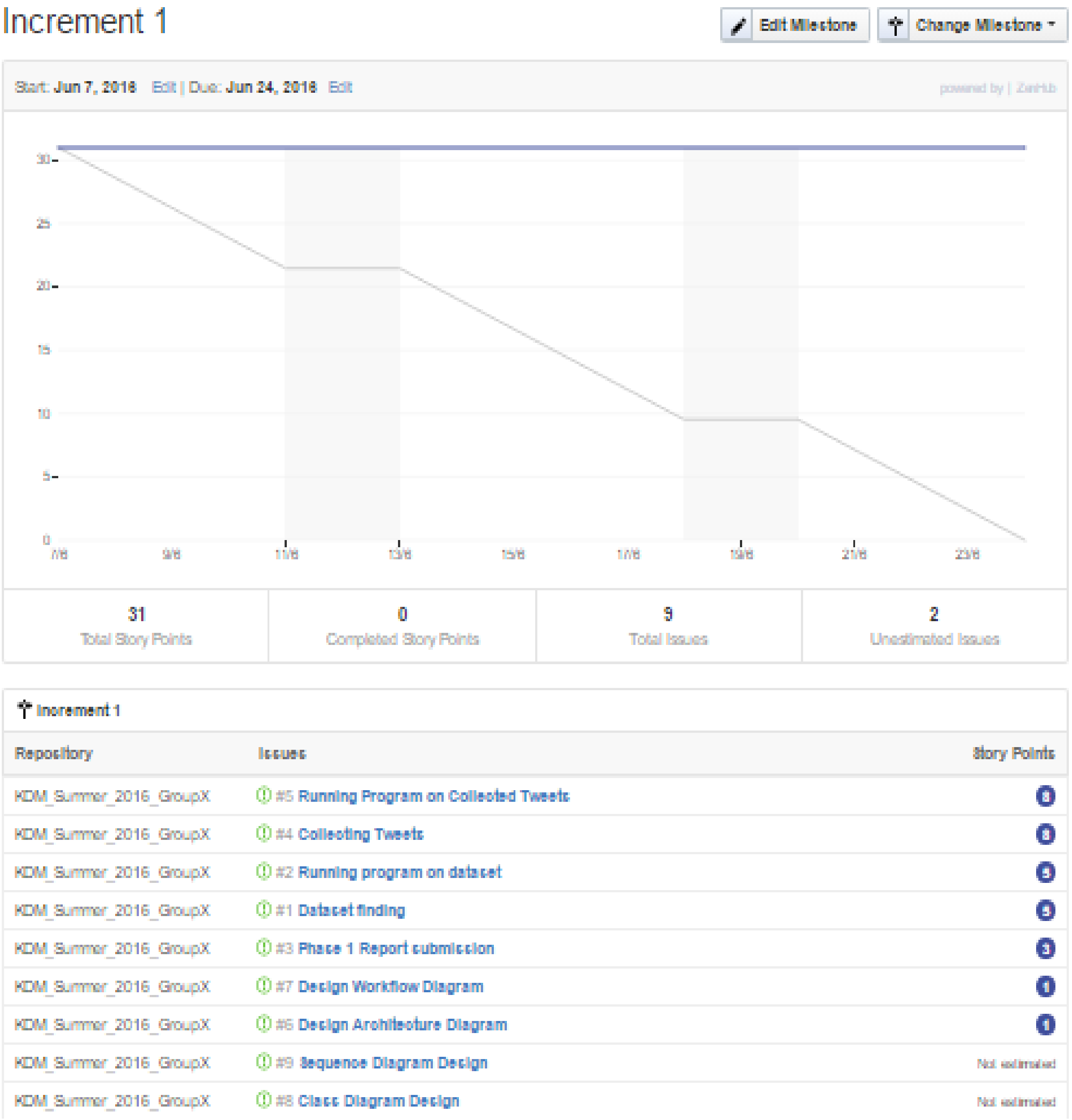
8 Closed 4

KDM_Summer_2016_GroupX #8
Class Diagram Design

KDM_Summer_2016_GroupX #9
Sequence Diagram Design

KDM_Summer_2016_GroupX #6
Design Architecture Diagram
1

KDM_Summer_2016_GroupX #1
Dataset finding
5



Increment 1

Start: Jun 7, 2016 Edit | Due: Jun 24, 2016 Edit

31	0	9	2
Total Story Points	Completed Story Points	Total Issues	Unestimated Issues

Increment 1

Repository	Issues	Story Points
KDM_Summer_2016_GroupX	#5 Running Program on Collected Tweets	8
KDM_Summer_2016_GroupX	#4 Collecting Tweets	8
KDM_Summer_2016_GroupX	#2 Running program on dataset	5
KDM_Summer_2016_GroupX	#1 Dataset finding	5
KDM_Summer_2016_GroupX	#3 Phase 1 Report submission	3
KDM_Summer_2016_GroupX	#7 Design Workflow Diagram	1
KDM_Summer_2016_GroupX	#6 Design Architecture Diagram	1
KDM_Summer_2016_GroupX	#9 Sequence Diagram Design	Not estimated
KDM_Summer_2016_GroupX	#8 Class Diagram Design	Not estimated

6.3 Concerns/Issues

Relating all the technologies and merging them to get final output

6.4 Future Work

Implementing the datasets and live data on machine learning and ontologies programs, implementing voice to text search

7. References

- https://github.com/DiseaseOntology/HumanDiseaseOntology
- In Class Tutorials
- https://dev.twitter.com/overview/api
- http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2
- http://www.statmt.org/lm-benchmark/1-billion-word-language-modeling-benchmark-r13output.tar.gz
- http://ebiquity.umbc.edu/redirect/to/resource/id/351/UMBC-webbase-corpus
- http://blog.wolframalpha.com/category/health-med/
- http://archive.ics.uci.edu/ml/datasets/Hepatitis