CS5560 Knowledge Discovery and Management

Project 1 – Report

Dynamic Question Answering System using Natural Language Processing

Team 2

Mudunuri, Sai Sarat Chandra Varma (14)

Pabolu, Sandeep (19)

Pallepati, Rakesh Reddy (20)

Surparaju, Lava Kumar (27)

Motivation:

On the onset, man is after the quest for the knowledge. This thrust led us to the invention of lot many things including the most precious computers. This thrust starts with some questions like" Why?"," What"," Who" and so on. This finally led the working on question answering system.

Objective:

The main objective of the project is to set up a code which generates approximate answers for user's quest based on the dataset. This sharing of ideas, experience and information should be available in the right place at the right time thereby reducing the user's wait time.

Significance:

The major significances can be listed out as under:

- It increases the interactivity of the user with the database.
- Searched answers can be viewed frequently.
- Sentiment analysis can be performed for the better understanding.
- A proper usage of the linguistic resources.

Domain and Datasets:

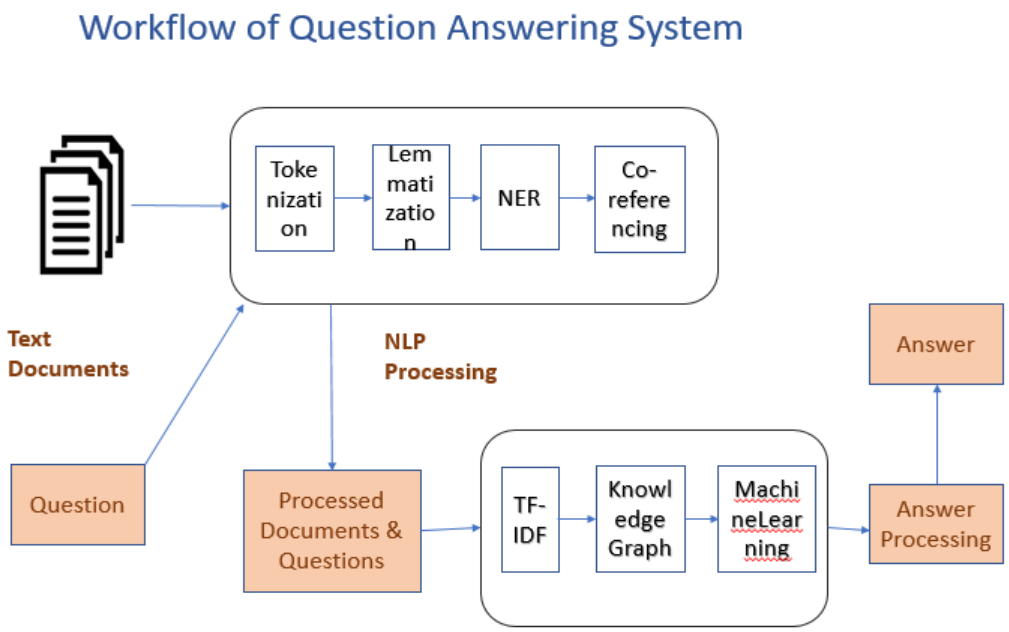
The data sets that we used for this project are available in the links followed. Basically, we have used two datasets in this project one of which is the news dataset and the other being sports. The links to the datasets are posted here.

BBC News - http://mlg.ucd.ie/datasets/bbc.html

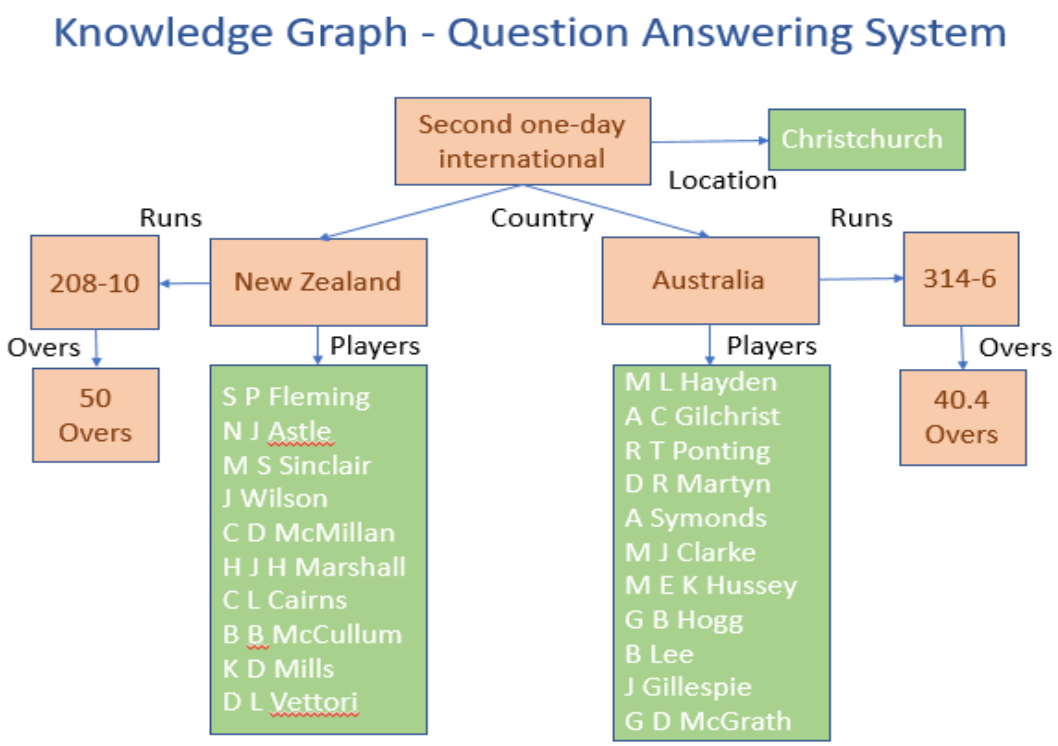
BBC Sports - http://mlg.ucd.ie/datasets/bbc.html

Design:

a. Workflow:



b. Knowledge Graph:



c. **Set of questions and answers:**

- What sport is discussed in the dataset?
- How many players are playing from Australia?
- Who has won the match?
- When was the match played?
- Which player has scored most number of runs?
- Are there any players injured in the match?
- Is there any possibility of Australia winning the match based on the commentary?

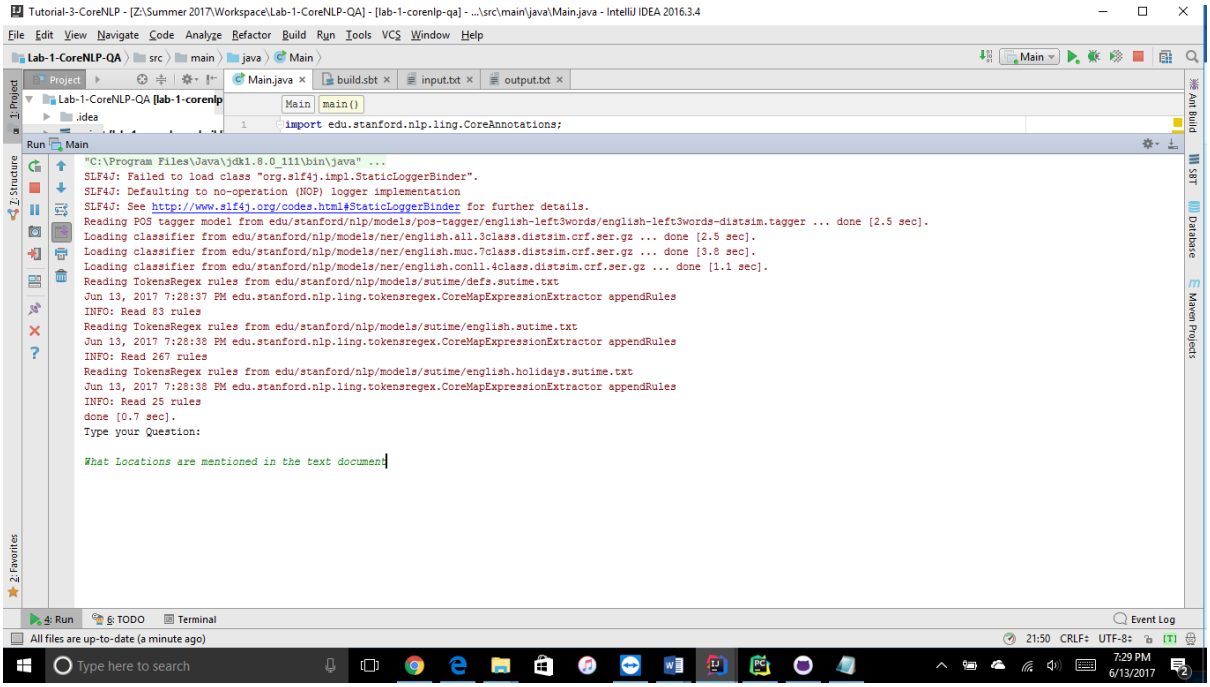
Implementation:

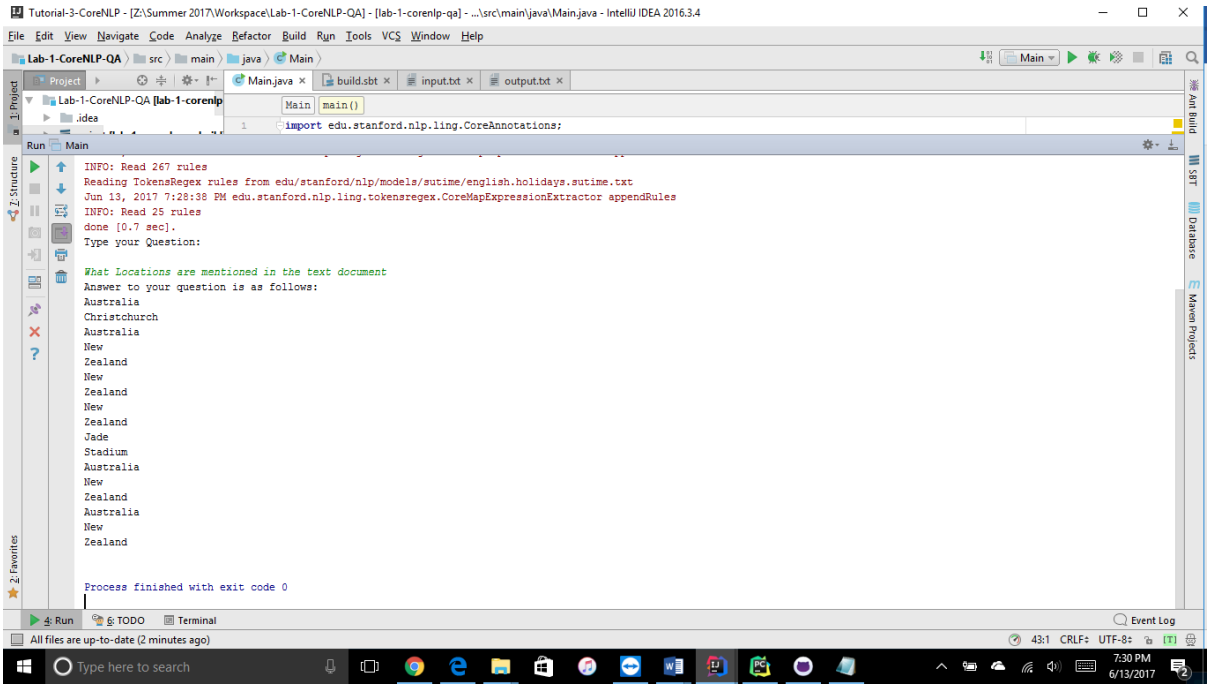
- We have used Natural Language processing (NLP) for tokenization, lemmatization, POS tagging, NER and Co-referencing.
- Used Stanford Core NLP for processing the selected dataset.
- The dataset used is related to BBC sports data on a match between Australia and New Zealand.
- Processed the input data taken from a text document and stored the information in different text files.
- Once after processing the question, we have used the data stored the output text files for answering them.

Technologies Used: Java, Python and Scala.

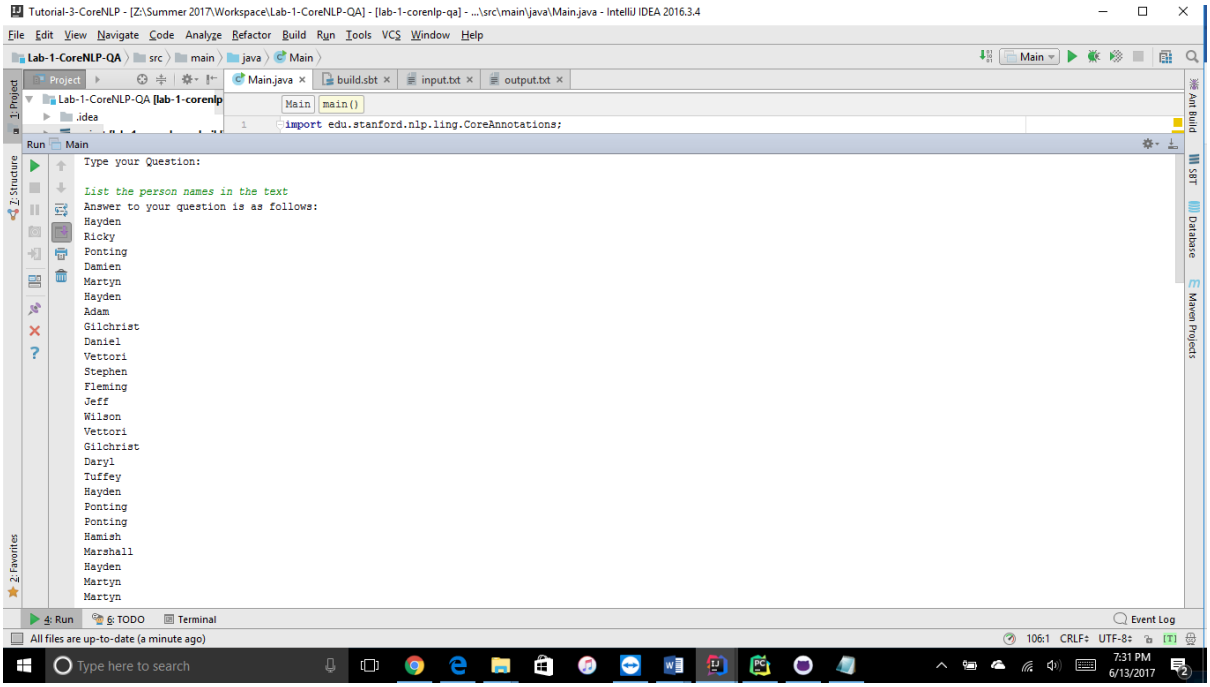
Please find the implemented Question Answering System in the below screenshots:

Question 1:

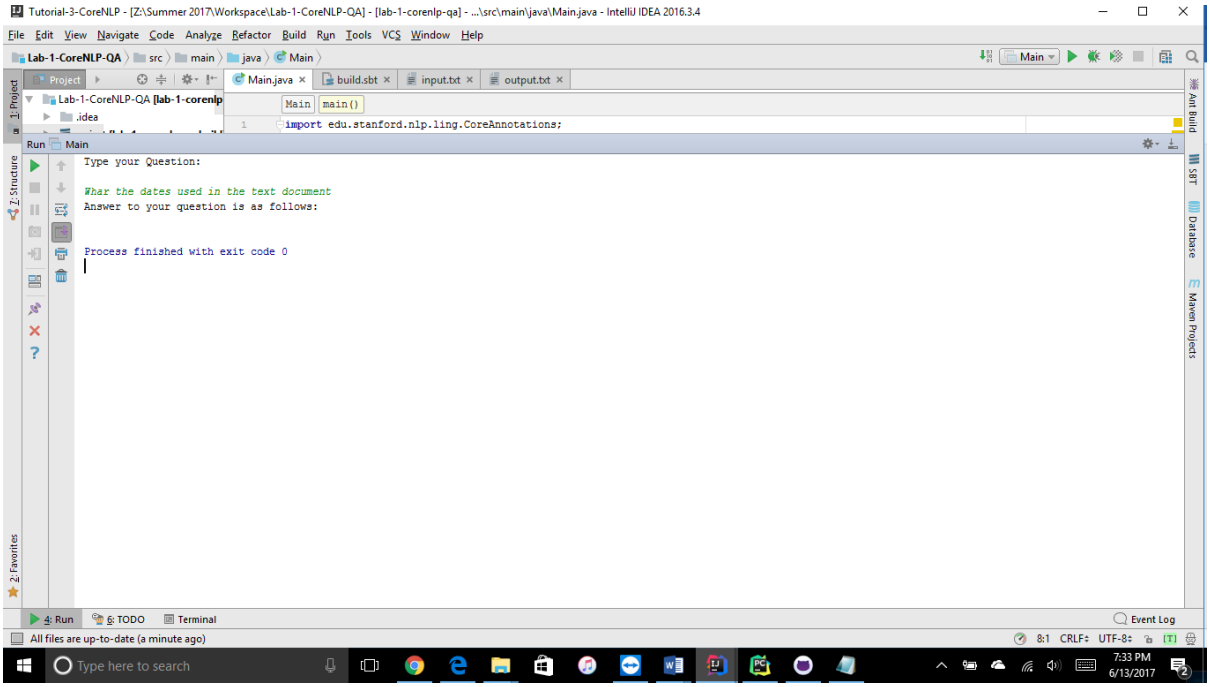

Tutorial-3-CoreNLP - [Z:\Summer 2017\Workspace\Lab-1-CoreNLP-QA] - [lab-1-corenlp-qa] - ...\src\main\java\Main.java - IntelliJ IDEA 2016.3.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Lab-1-CoreNLP-QA src main java Main
Main.java build.sbt input.txt output.txt
import edu.stanford.nlp.ling.CoreAnnotations;
Run Main
"C:\Program Files\Java\jdk1.8.0_111\bin\java" ...
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done [2.5 sec].
Loading classifier from edu/stanford/nlp/models/ner/english.all.3class.distsim.crf.ser.gz ... done [2.5 sec].
Loading classifier from edu/stanford/nlp/models/ner/english.muc.7class.distsim.crf.ser.gz ... done [3.8 sec].
Loading classifier from edu/stanford/nlp/models/ner/english.conll.4class.distsim.crf.ser.gz ... done [1.1 sec].
Reading TokensRegex rules from edu/stanford/nlp/models/sutime/defs.sutime.txt
Jun 13, 2017 7:28:37 PM edu.stanford.nlp.ling.tokensregex.CoreMapExpressionExtractor appendRules
INFO: Read 83 rules
Reading TokensRegex rules from edu/stanford/nlp/models/sutime/english.sutime.txt
Jun 13, 2017 7:28:38 PM edu.stanford.nlp.ling.tokensregex.CoreMapExpressionExtractor appendRules
INFO: Read 267 rules
Reading TokensRegex rules from edu/stanford/nlp/models/sutime/english.holidays.sutime.txt
Jun 13, 2017 7:28:38 PM edu.stanford.nlp.ling.tokensregex.CoreMapExpressionExtractor appendRules
INFO: Read 25 rules
done [0.7 sec].
Type your Question:
What Locations are mentioned in the text document
All files are up-to-date (a minute ago)
7:29 PM
6/13/2017


Tutorial-3-CoreNLP - [Z:\Summer 2017\Workspace\Lab-1-CoreNLP-QA] - [lab-1-corenlp-qa] - ...\src\main\java\Main.java - IntelliJ IDEA 2016.3.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Lab-1-CoreNLP-QA src main java Main
Main.java build.sbt input.txt output.txt
import edu.stanford.nlp.ling.CoreAnnotations;
Run Main
INFO: Read 267 rules
Reading TokensRegex rules from edu/stanford/nlp/models/sutime/english.holidays.sutime.txt
Jun 13, 2017 7:28:38 PM edu.stanford.nlp.ling.tokensregex.CoreMapExpressionExtractor appendRules
INFO: Read 25 rules
done [0.7 sec].
Type your Question:
What Locations are mentioned in the text document
Answer to your question is as follows:
Australia
Christchurch
Australia
New
Zealand
New
Zealand
New
Zealand
Jade
Stadium
Australia
New
Zealand
Australia
New
Zealand
Process finished with exit code 0
All files are up-to-date (2 minutes ago)
7:30 PM
6/13/2017

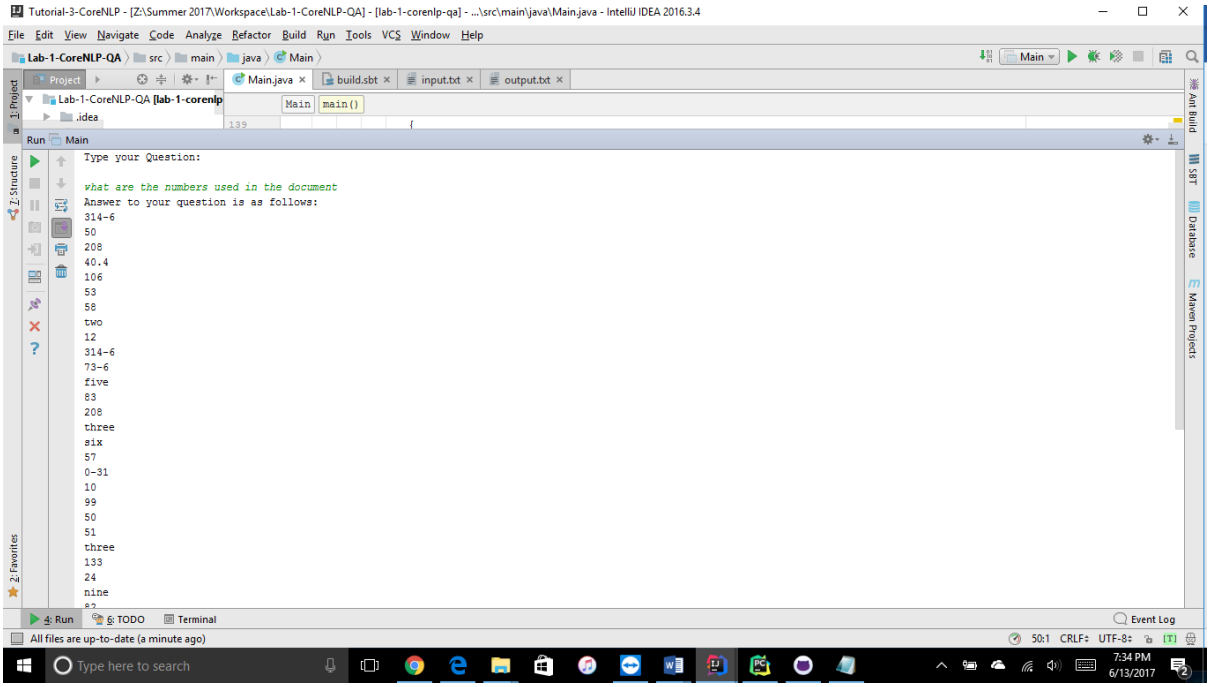
Question 2:



Question 3:



Question 4:


Tutorial-3-CoreNLP - [Z:\Summer 2017\Workspace\Lab-1-CoreNLP-QA] - [lab-1-corenlp-qa] - ...\src\main\java\Main.java - IntelliJ IDEA 2016.3.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Lab-1-CoreNLP-QA src main java Main
Main.java build.sbt input.txt output.txt
Run Main
Type your Question:
what are the numbers used in the document
Answer to your question is as follows:
314-6
50
208
40.4
106
53
58
two
12
314-6
73-6
five
83
208
three
six
57
0-31
10
99
50
51
three
133
24
nine
All files are up-to-date (a minute ago)
Type here to search
7:34 PM
6/13/2017

Project management:

a. **Contribution of each member:**

Name	Work Contribution	Percentage Contributed
Sandeep Pabolu	Spark Implementation, CoreNLP, Dataset Processing, Documentation	25%
Sri Sai Sarat Chandra Varma Mudunuri	Java Implementation, CoreNLP, Dataset Processing, Documentation	25%
Lava Kumar Surparaju	Java Implementation, CoreNLP, Dataset Processing, Documentation	25%
Rakesh Reddy Pallepati	Spark Implemenataion, CoreNLP, Dataset Processing, Documentation	25%

b. ZenHub and GitHub 's:

https://github.com/SaratM34/CS5560-KDM-Final-Project

Clear current search query, filters, and sorts

0 Open ✓ 9 Closed | Author | Labels | Projects | Milestones | Assignee | Sort

- **Exceptions** 3
  #9 by SaratM34 was closed 2 days ago · Testing 3
- **Function failure** 13
  #8 by SaratM34 was closed 2 days ago · Testing 3
- **Network Failure** 5
  #7 by SaratM34 was closed 2 hours ago · Testing 3
- **Validation errors** 5
  #6 by SaratM34 was closed 2 days ago · Testing 2
- **Login errors** 21
  #5 by SaratM34 was closed 2 days ago · Testing 2
- **Compile time errors** 8
  #4 by SaratM34 was closed 2 hours ago · Testing 2
- **Errors** 13
  #3 by SaratM34 was closed 2 days ago · Testing 1
- **Bug Fixing** 8
  #2 by SaratM34 was closed 2 days ago · Testing 1
- **UI Enhancements** 5
  #1 by SaratM34 was closed 2 hours ago · Testing 1

SaratM34 / KDM-Lab-Assignments

Unwatch 1 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Boards | Reports | Projects 0 | Wiki | Insights

View | Repos (1/1) | Show one | Labels | Milestones | Assignees | Epics | Releases | Search (/) | New Issue +

In Progress	Review/QA	Done	Closed
0 issues - 0 story points	0 issues - 0 story points	0 issues - 0 story points	9+ issues - 81 story points
			KDM-Lab-Assignments #4 Compile time errors · Testing 2 · 8
			KDM-Lab-Assignments #1 UI Enhancements · Testing 1 · 5
			KDM-Lab-Assignments #7 Network Failure · Testing 3 · 5

Add a Pipeline ...

Labels | Milestones

New milestone

3 Open ✓ 0 Closed

Sort

Testing 2

Past due by 1 day Last updated about 2 hours ago

100% complete 0 open 3 closed

Edit Close Delete

Testing 1

Past due by 1 day Last updated about 2 hours ago

100% complete 0 open 3 closed

Edit Close Delete

Testing 3

Past due by 1 day Last updated about 2 hours ago

100% complete 0 open 3 closed

Edit Close Delete

Burndown | Velocity tracking | Release report

Testing 1

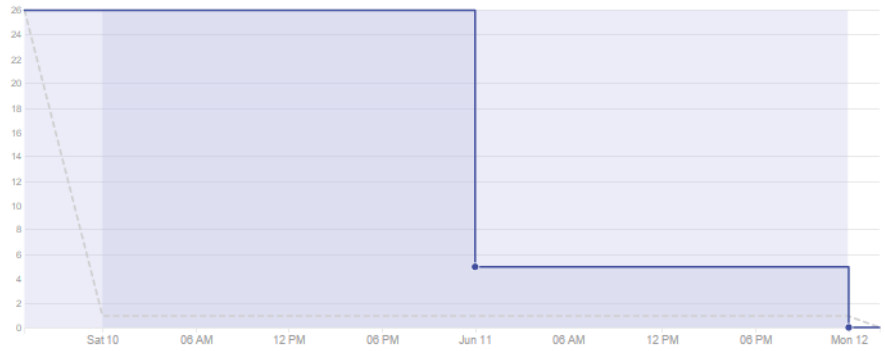
Start: **Jun 10, 2017** Edit Due: **Jun 12, 2017** Edit

Edit Milestone Milestones

Labels Hide Pull Requests Burn Pipelines

Burndown report

Weekends Ideal Completed



26 Total Story Points
26 Completed / 0 Remaining

3 Total Issues and Pull Requests
3 Completed / 0 Remaining

Burndown | Velocity tracking | Release report

Testing 2

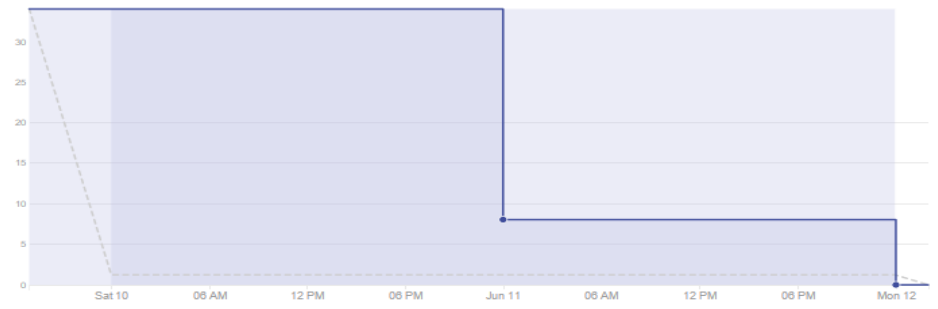
Start: **Jun 10, 2017** Edit Due: **Jun 12, 2017** Edit

Edit Milestone | Milestones

Labels | Hide Pull Requests | Burn Pipelines

Burndown report

Weekends | Ideal | Completed



34 Total Story Points
34 Completed / 0 Remaining

3 Total Issues and Pull Requests
3 Completed / 0 Remaining

Burndown | Velocity tracking | Release report

Testing 3

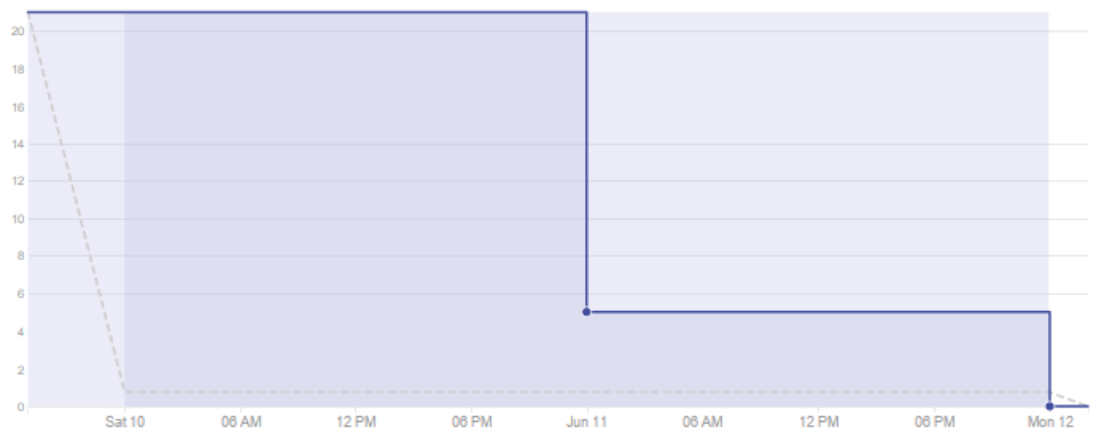
Edit Milestone | Milestones

Start: **Jun 10, 2017** Edit Due: **Jun 12, 2017** Edit

Labels | Hide Pull Requests | Burn Pipelines

Burndown report

Weekends | Ideal | Completed



21 Total Story Points
21 Completed / 0 Remaining

3 Total Issues and Pull Requests
3 Completed / 0 Remaining

c. Concerns and Issues:

- Notable issues are regarding processing data for some set of questions.
- Filtering stop words, synonyms.
- Co-referencing for complex datasets.
- Relation between the persons in the text.
- Related words questions need to be achieved.

d. Future Work:

In our next project report we will be extending our current questions set to much more complex set. Also, Processing data using N-gram, W2V. We need to implement the relation analysis, enrich the question type with yes/no answers. Full-Fledged Implementation of the Knowledge Graph for the Dataset . Text classification for the dataset using Shallow/Deep Learning.