TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

KNOWLEDGE DISCOVERY AND MANAGEMENT

SUMMARIZATION

INSTRUCTOR:

Dr. Yugyung Lee

TEAM 8:

VILAS MAMIDYALA

VIKESH PADARTHI

DINESH KUMAR BANDAM

BHUMIREDDY RANJITHA REDDY

TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

FIRST INCREMENT REPORT - SUMMARIZATION

1. Motivation:

We know that the whole world is awaiting to hear the result of US election which are going to be released by the end of this year. Everyone would like to see how these elections are going to be held. One has an anxiety that who is going to win and what actually the people opinion is and who has more probability to win. These questions stimulate our work towards collecting data about politics which clears all our skeptic things about elections. Since many of the things related to students and their future who have more excitement and worry to get to know the result. Our main motivation behind this project is to analyze the data present in social media like twitter and plot some graphs which shows about which candidate is more famous in social media, the probability of who will be getting elected.

Objective:

Main objective of this project is to use NLP, machine learning knowledge to predict the outcome of election result. Using these we can summarize the result of various blogs, news, and editorial matters in newspapers which are related to elections. We will first plot some graphs based on the twitter data which we have collected. And we want to analyze various text data present in the World Wide Web like Wikipedia and summarize these papers.

Expected outcomes:

By performing these operations using NLP, Machine Learning we want to predict the outcome of the US elections and various views about US elections by the people around the world. The output will be ontology graphs which are developed by analyzing the data sets which are related to US elections.

2. Domain:

Data Set: Twitter Data, provided data sets by Lee.

Technologies: Java, Scala.

Topic: US Politics

IDE : IntelliJ

3. Data Collection:

Twitter data using JAVA and Linux.

4. Task and Features:

- Collected Twitter data using Java code.
- Link for the source code is:

  https://github.com/vilasmamidyala/KDM_SM16_SM/tree/master/Source/twit

- NLP processing has been applied to the sample input collected above .

  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Nlp%20Output.txt
  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Simplecorenlpoutput.txt

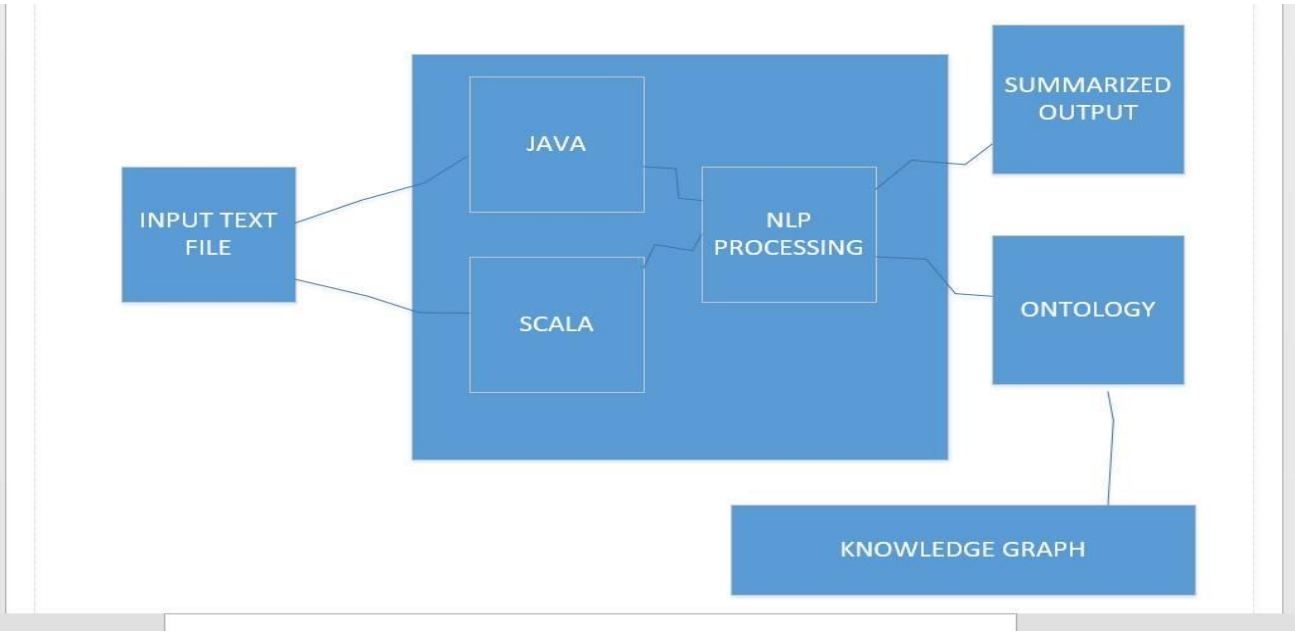
- Word count has been applied to the given same input :

  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/wordcount_output.txt

- Information Extraction/Retrieval technologies :
  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/wordcount_TFID.txt

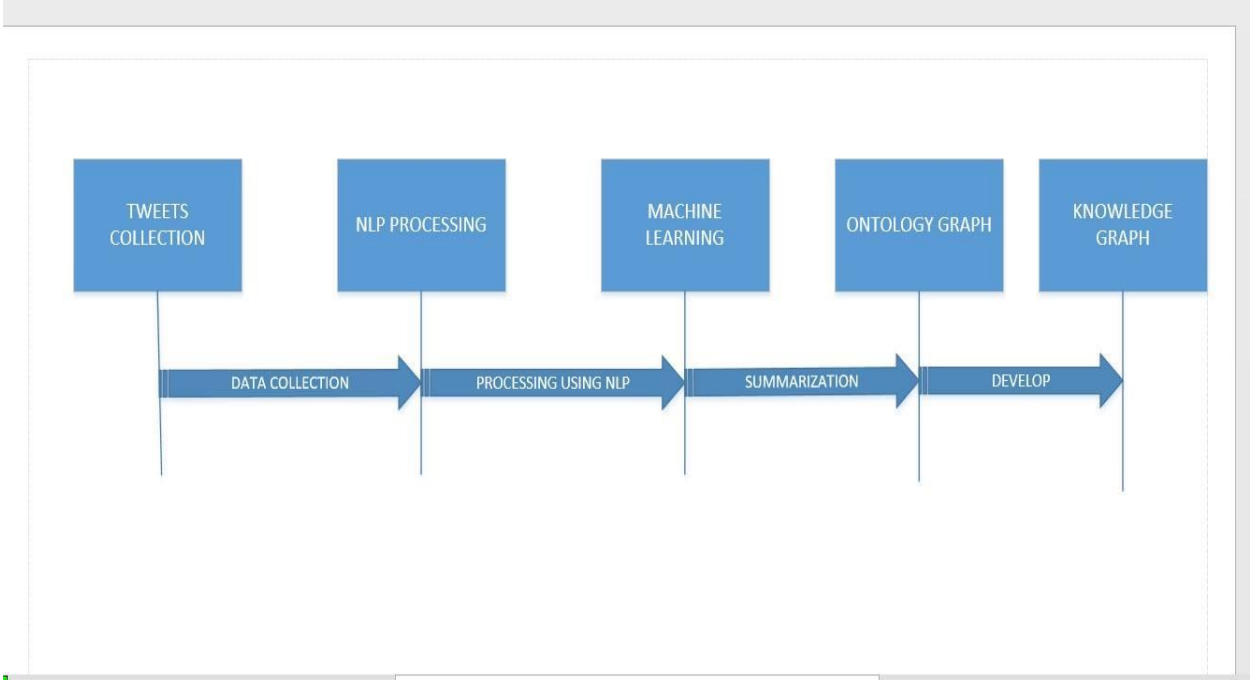
5. Implementation specification:

Software Architecture :

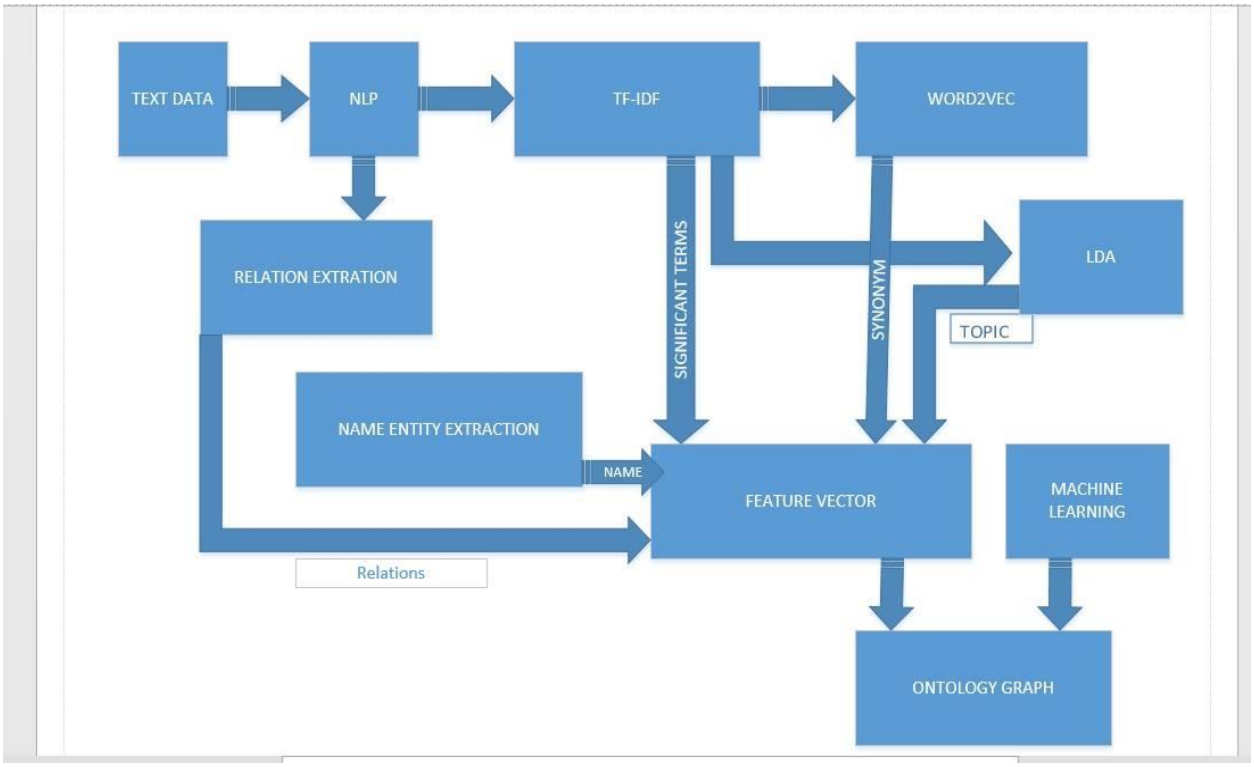


TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

SEQUENCE DIAGRAM :



WORKFLOW DIAGRAM :



Existing Services Used:

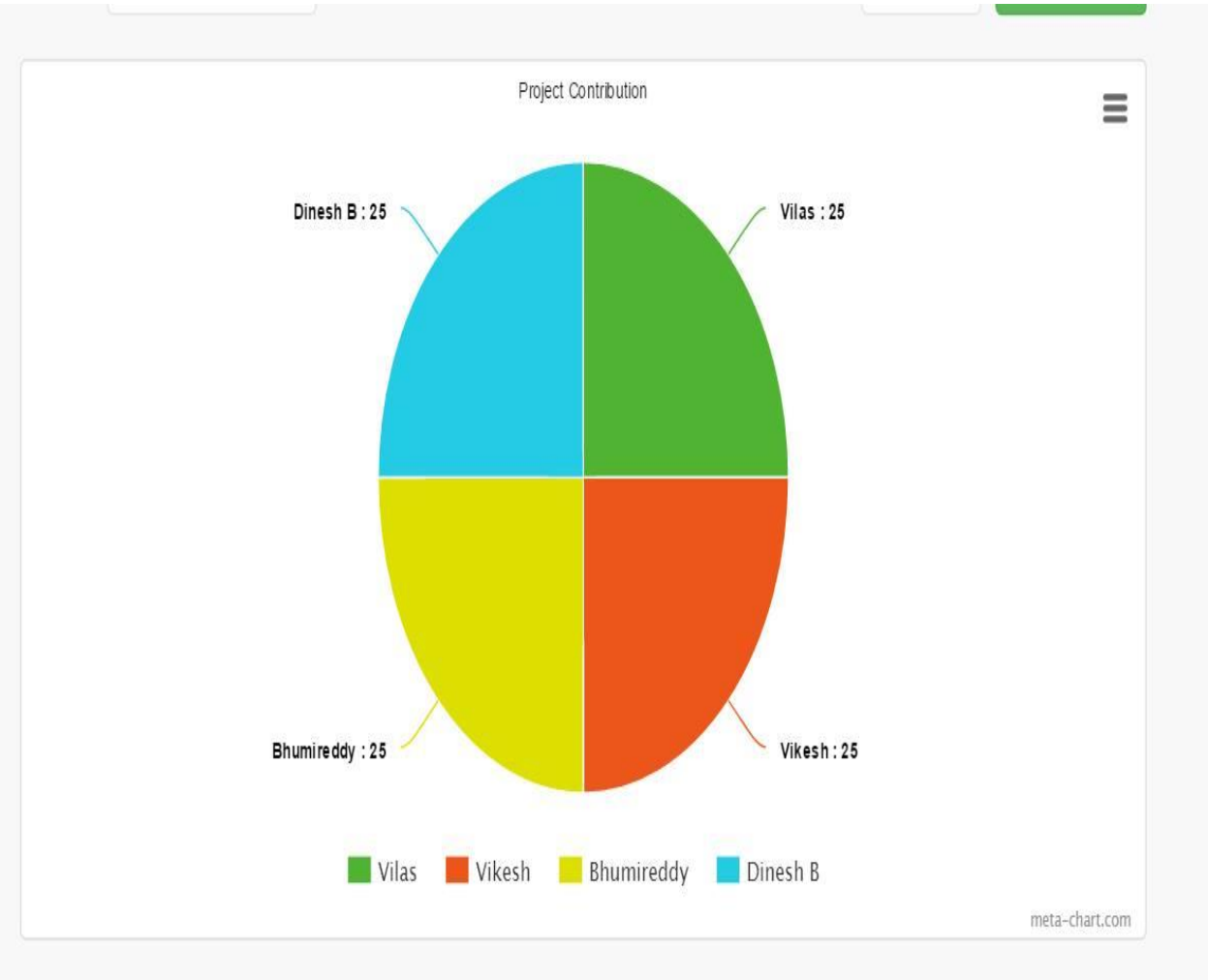
- Implemented word count program using Scala.
- Implemented NLP program.
- Implemented TF-IDF.

New Services:

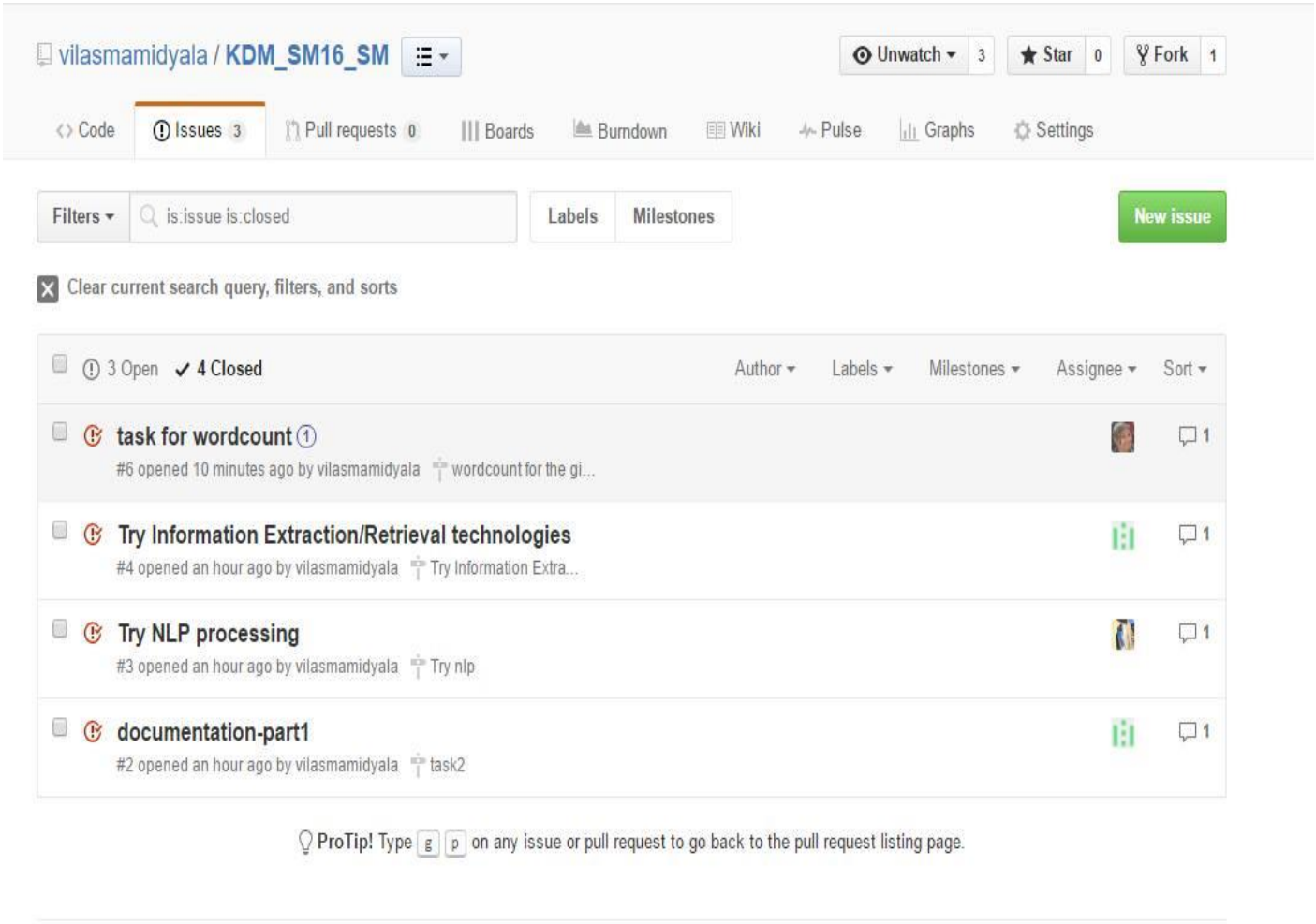
Tweet collection using Java Code.

6. Project Management:

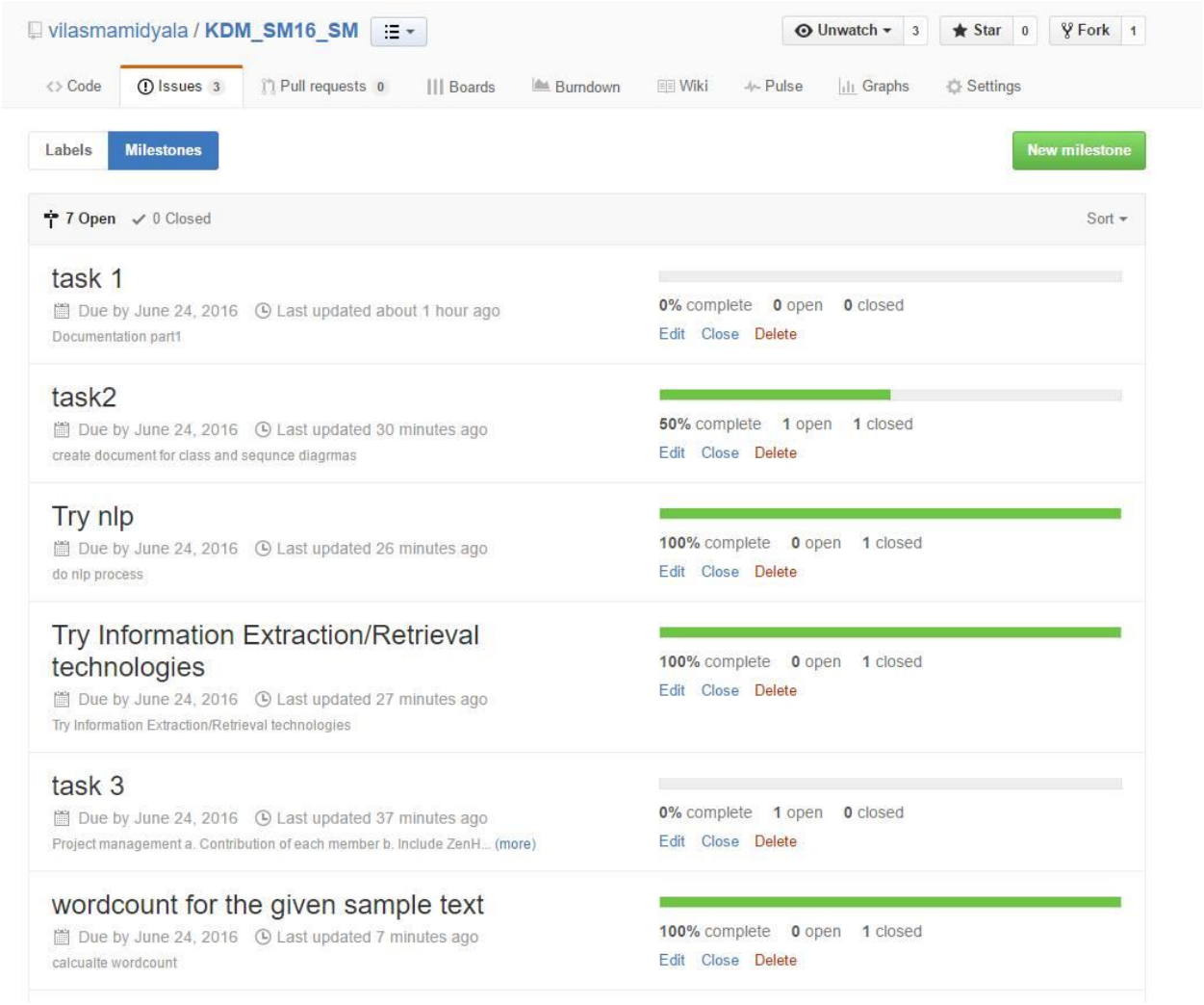
Contribution of Each member:



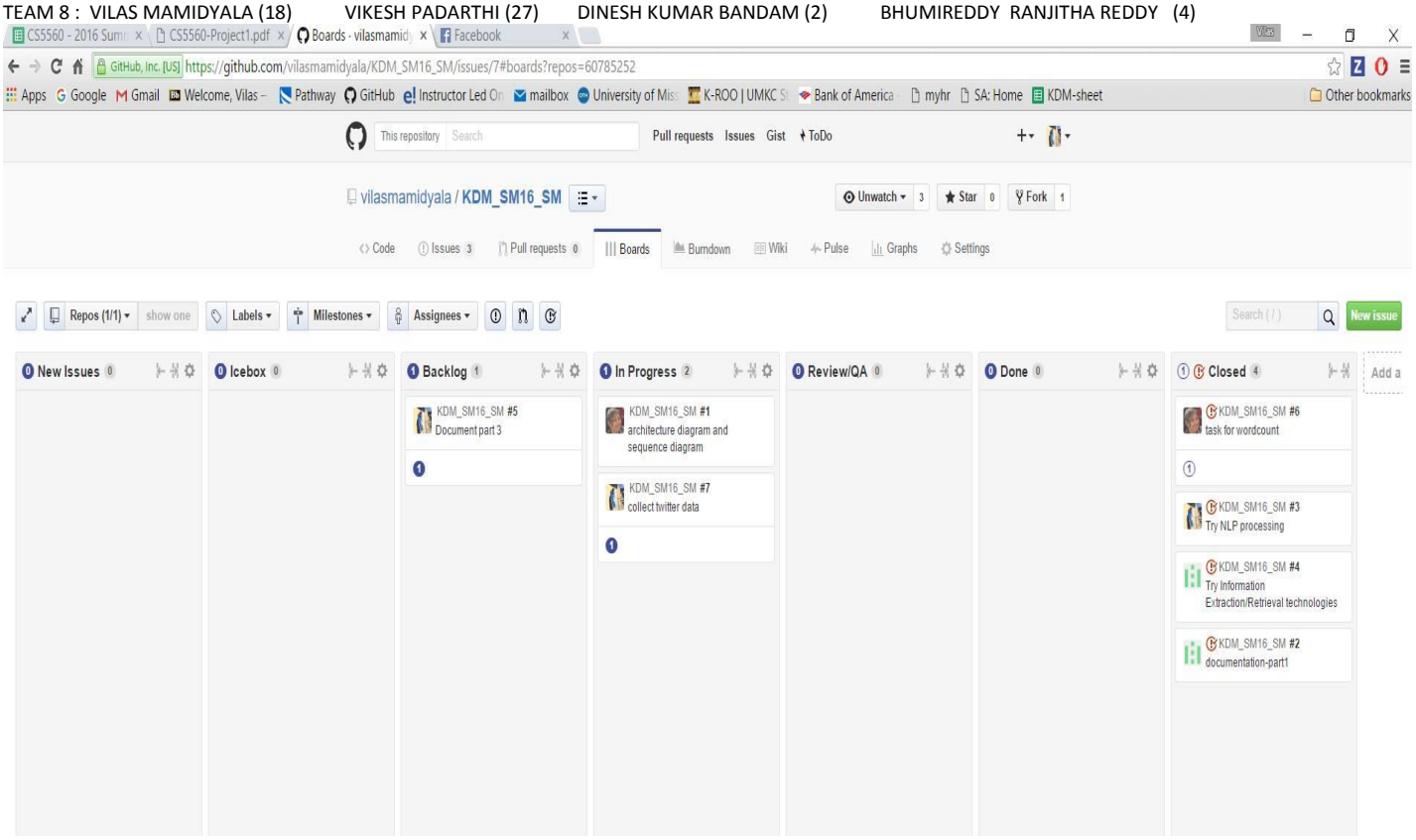
Zenhub and Github Screen shots:



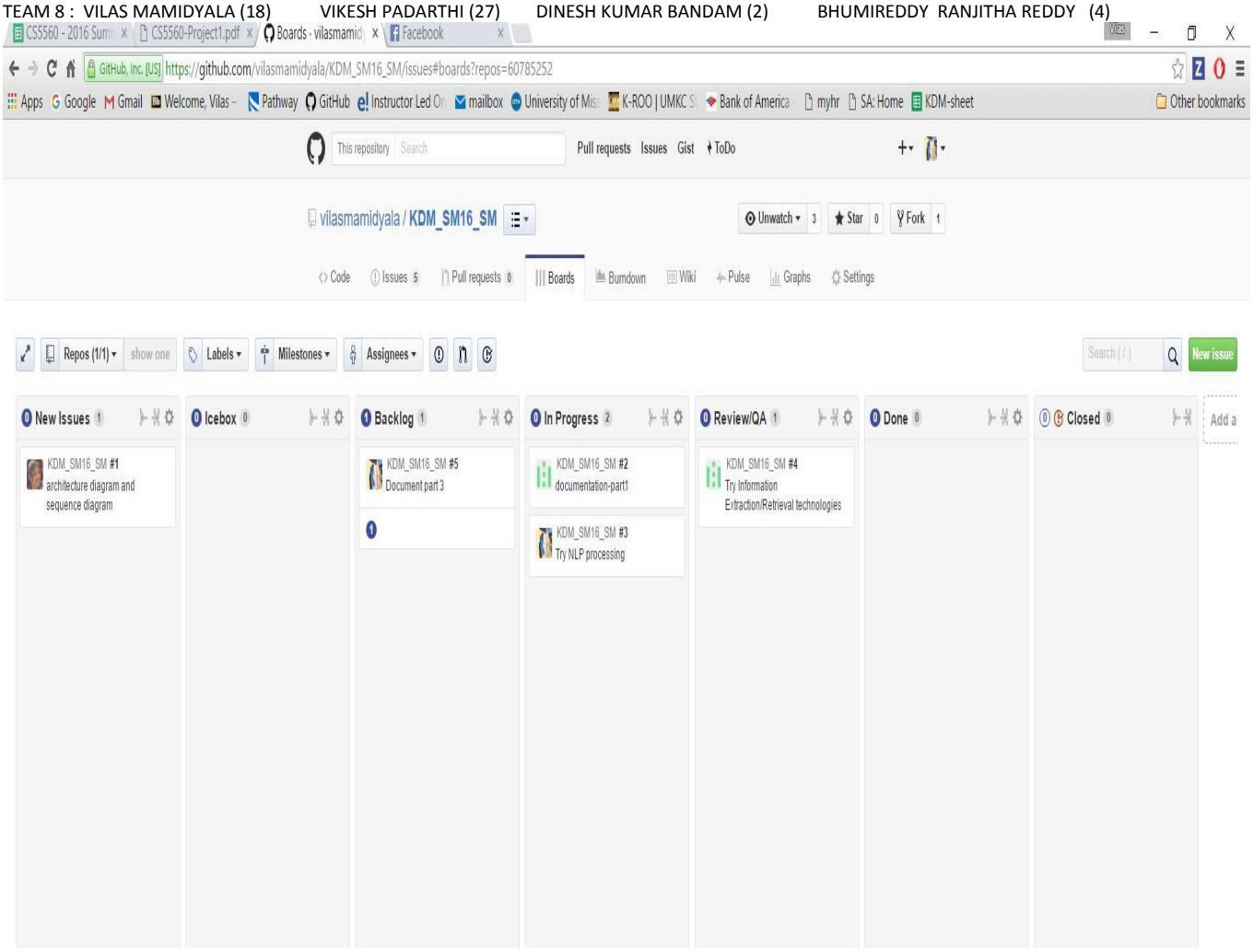
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)



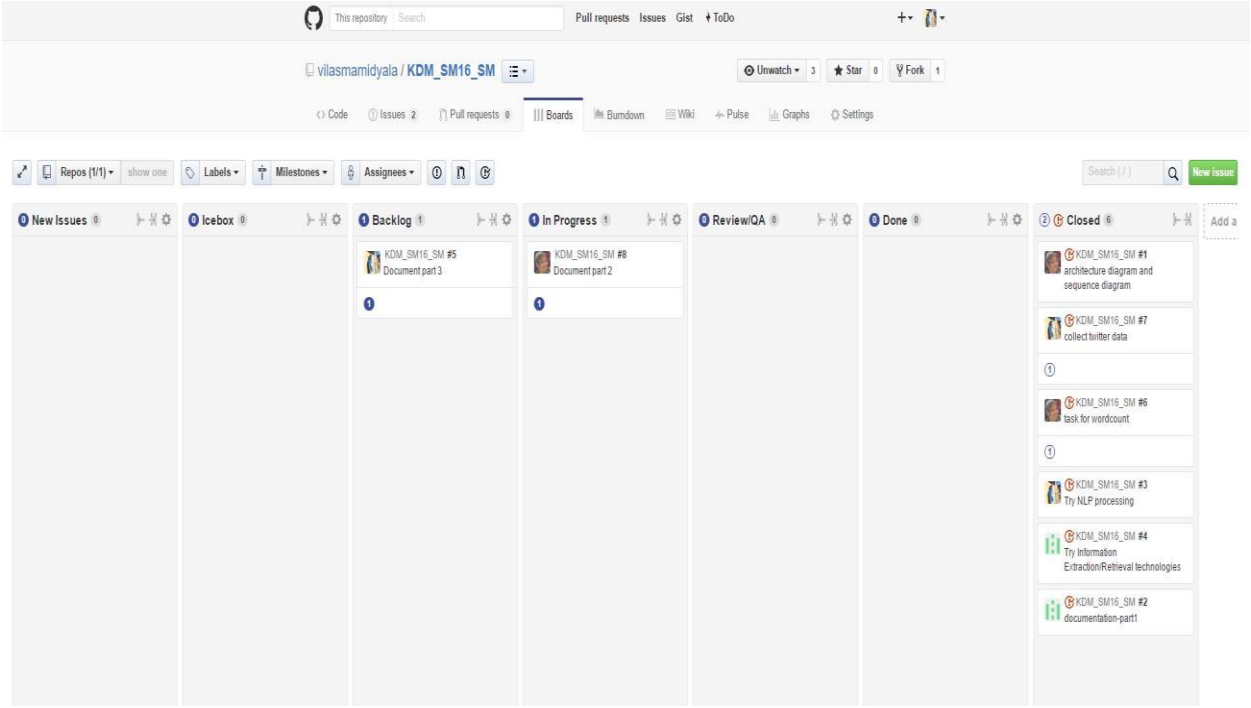
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)



TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)



TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)



Feature concerns/Issues:

1) For small amount of data given as input for NLP processing an d for other code executions. We found that these programs are working well and giving better results. The issue has occurred when we had tried implement NLP operation on large amount of data the programs were not able to run properly.
2) We considered taking Twitter data for the first phase. But we want to know whether twitter data can be useful for summarization? Because each tweet will be independent of the other tweets most of the times. This data alone might not help us for summarization. we think we need to take other different source s of data as well. we will try to figure out about what are the other sources that can be included.

Future Work:

In our further increments we would like focus on how to implement NLP operations on a bit of huge amount of data. We would like to do Word2Vec and LDA analysis on our data and then to get the feature vector for the data. We would like to implement Machine learning and ontology to derive final graphs.