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)

FINAL 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

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

3. Data Collection:

Twitter data using JAVA and Linux.

4. Task and Features:

- Source data
  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/output_word2vec.txt
- 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

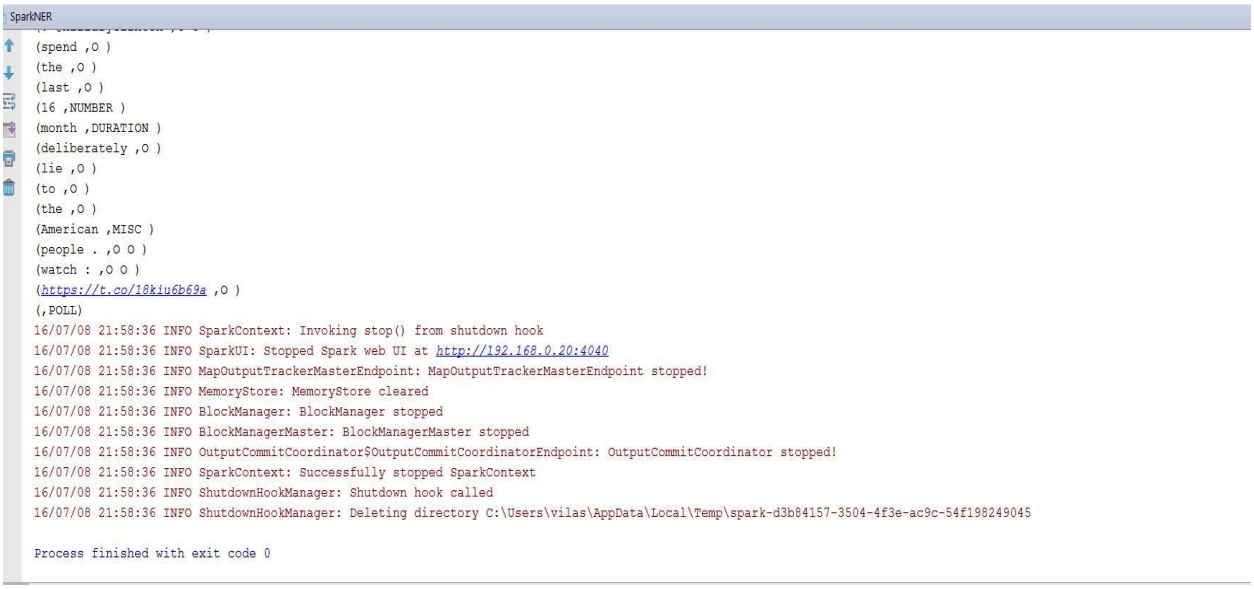
- Word count has been applied to the given same input :

  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

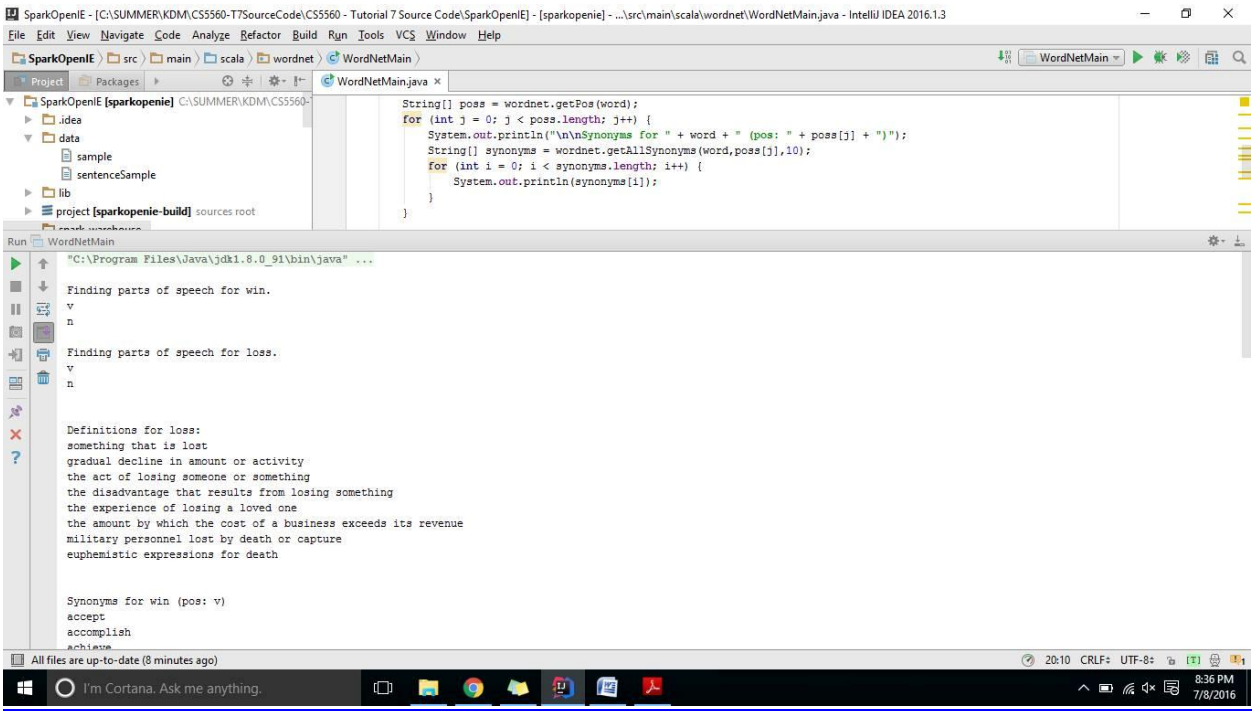
- Name Entity Extraction/Relation Extraction
  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/sparkner.pdf

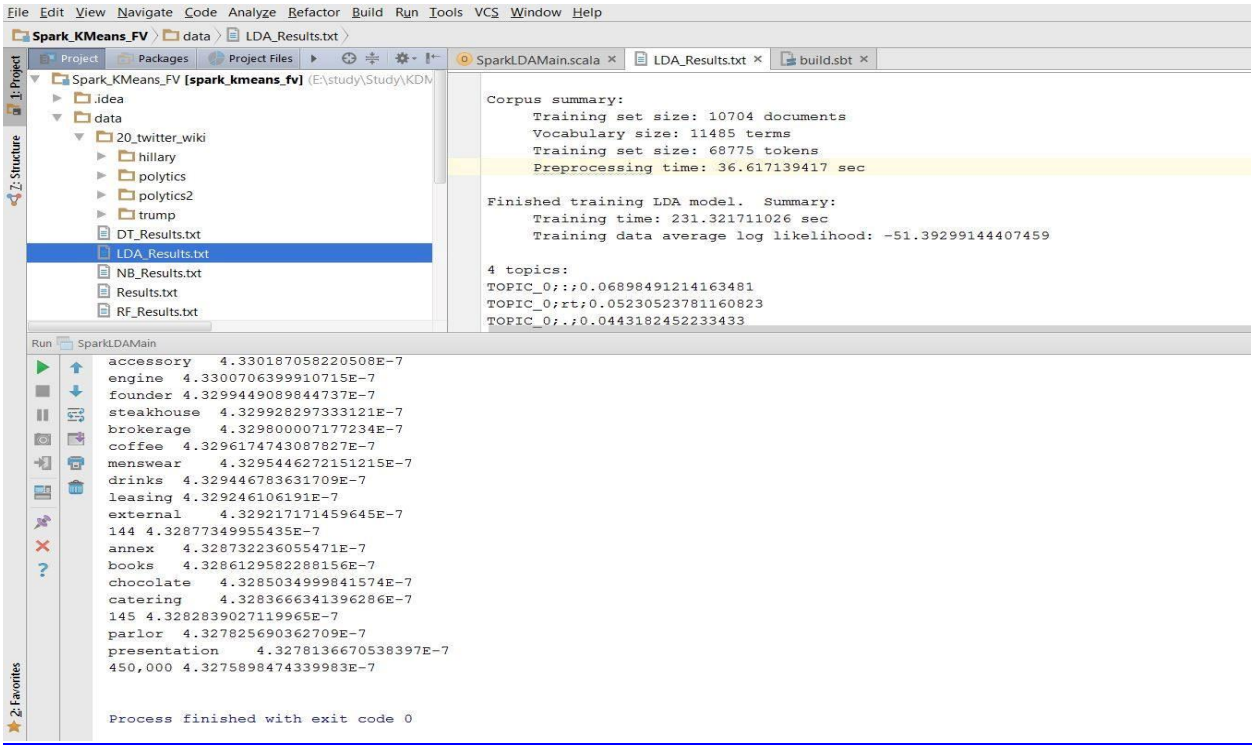
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)



- WordNet
  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Wordnet_output.docx
- Topic Discovery
  LDA:
  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/LDA_Results.txt

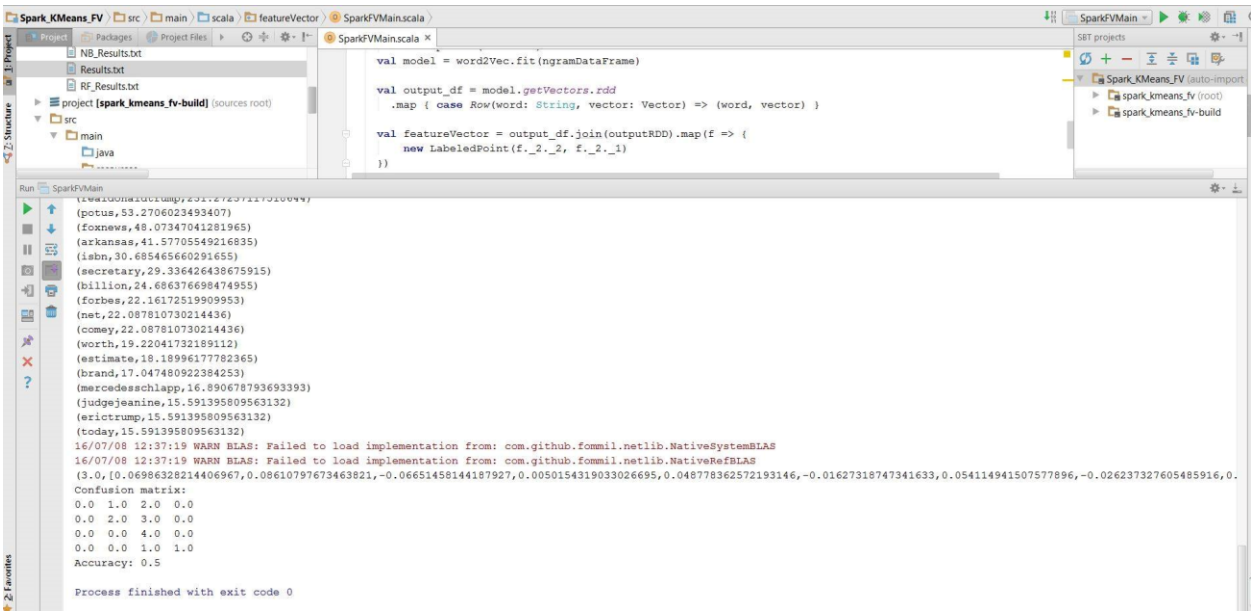
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)





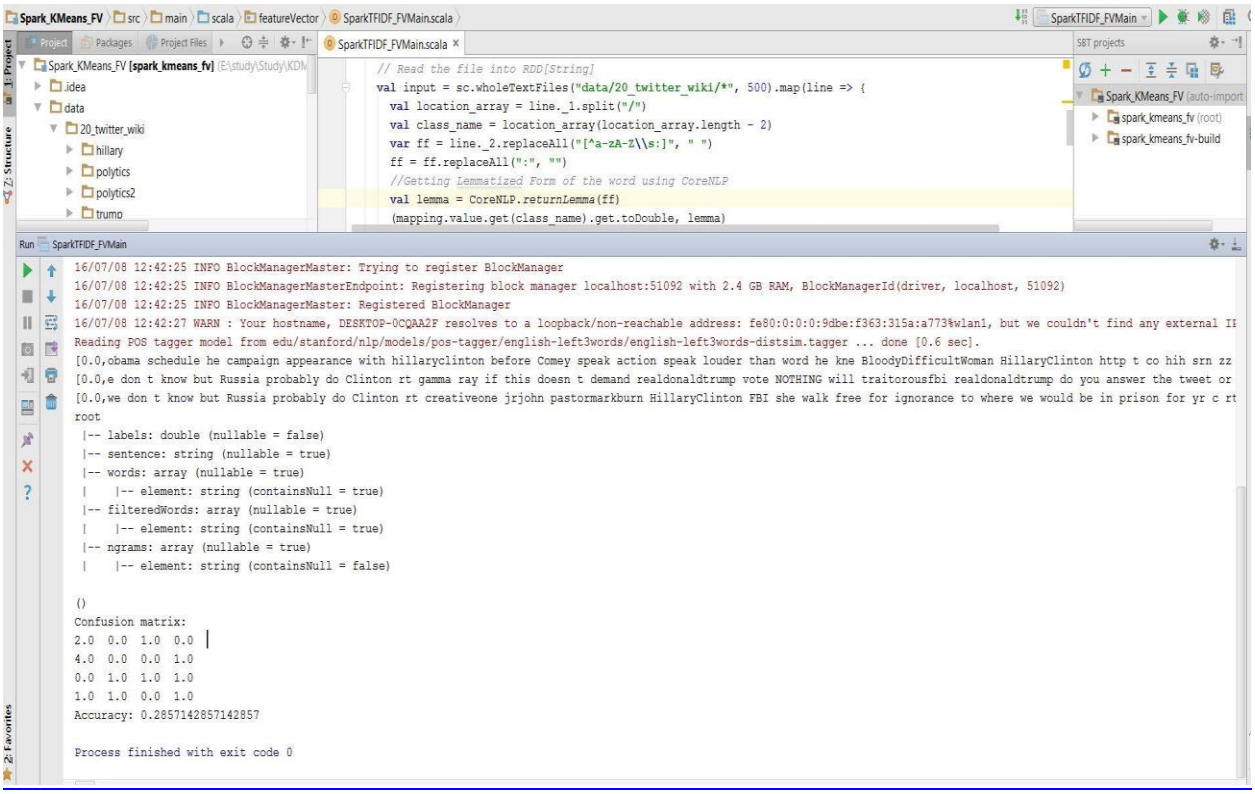
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

- Feature Vector for Machine Learning
  https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/spark_fv_output.pdf



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

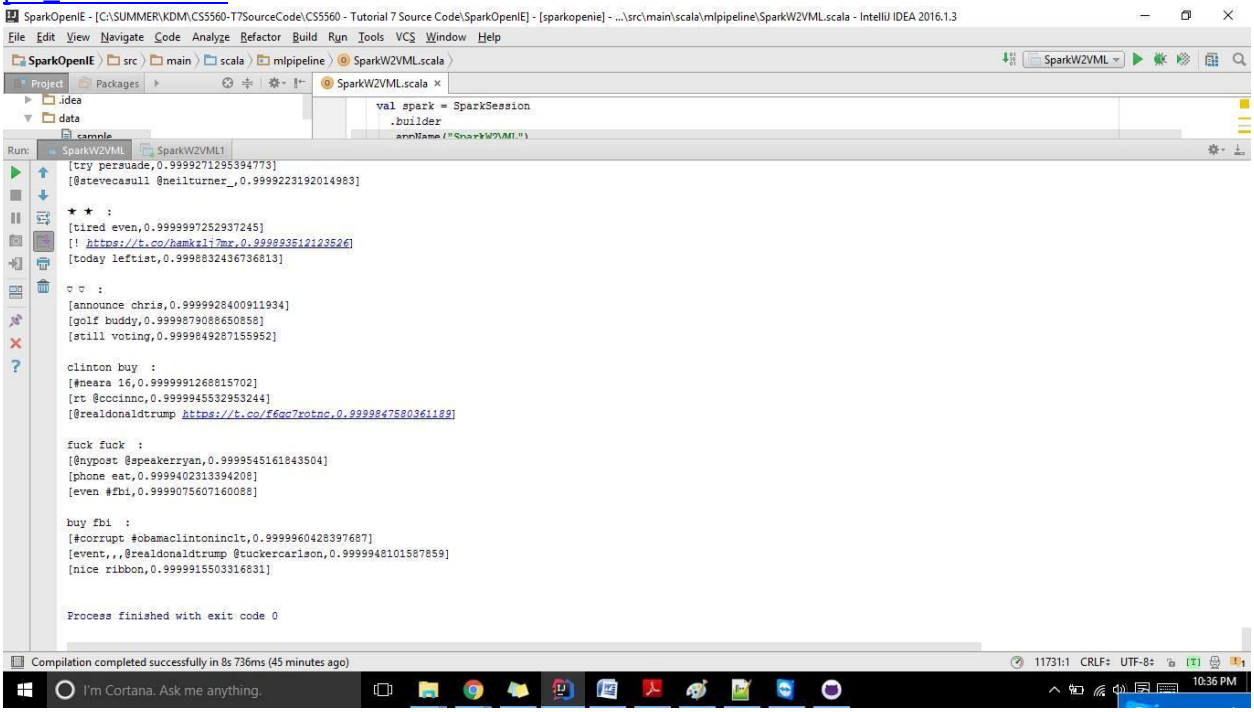
https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/spark_fv_output.pdf



- NGram and Word2Vec:

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

- https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/output_word2vec.txt



Ontology Execution outputs:

https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Ontology%20execution%20outputs.pdf

SOURCE CODE URL:

https://github.com/vilasmamidyala/KDM_SM16_SM/tree/master/Source/Spark2Ont

```
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done [1.0 sec].
root
 |-- location: string (nullable = true)
 |-- docs: string (nullable = true)
 |-- rawTokens: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- tokens: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- features: vector (nullable = true)
 |-- idfFeatures: vector (nullable = true)
```

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

```
Corpus summary:
     Training set size: 2 documents
     Vocabulary size: 13 terms
     Training set size: 15 tokens
     Preprocessing time: 0.558577732 sec

Finished training LDA model.  Summary:
     Training time: 4.352653153 sec
     Training data average log likelihood: -20.115197054435626

3 topics:
president win usa election state currently majority female clinton lead oppostion trump virtue
usa win president virtue trump oppostion lead majority female currently clinton election state
usa win president clinton state lead currently election female majority virtue trump oppostion
0.0
0.0
0.0
```

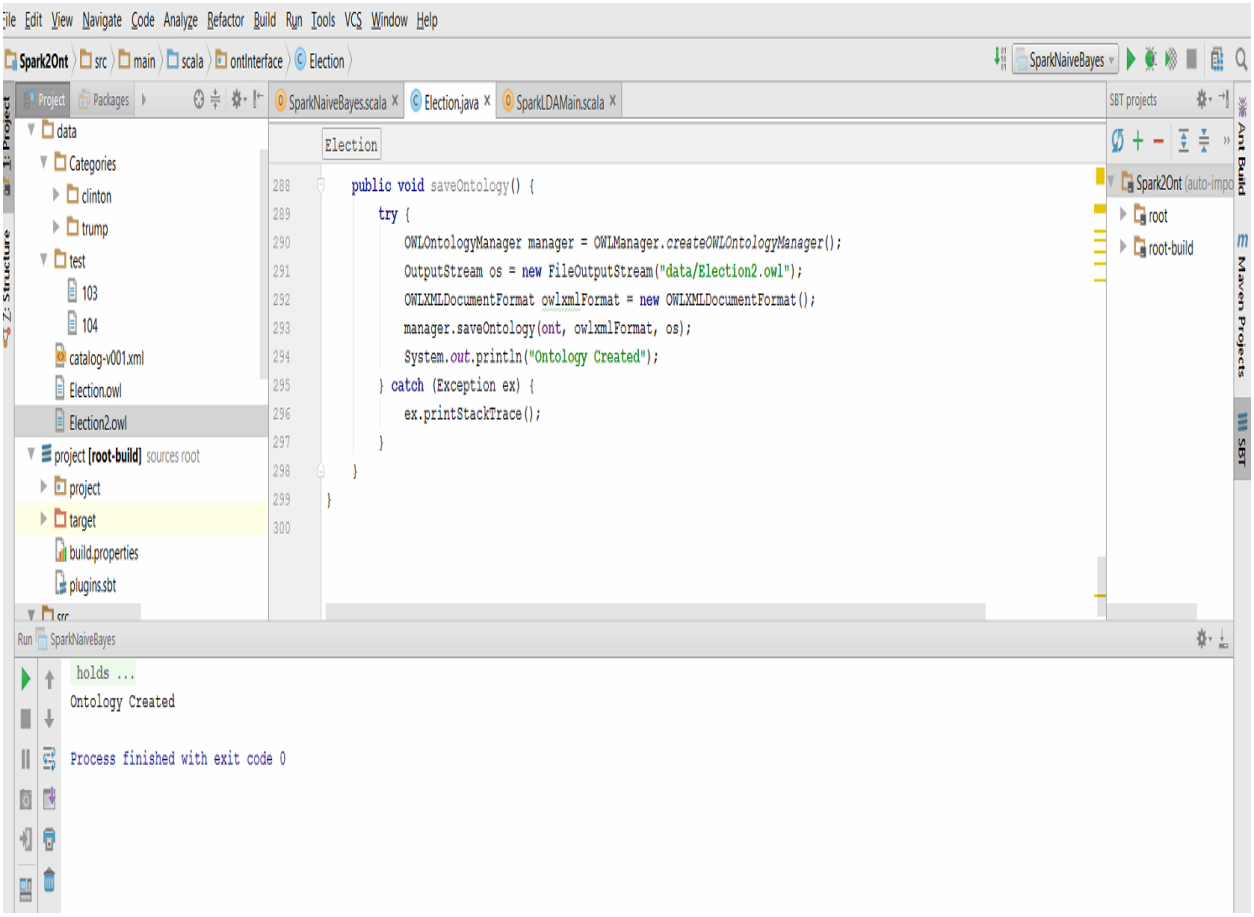
SparkNaiveBayes

```
holds ...
Ontology Created


Process finished with exit code 0
```

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

Election.java is where we write code for generating our ontology using the classes mentioned in the spark program using naïve Bayes algorithm



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

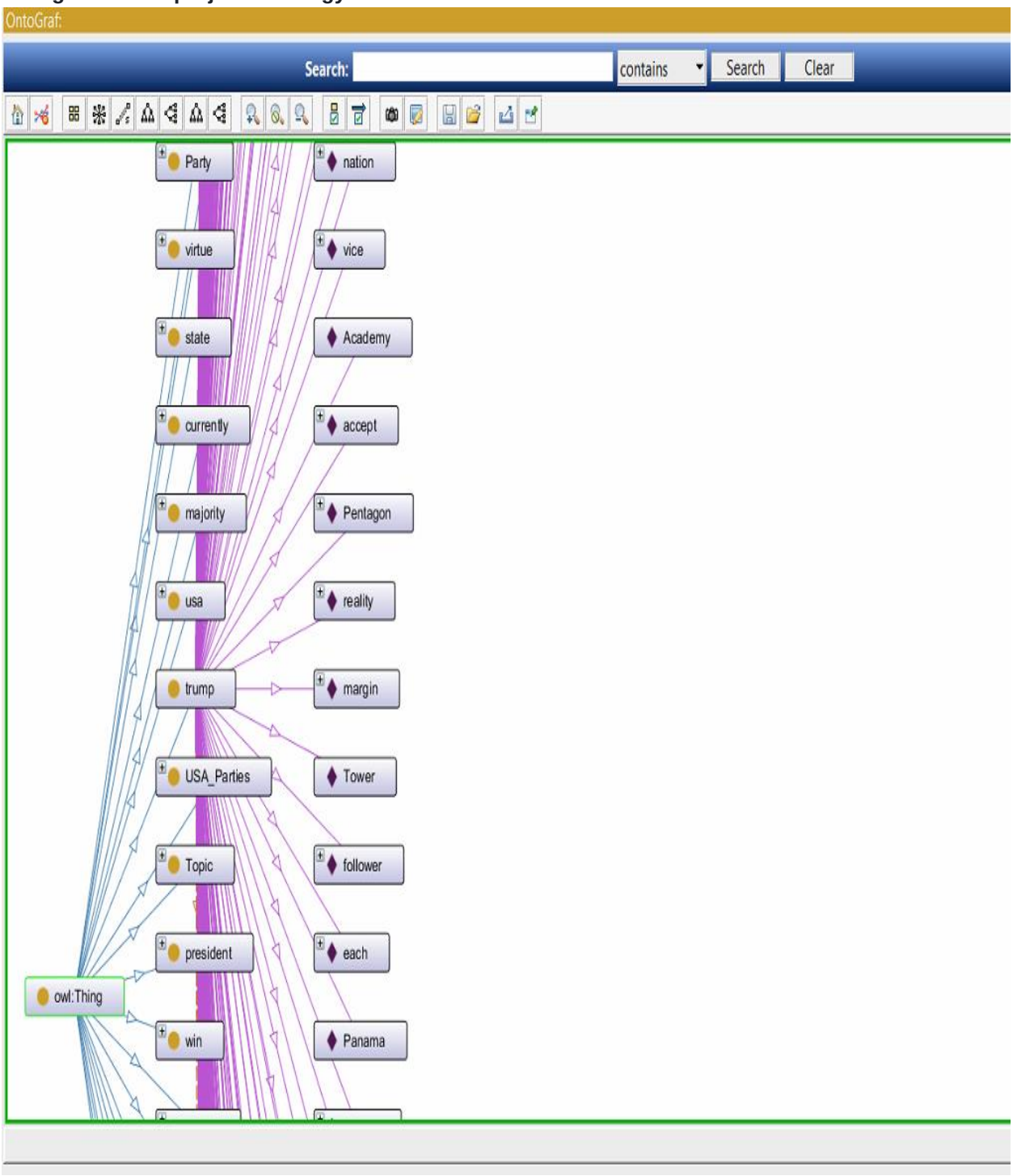
Ontology Election OWL File in Protégé Outputs:

https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Ontoin%20protege_outputs.pdf

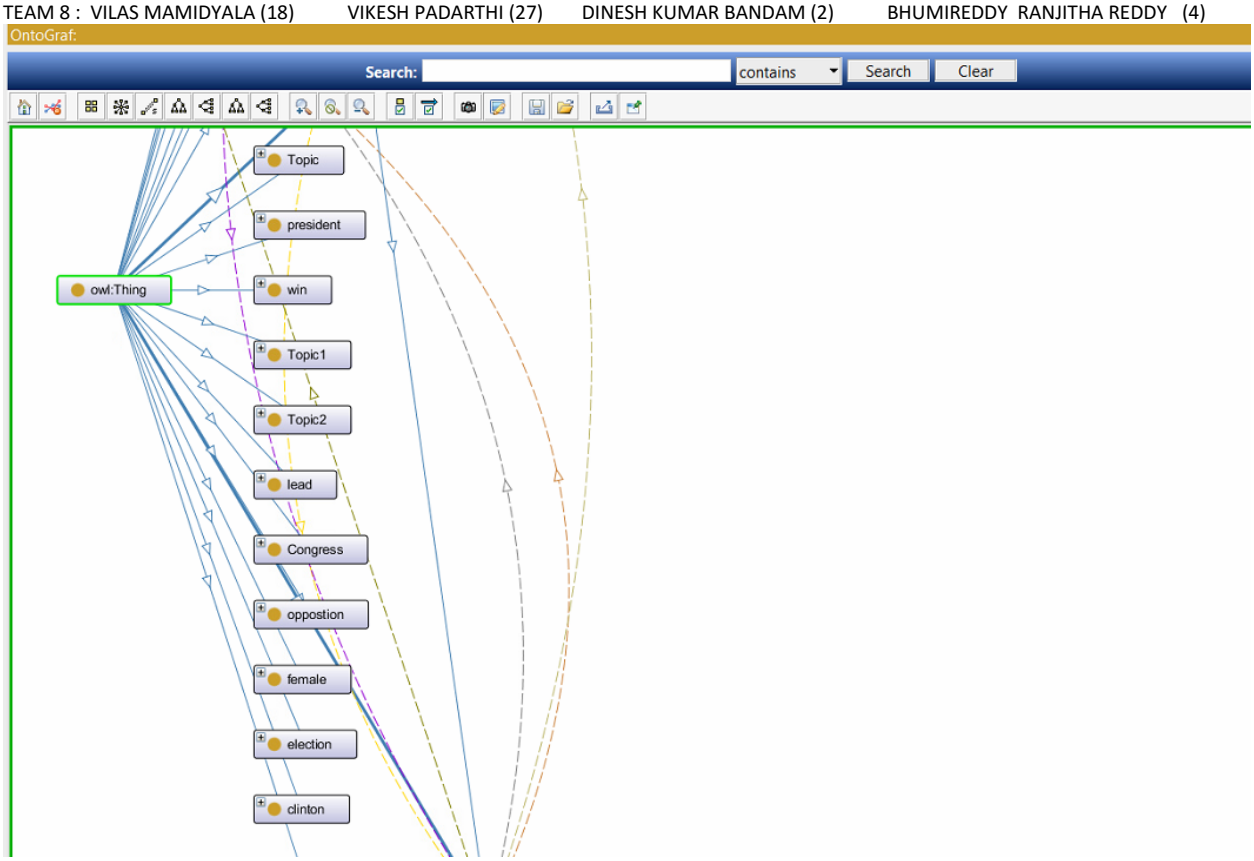
Class Hierarchy:

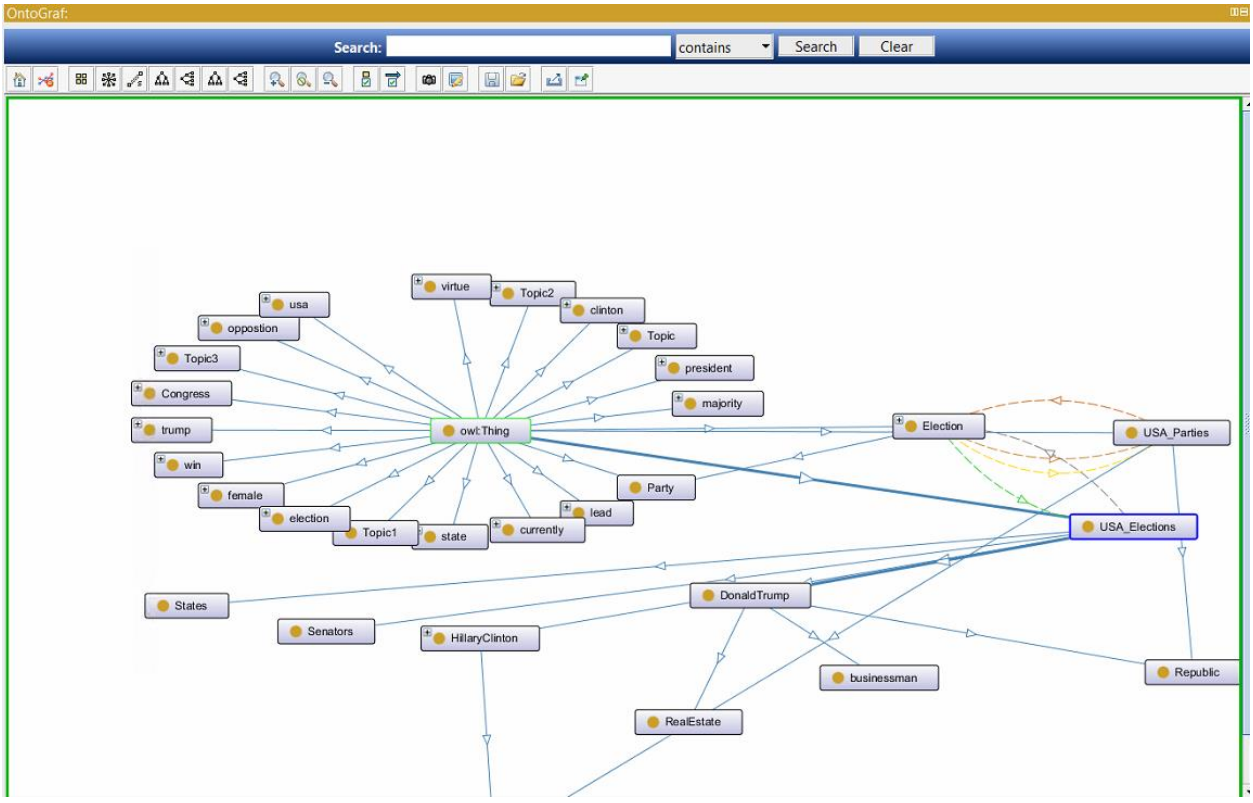
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

Ontograf for our project Ontology:



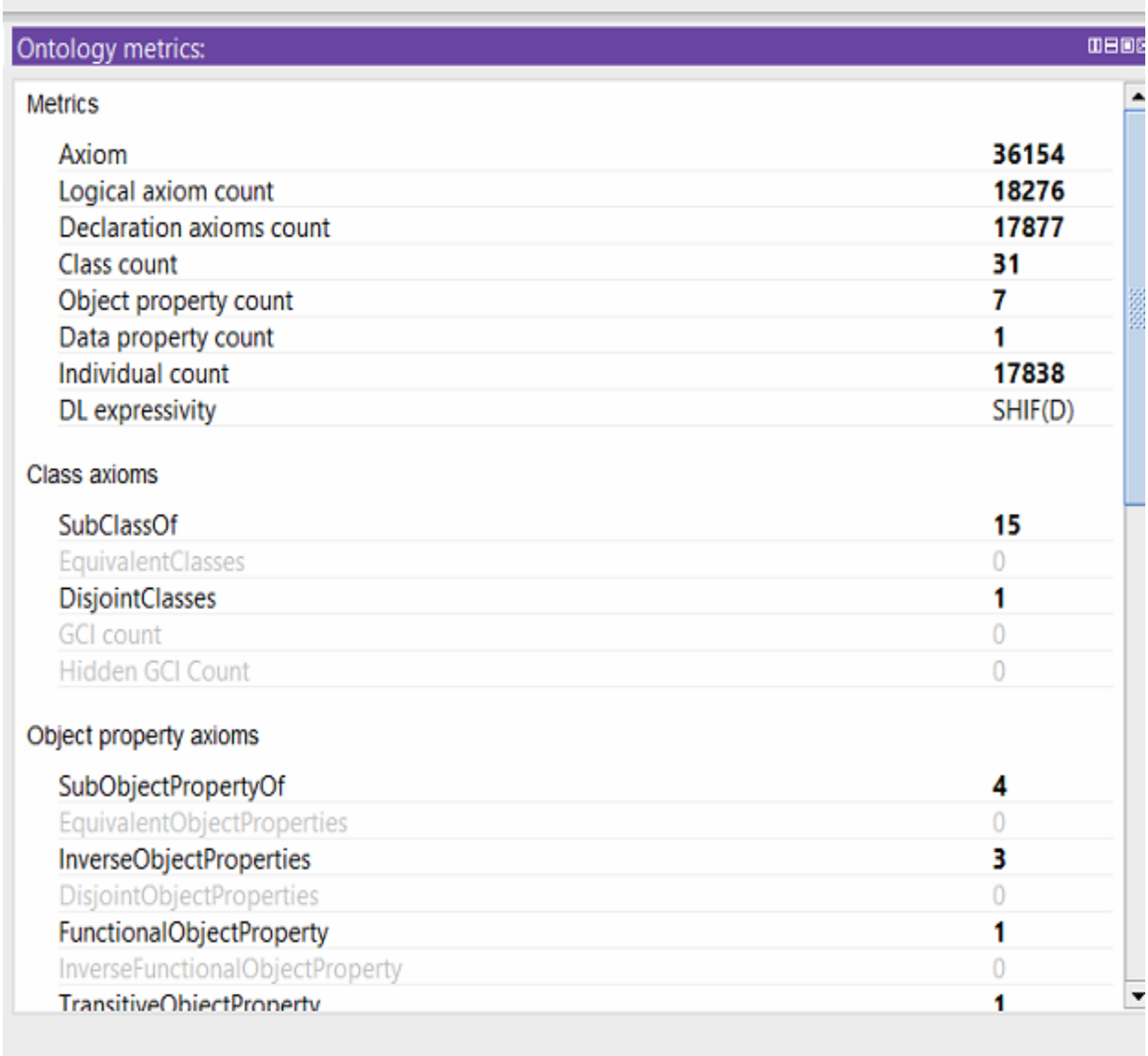
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)





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

Ontology Metrics:

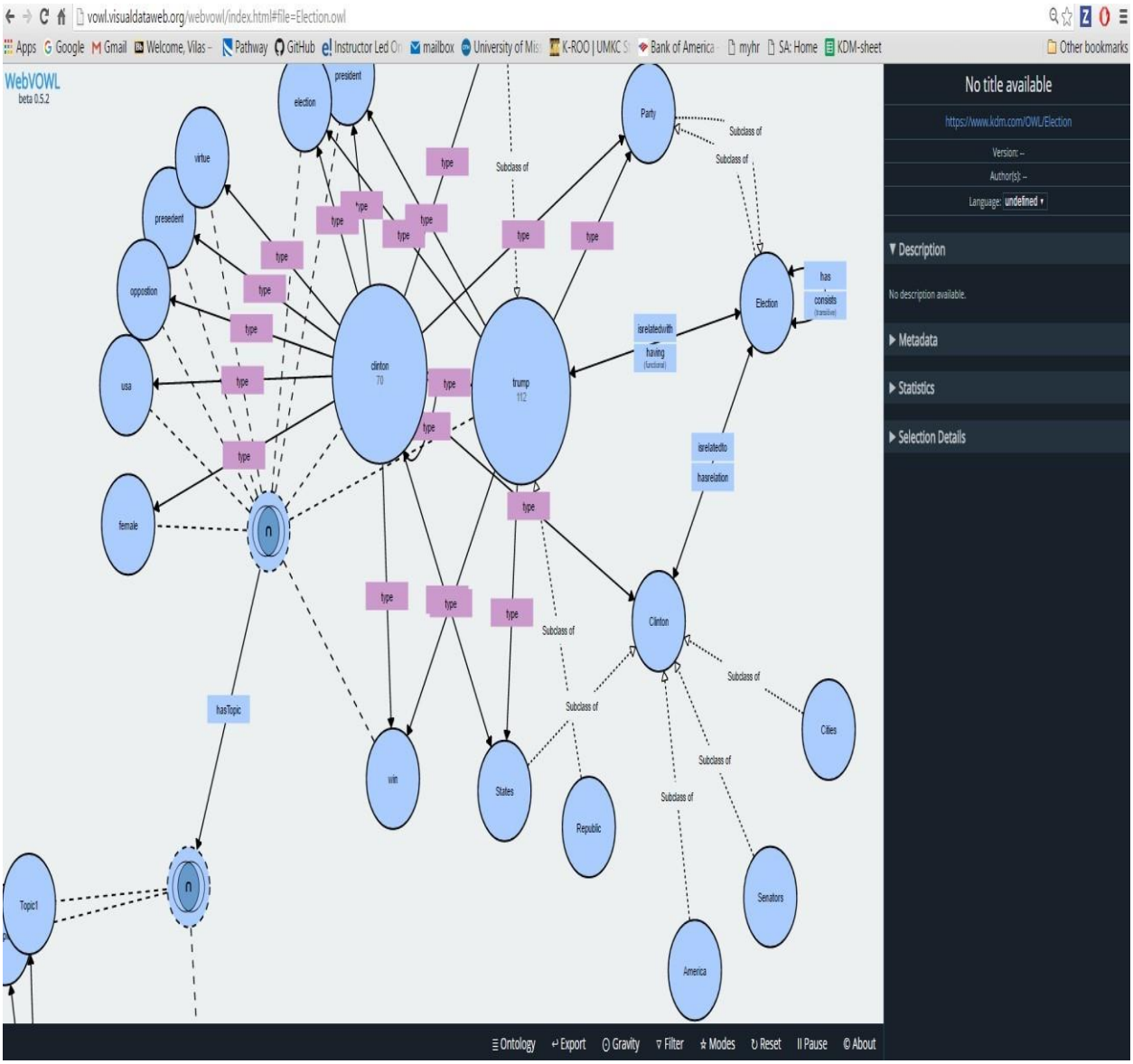


Ontology metrics:

Metrics	
Axiom	36154
Logical axiom count	18276
Declaration axioms count	17877
Class count	31
Object property count	7
Data property count	1
Individual count	17838
DL expressivity	SHIF(D)
Class axioms	
SubClassOf	15
EquivalentClasses	0
DisjointClasses	1
GCI count	0
Hidden GCI Count	0
Object property axioms	
SubObjectPropertyOf	4
EquivalentObjectProperties	0
InverseObjectProperties	3
DisjointObjectProperties	0
FunctionalObjectProperty	1
InverseFunctionalObjectProperty	0
TransitiveObjectProperty	1

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

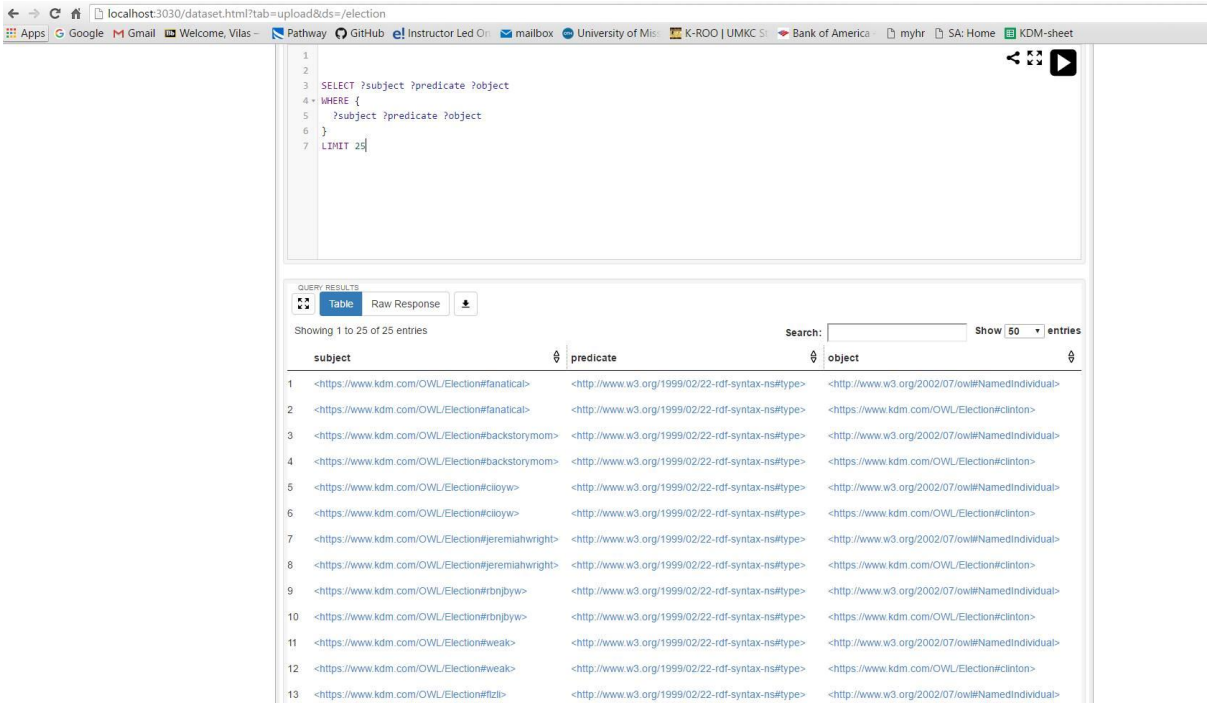
Web View Of our OWL file:



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

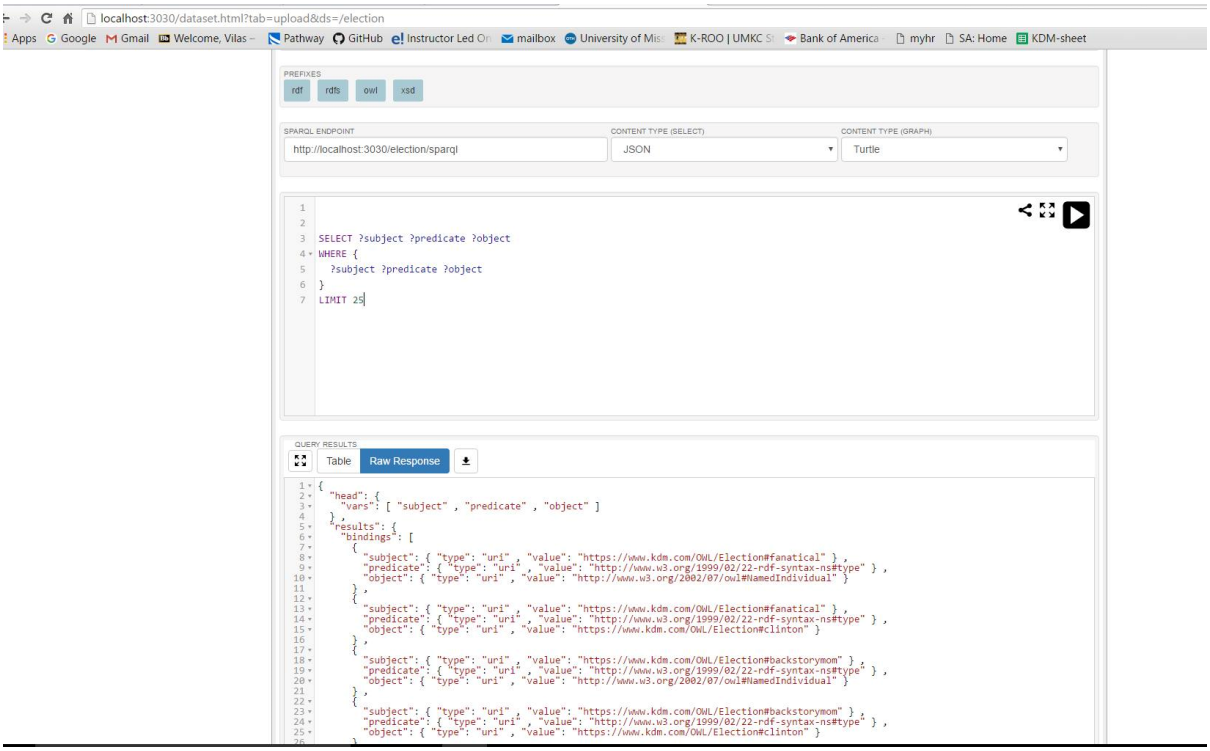
Apache jena fuseki and sparql:

1. Sample Select query and its output in Table Response format



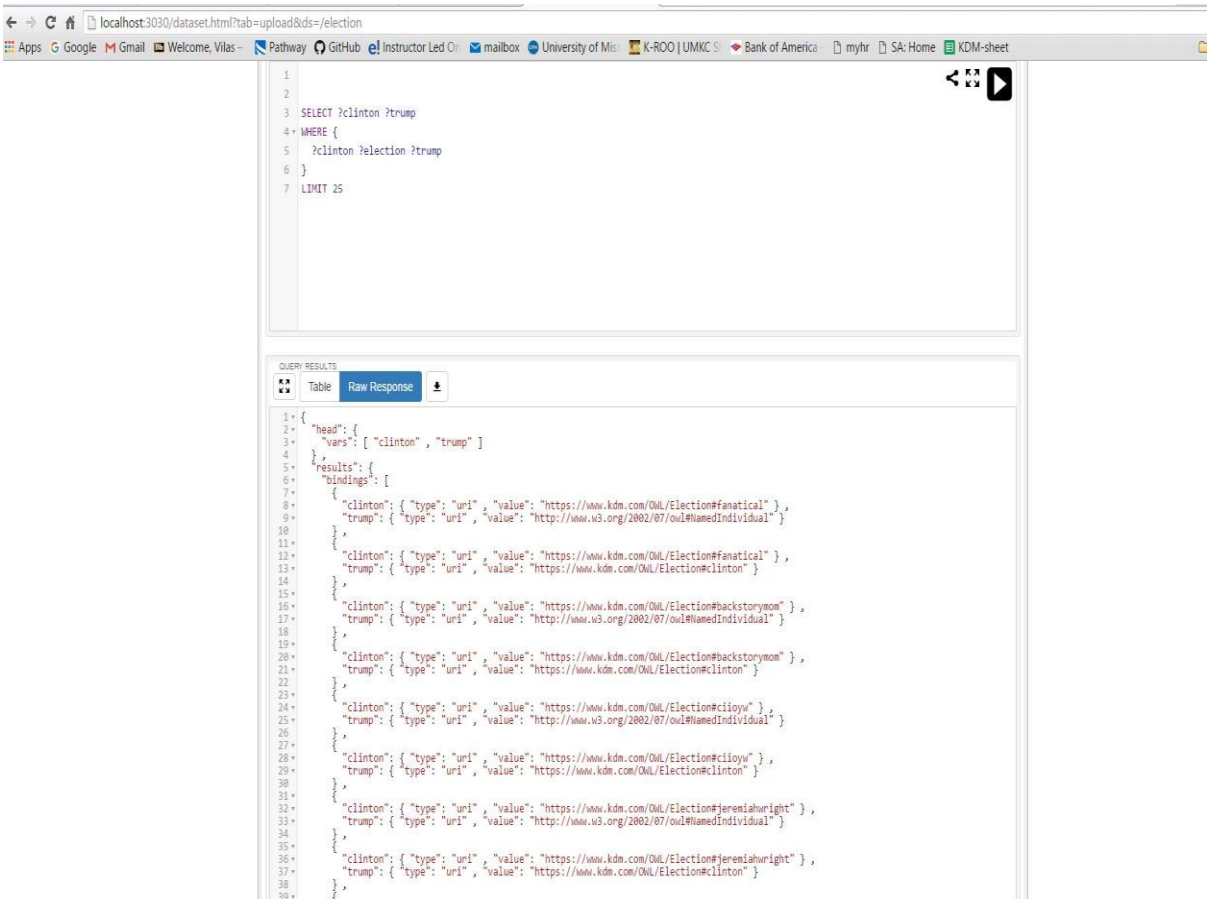
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

2. Sample select query and its output in Raw Response format



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

3. Select query on subjects Clinton, trump and using predicate Election



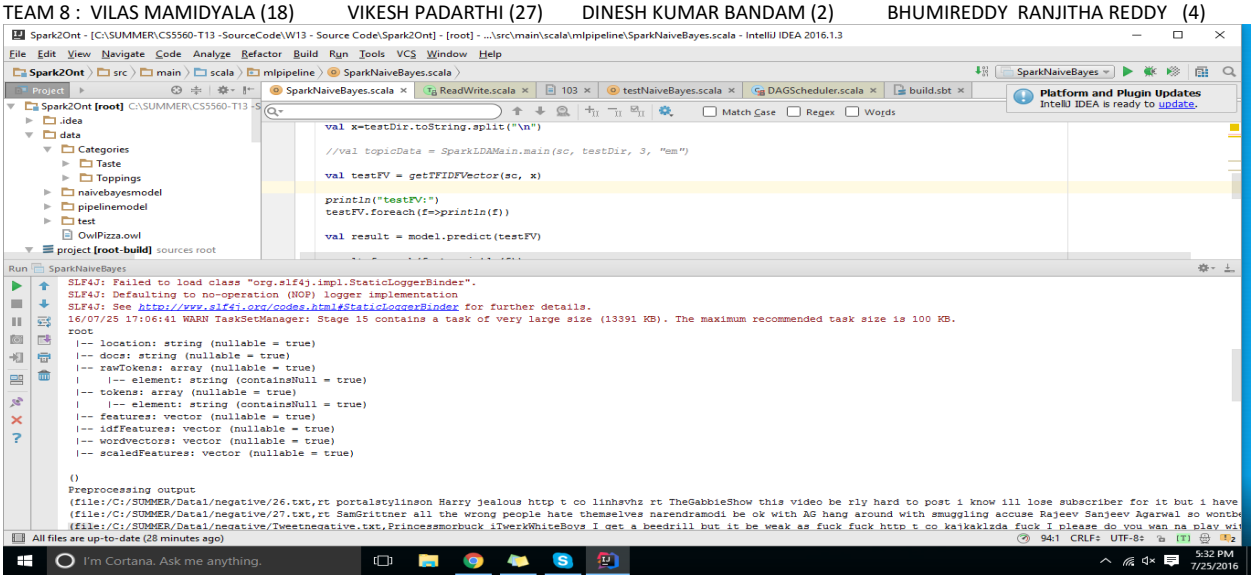
Naïve Bayes Approach for Twitter Sentimental Analysis :

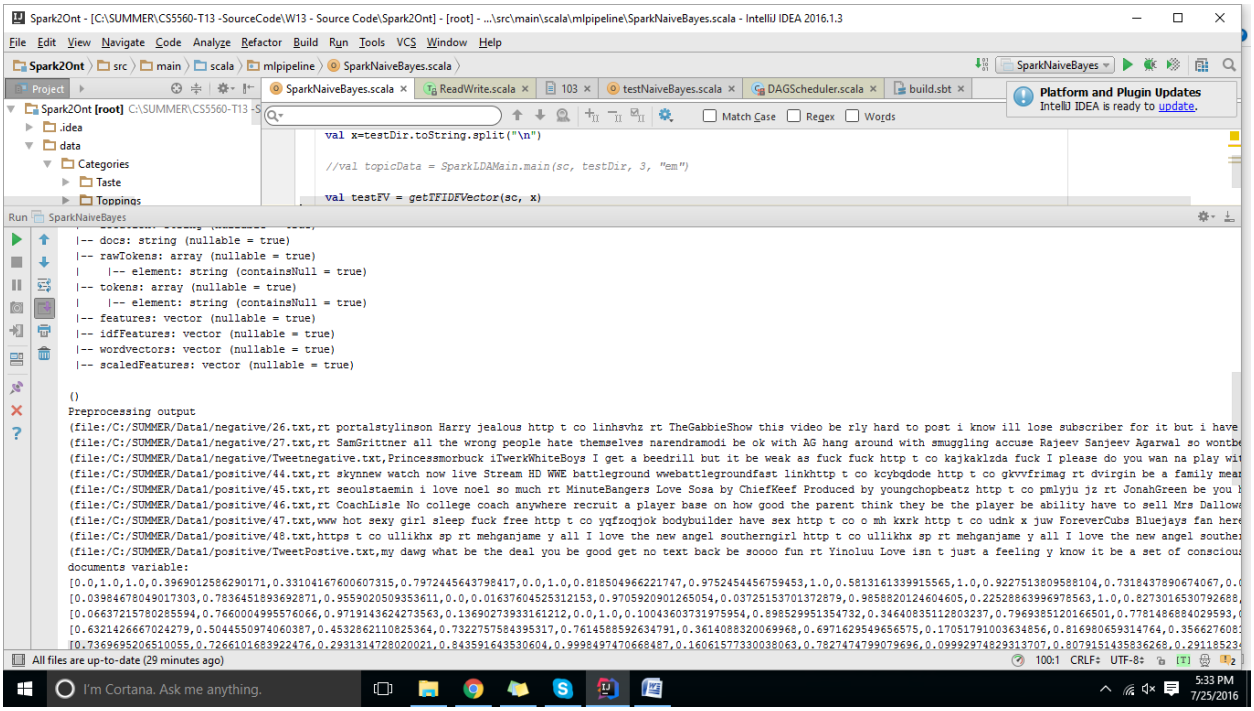
https://github.com/vilasmamidyala/KDM_SM16_SM/tree/master/Source/NaiveBayes

Naive Bayes Model output:

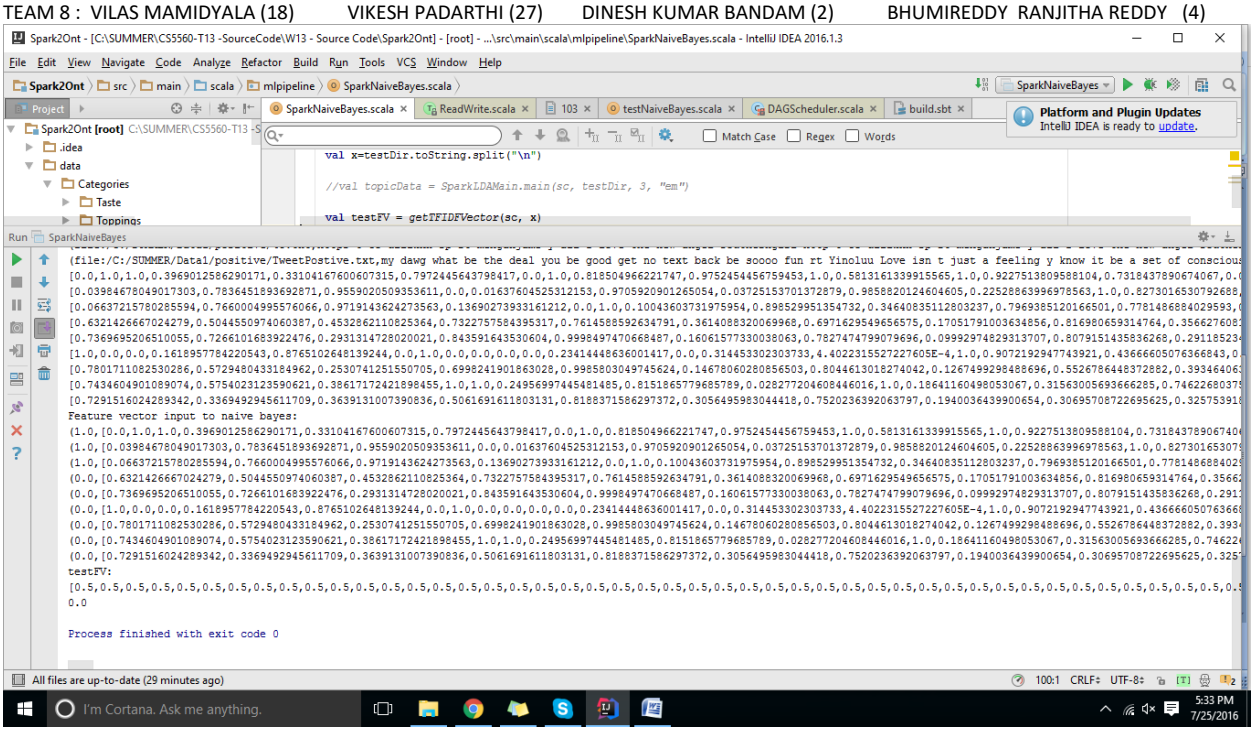
The below images refers to the training data output. We are here first training the model by giving it both positive tweets and negative tweets. Once we execute the program we will get the below output.

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)



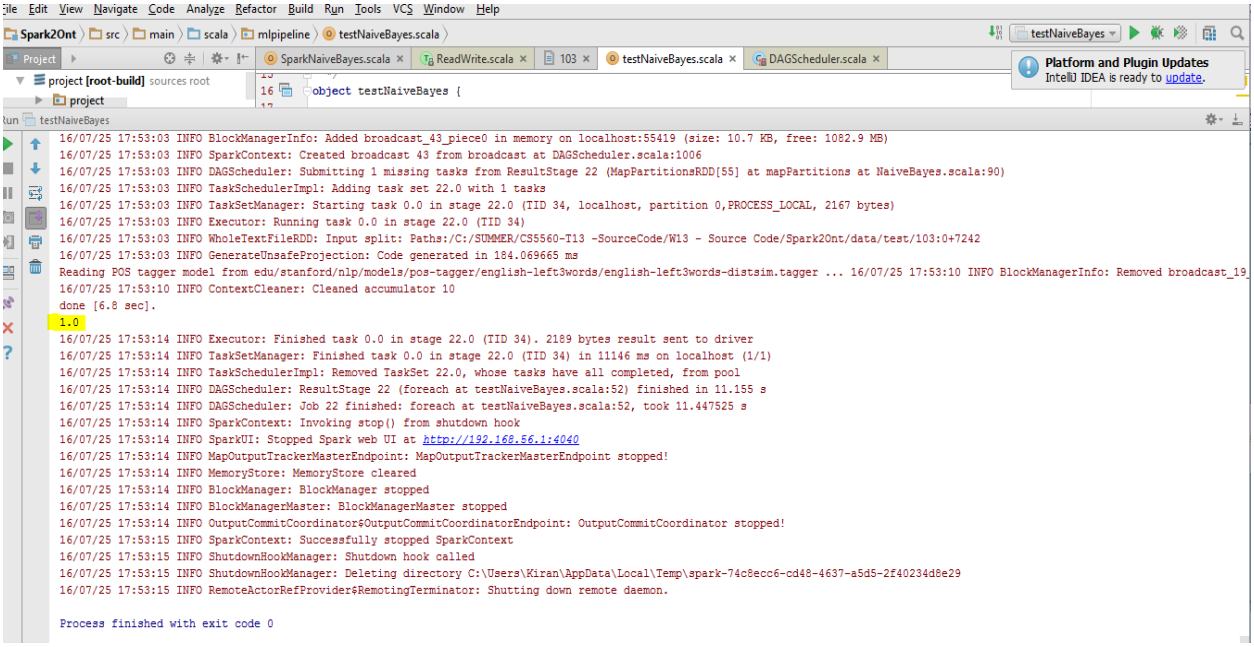
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

After training this model, we have checked the working of this model by giving the positive tweets and negative tweets as input. It worked perfectly. Below image is the example of Negative tweets. Here we are getting 1.0 as output which tells us the file contains negative tweets.

Output:

https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Twitter_Sentiment%20_Analysis.pdf

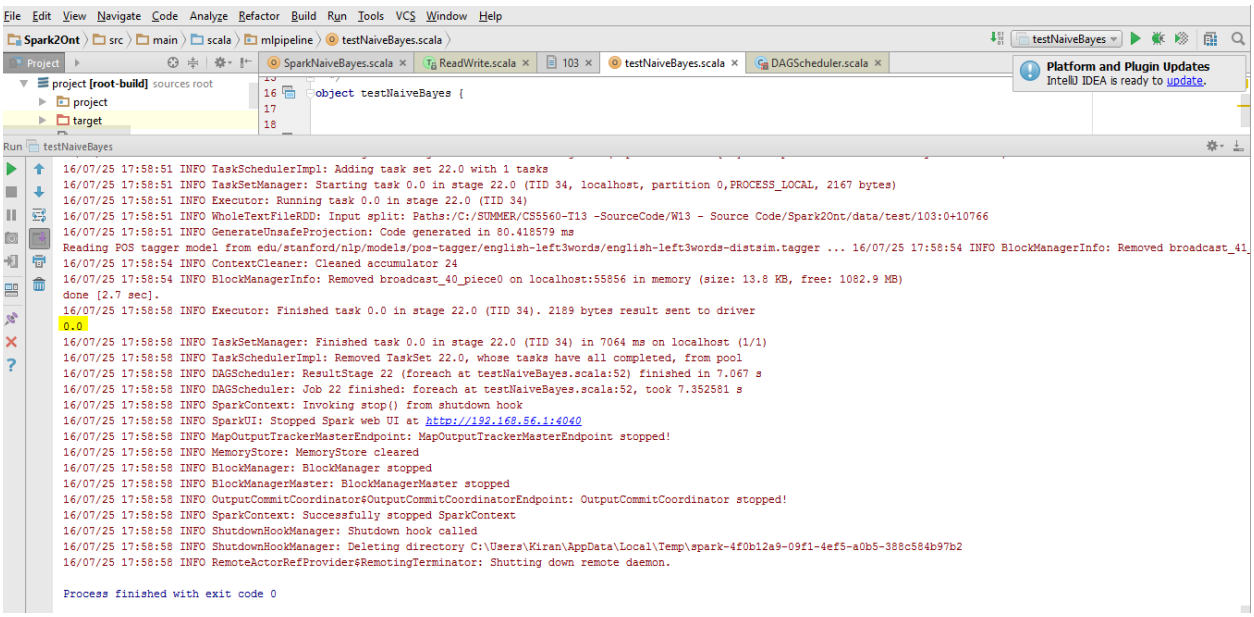
when negative tweets as given input:



we can see 1.0 as output.

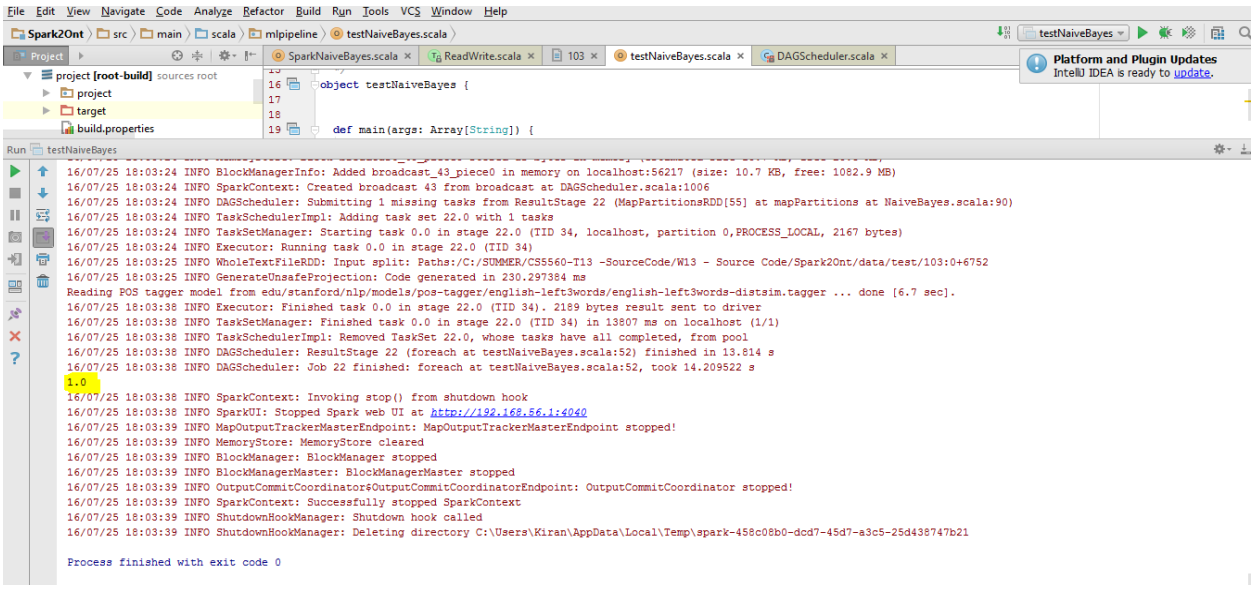
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

when positive tweets are given as input



we can see we got 0.0 as output

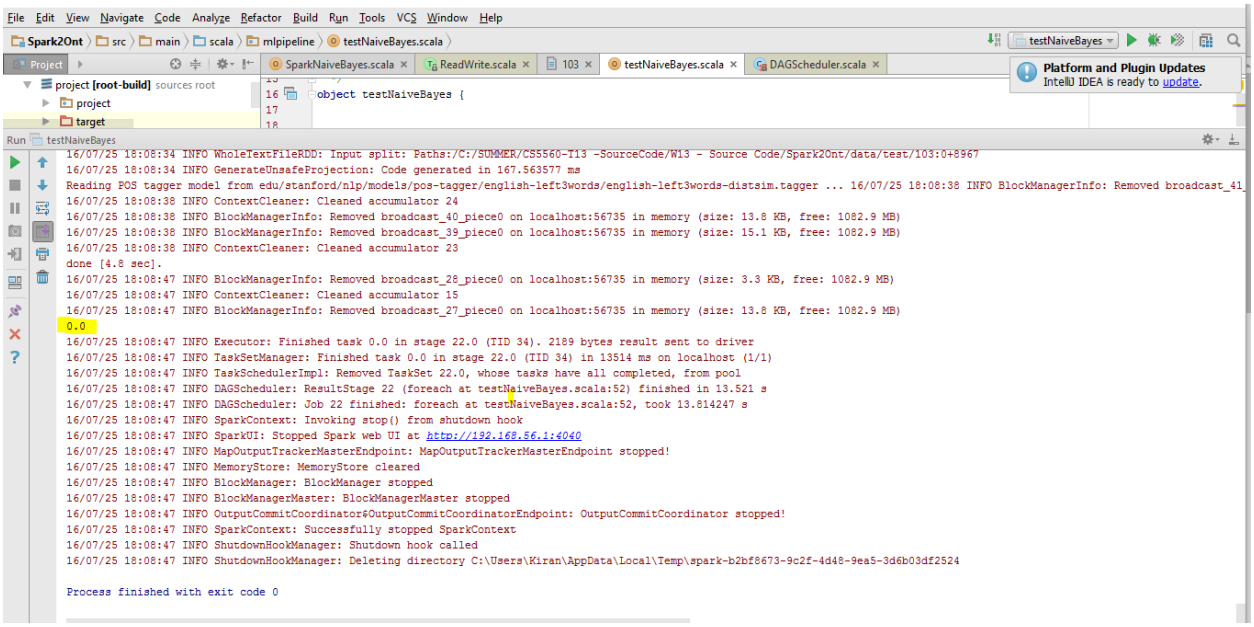
Now we have given 30 positive and 30 negative tweets as input to know how this model works. It has given output as 1.0 which tells us that if the file contains negative data with equal number of positive tweets, it considers as negative.



We are getting 1.0 as output which says us that data is having /having meaning like negative tweets.

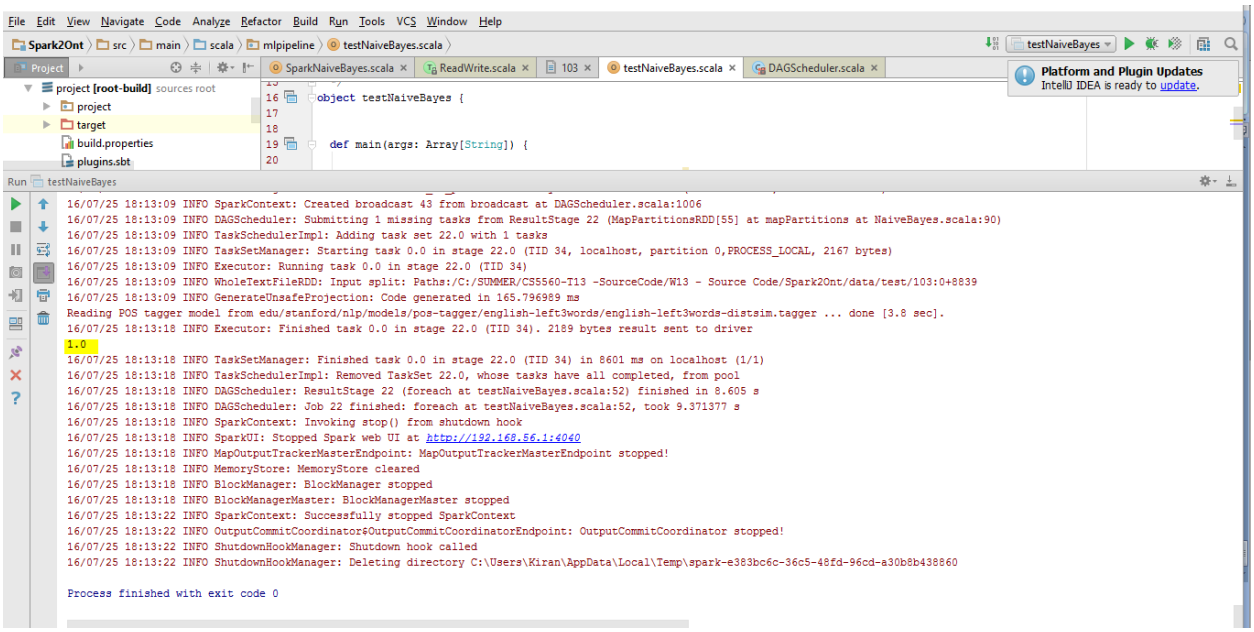
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

when positive 50 negative 30 tweets are given as input



we can see that output is 0.0 which tells us that here positive tweets are more in number and it is domination the negative tweets.

when negative 50 positive 30 is given as input



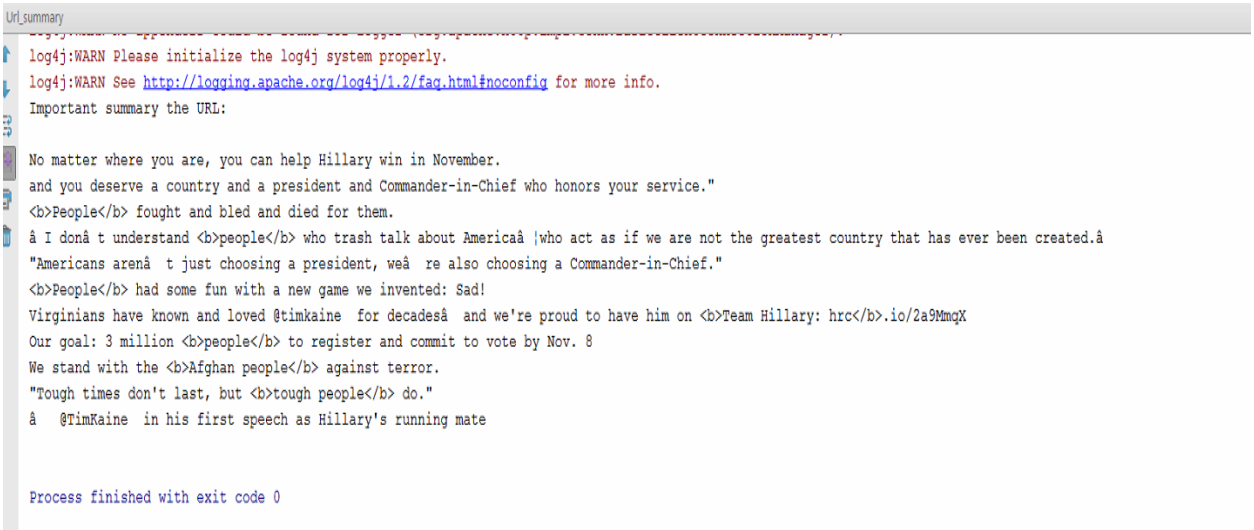
we can see that output is 1.0 which tells us that here positive tweets are more in number and it is domination the negative tweets.

Summarization outputs:

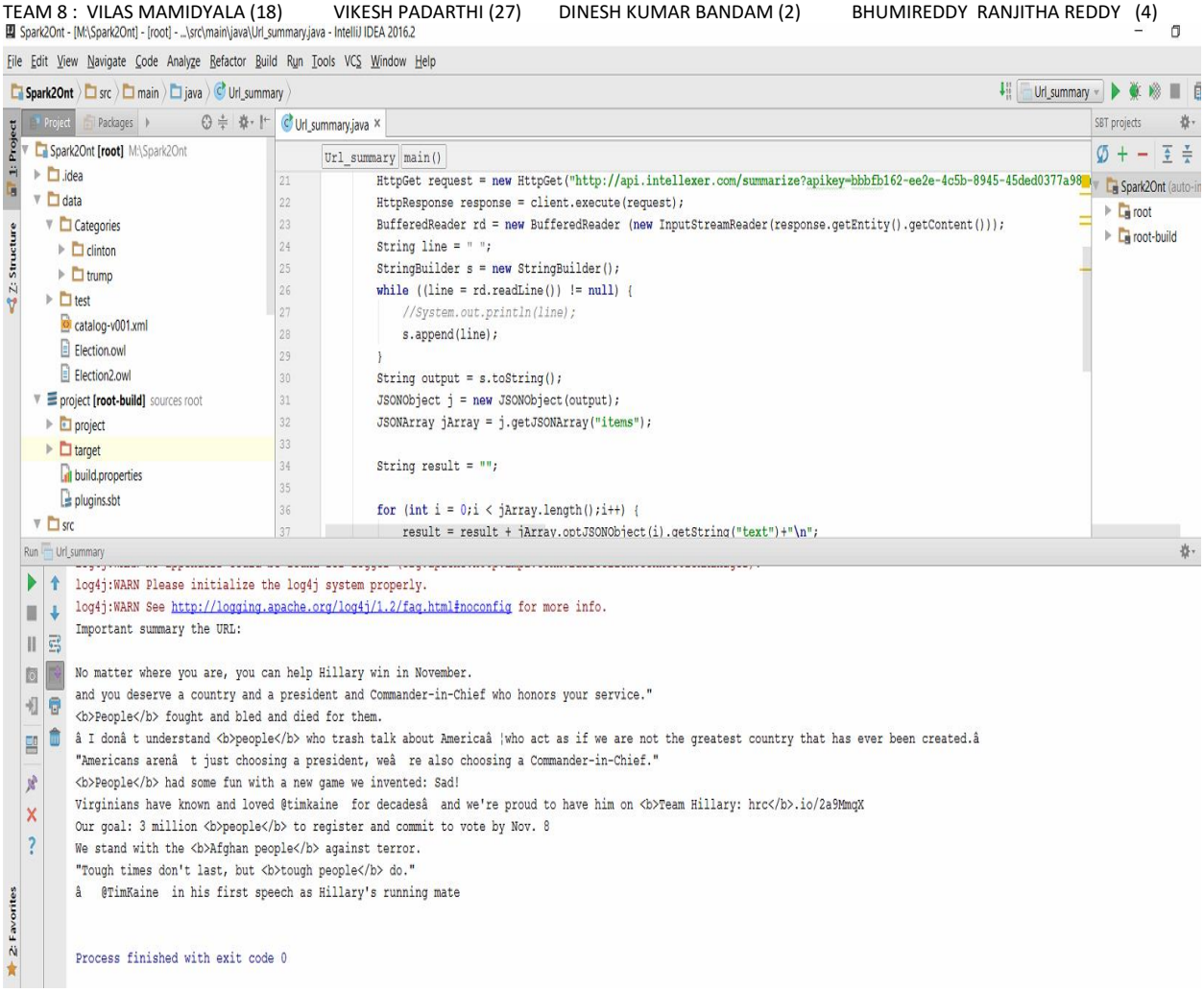
https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Summarization_output.pdf

Source code URL:
https://github.com/vilasmamidyala/KDM_SM16_SM/tree/master/Source/Spark2Ont/src/main/java

Summarization For a given URL:

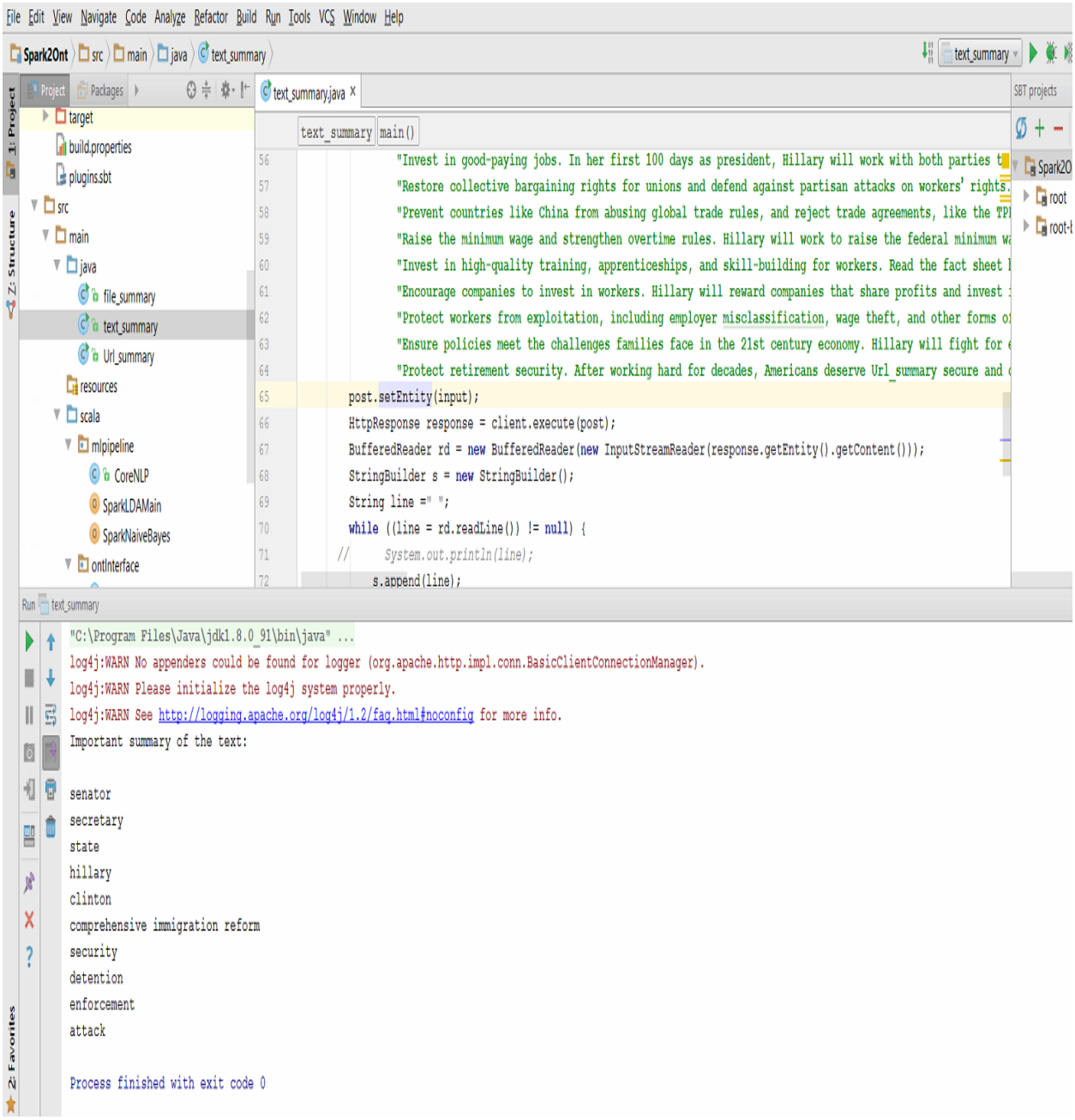


TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)



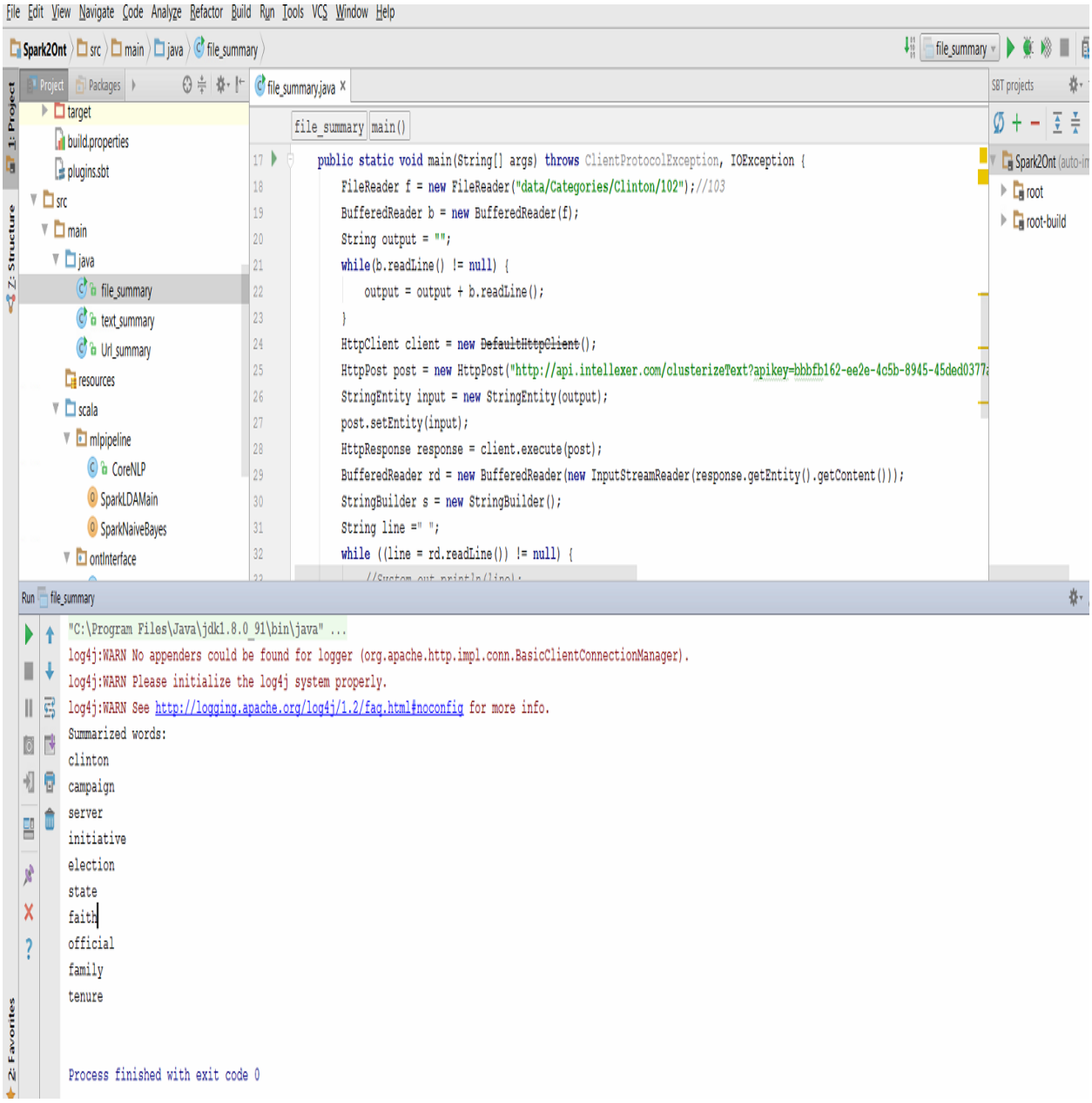
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

Summarization For a given Text:



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

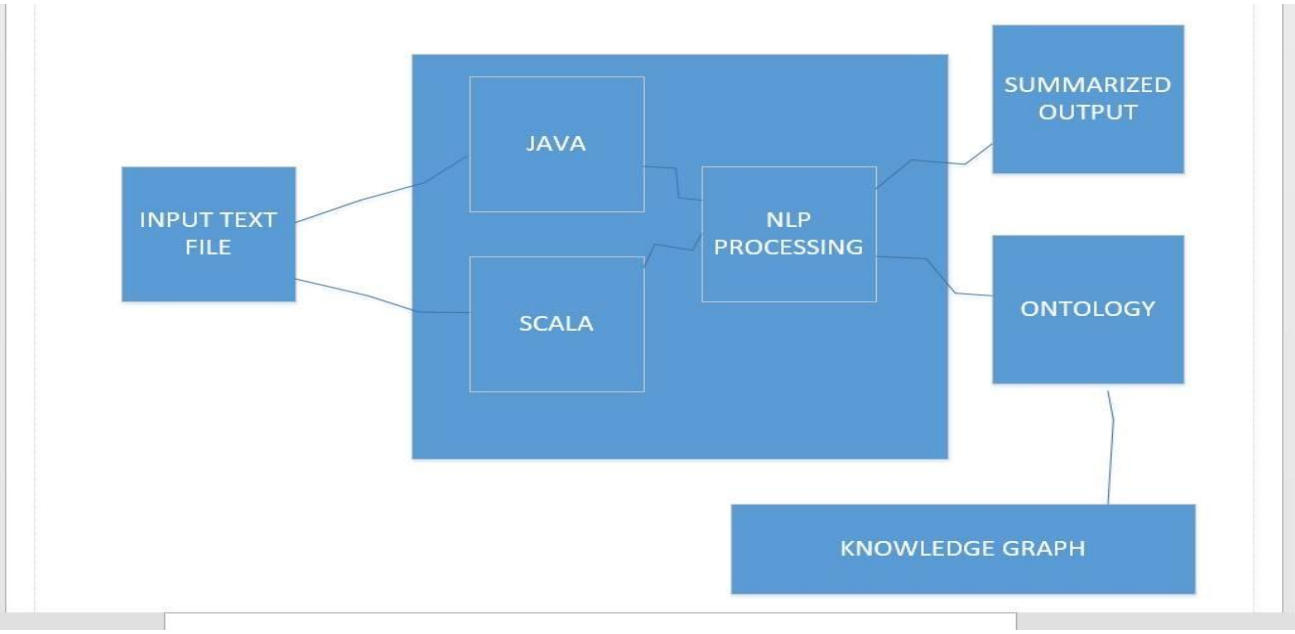
Summarization For a given File:



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

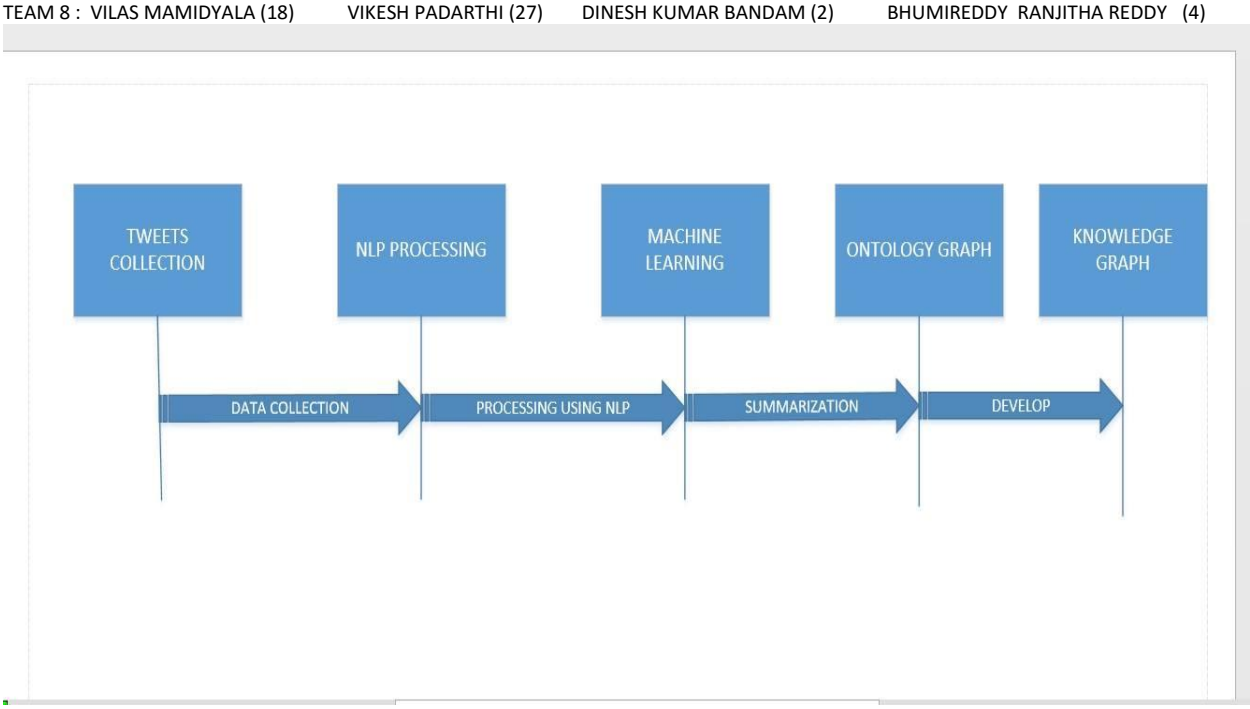
5. Implementation specification:

Software Architecture:

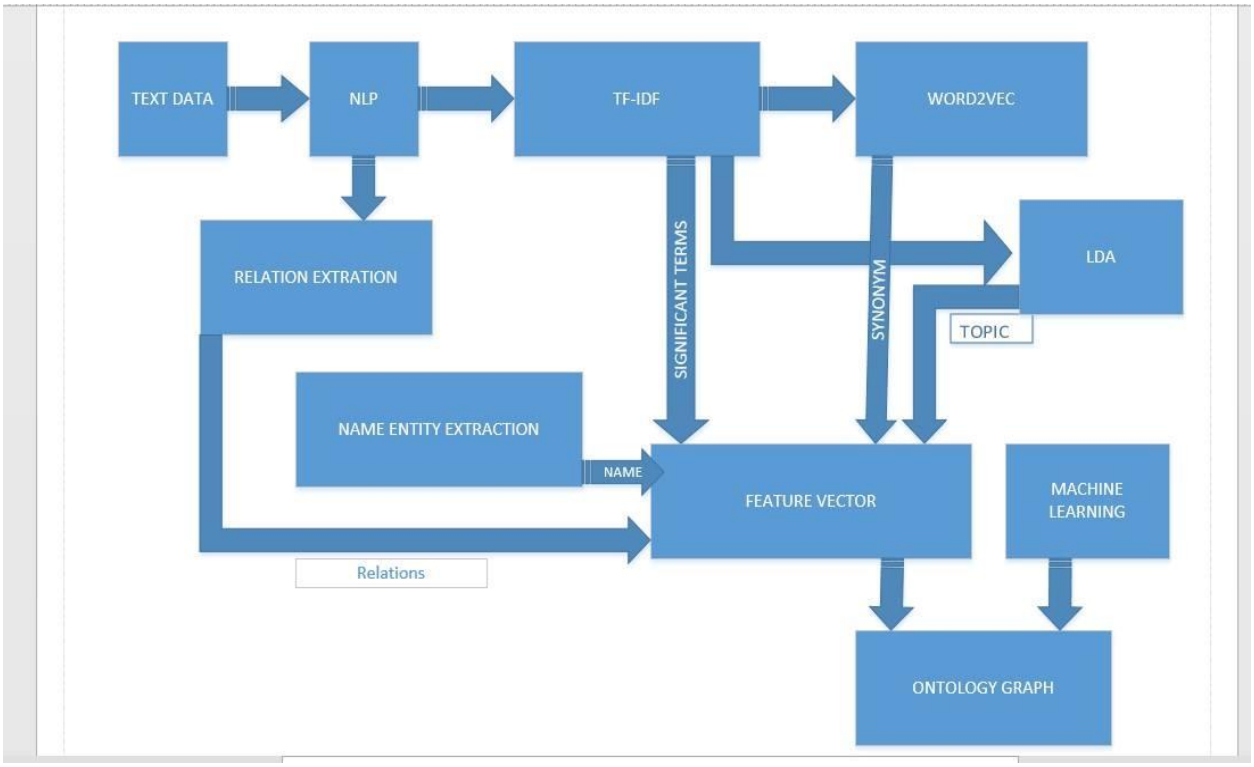


Sequence Diagram:

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



Workflow Diagram:



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

Existing Services Used:

- Implemented word count program using Scala.
- Implemented NLP program.
- Implemented TF-IDF.
- Implemented Word2vec
- Implemented Wordnet
- Implemented NER
- Implemented Feature vector generation
- Generated ontology and used protégé tool for its visualization.
- API services are implemented for summarization

New Services:

Tweet collection using Java Code.

API Services for summarization

6. Results and Evaluation:

- **Accuracy:**

0.08015114814043045,0.07753439247608185,-0.19245687127113342,-0.008618769235908985,-0.002883358858525753,0.13036096096038818,0.07886315882205963,0.04046628996729851,-0.006241005845367908 5,0.08720554411411285,0.00431081373244524,-0.014776067808270454,0.008751815184950829,0.04563858360052109,-5.725307273678482E-4,0.04987286403775215,-0.0967053771018982,0.05071503296494484,0.12420203536748886,-0.06976064294576645,0.04827743023633957,0.10278143733739853,0.036327339708805084,0.055137474074089860916,0.0542316734790802,0.050917837768793106,-0.034305866807699203,-0.05986405536532402,0.026683181524276733,-0.09663253277540207,0.018960680812597275,0.020075950771570206,0.05938071385025978,-0.011799460276961327,-0.06337083876132965,0.010859455913305283,0.10203225165605545,0.045340344309806824,-0.029840435832738876,0.010707934387028217,0.1364029198884964,-0.016686882823705673])

Confusion matrix:

0.0 1.0 2.0 0.0

0.0 2.0 3.0 0.0

0.0 0.0 4.0 0.0

0.0 0.0 1.0 1.0

Accuracy: 0.5

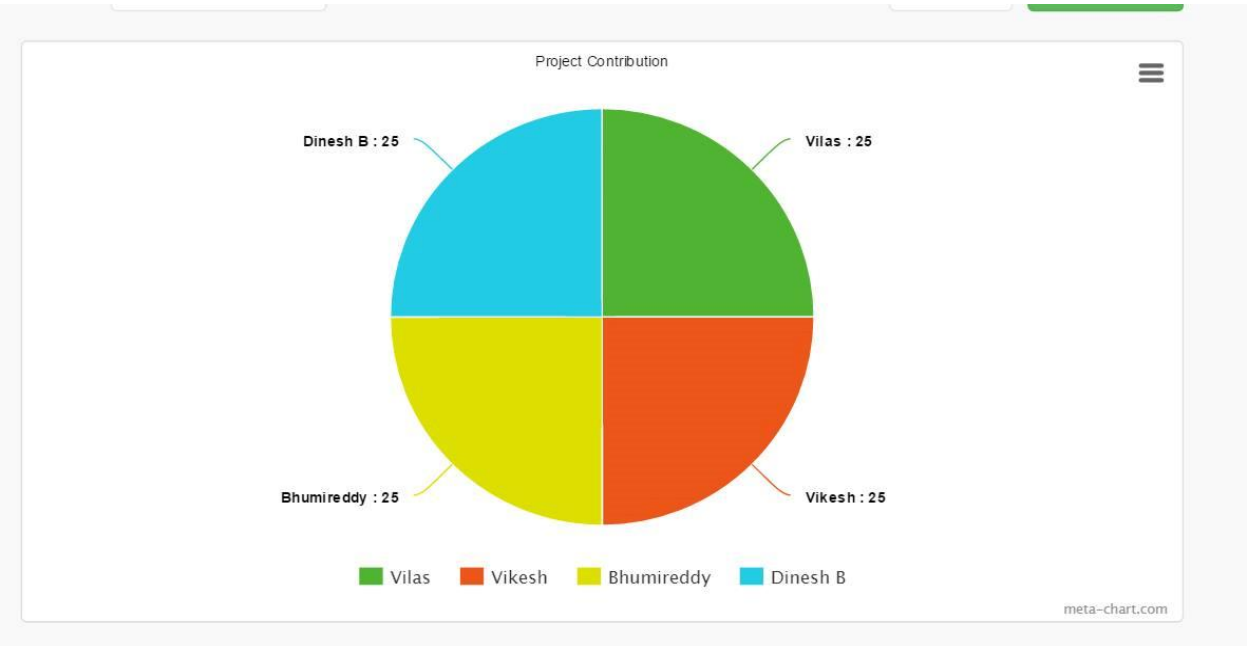
Here when we ran the Naïve bayes algorithm we got the above feature vector and the confusion matrix. We could also observe that it has given the accuracy of 0.5 as well.

When we performed clustering for the tweets with positive and negative we had achieved 100% accuracy for the classification. This is a good achievement with respect to clustering.

- **Run Time Performance:**
  - ✓ Ontology Run Time: 15 min
  - ✓ Summarization:
    - ➢ For a given file: 1minute
    - ➢ For a given URL: 1minute
    - ➢ For a given text: 1minute
  - ✓ Clustering of tweets with respect to positive and negative words using Naïve Bayes Algorithm: 7 min for training and 3 min for testing.
  - ✓ OWL file opening through Protégé: 3 min for viewing Onto graph.
  - ✓ SparQL query execution: 1minute.

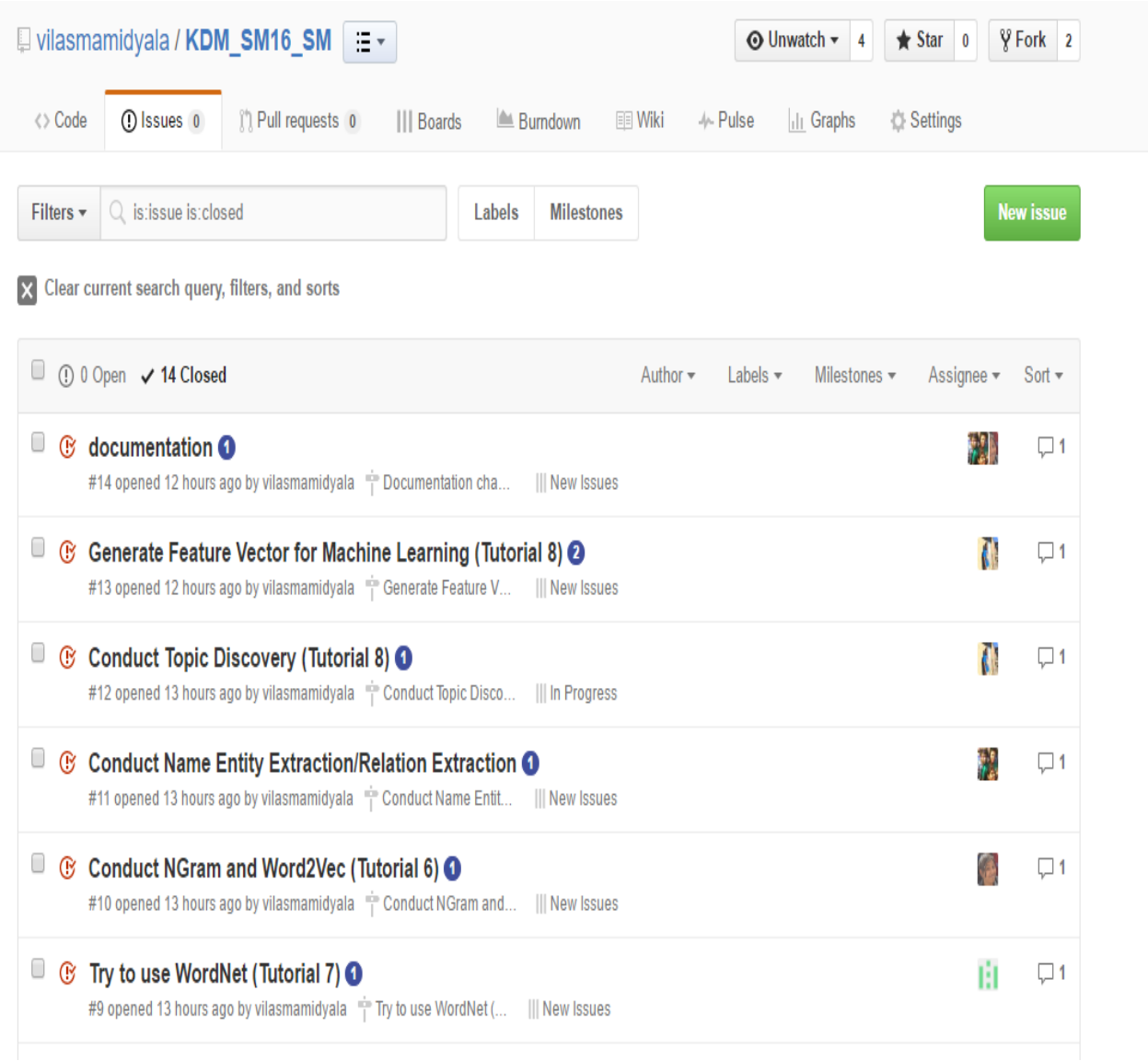
7. Project Management:

Contribution of Each member:



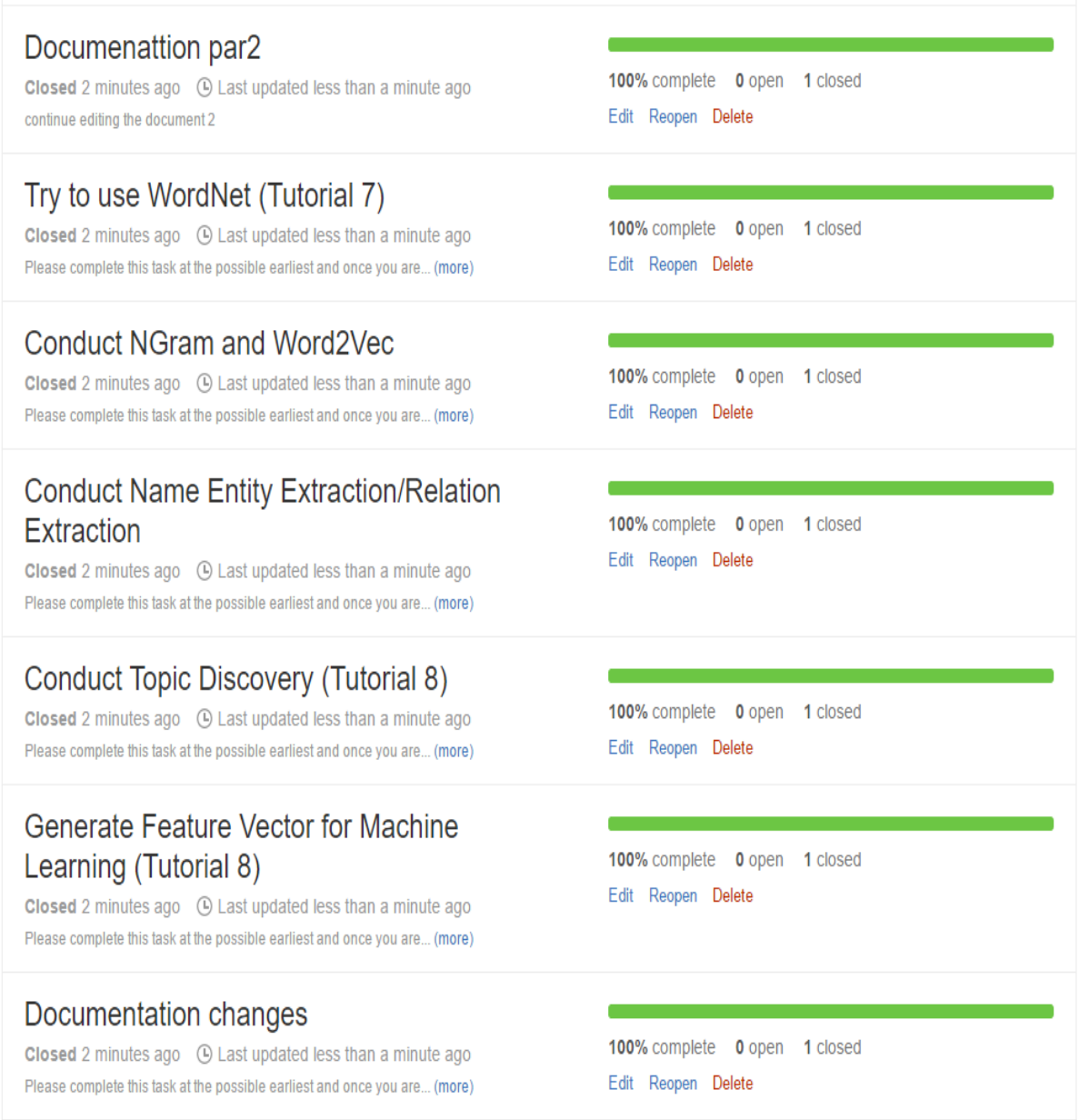
TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

Zenhub and Github Screen shots:

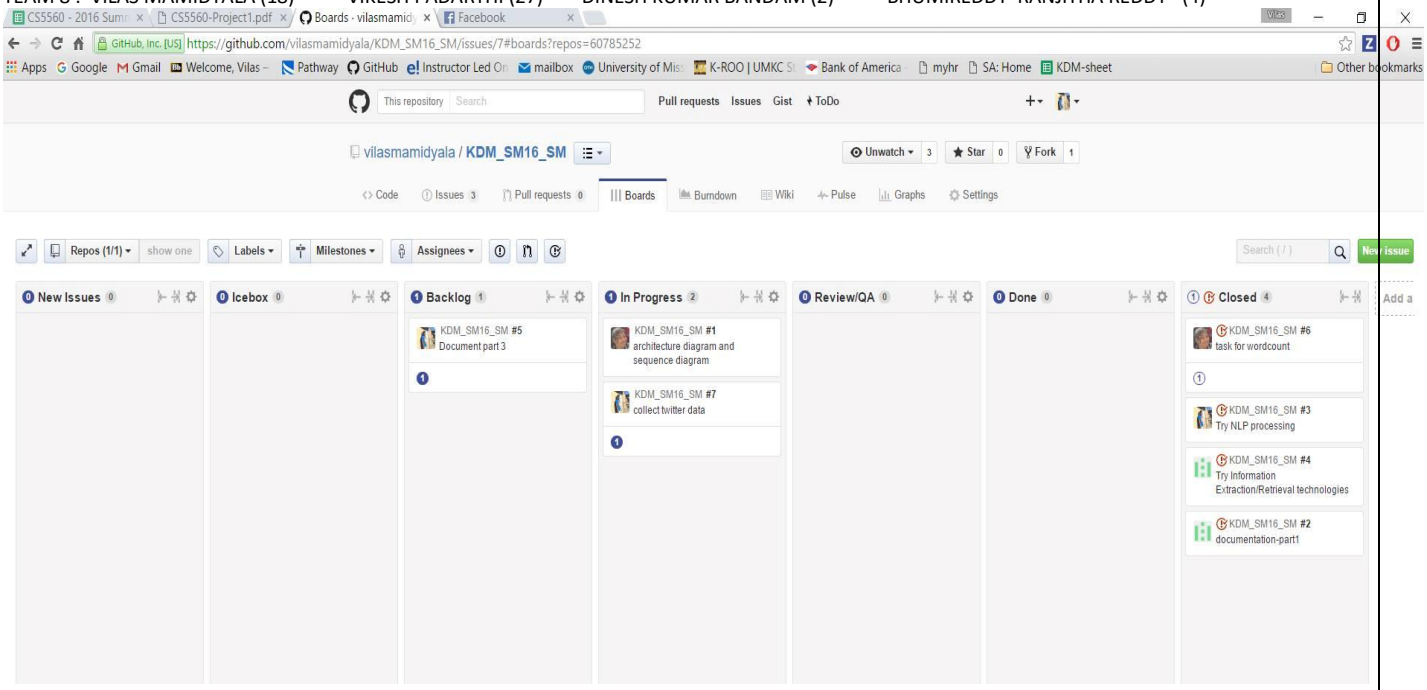


TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

Milestones:

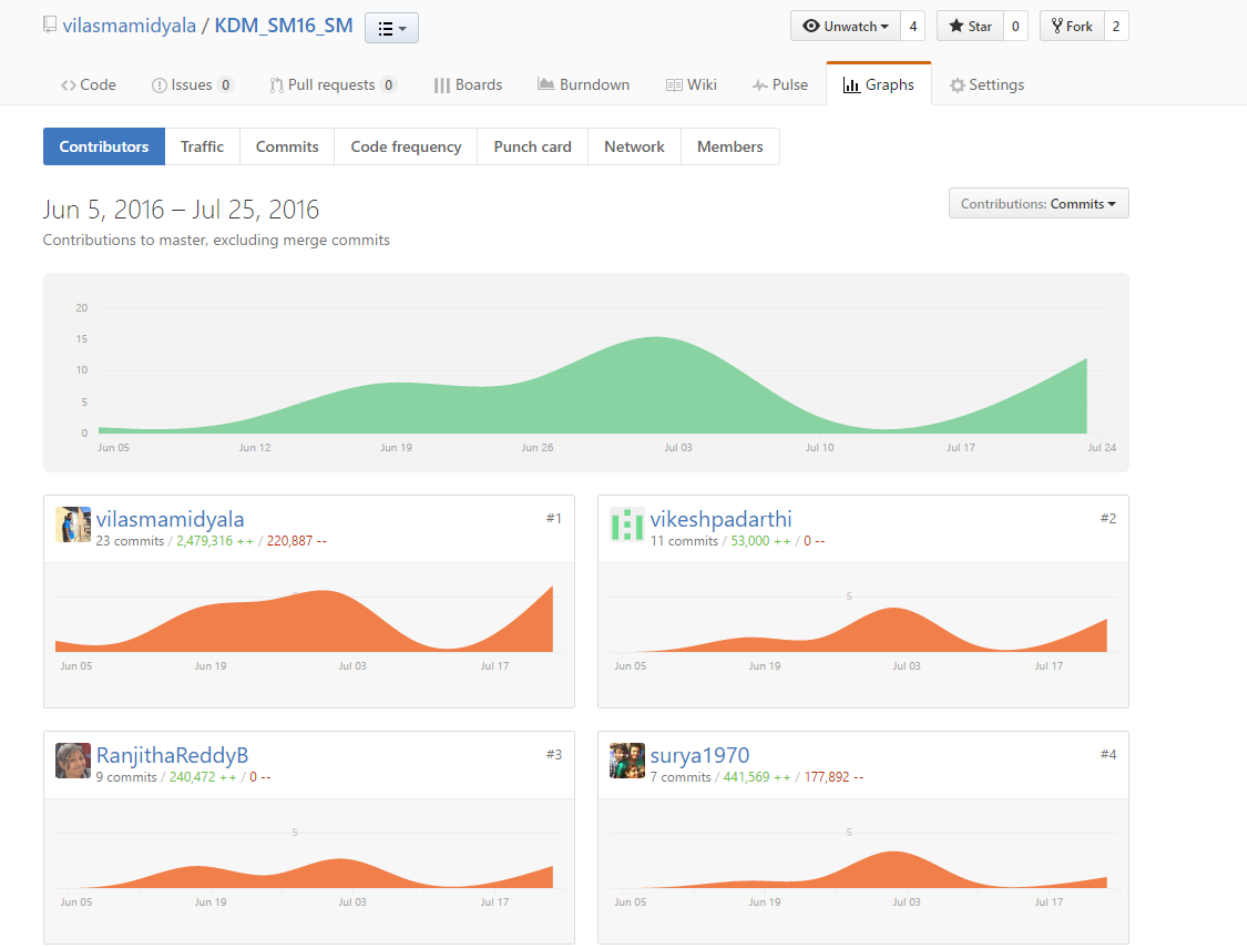


TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)



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

Contribution of source code GitHub:



8. 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.

References:

- SparQL: https://www.w3.org/TR/rdf-sparql-query/
- Ontology: http://homes.cs.washington.edu/~pedrod/papers/hois.pdf
- Protégé: http://protege.stanford.edu/
- NaiveBayes Algorithm: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf
- Summarization: https://www.semanticscholar.org/paper/An-Ontology-Based-Approach-to-Text-Summarization-Hennig-Umbrath/ab138bc53af41bfc5f1a5b2ce5ab4f11973e50aa
- http://www.intellexer.com/