

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

SECOND 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

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)

```

Run SparkNER
(EI_Universal_Mx : ,O O )
(#entérate ,O )
(Lynch ,PERSON )
(y ,O )
(Comey ,PERSON )
(comparecerán ,O )
(ante ,O )
(Congreso ,PERSON )
(por ,O )
(caso ,O )
(HillaryClinton ,PERSON )
(https://t.co/wjhlpatd6 ,O )
( ,POLL)
(rt ,O )
(@Reince : ,O O )
( . @hillaryclinton ,O O )
(spend ,O )
(the ,O )
(last ,O )
(16 ,NUMBER )
(month ,DURATION )
(deliberately ,O )
(lie ,O )
(to ,O )
(the ,O )
(American ,MISC )
(people . ,O O )

```

```

SparkNER
(spend ,O )
(the ,O )
(last ,O )
(16 ,NUMBER )
(month ,DURATION )
(deliberately ,O )
(lie ,O )
(to ,O )
(the ,O )
(American ,MISC )
(people . ,O O )
(watch : ,O O )
(https://t.co/18kiu6b62a ,O )
( ,POLL)
16/07/08 21:58:36 INFO SparkContext: Invoking stop() from shutdown hook
16/07/08 21:58:36 INFO SparkUI: Stopped Spark web UI at http://192.168.0.20:4040
16/07/08 21:58:36 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/07/08 21:58:36 INFO MemoryStore: MemoryStore cleared
16/07/08 21:58:36 INFO BlockManager: BlockManager stopped
16/07/08 21:58:36 INFO BlockManagerMaster: BlockManagerMaster stopped
16/07/08 21:58:36 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
16/07/08 21:58:36 INFO SparkContext: Successfully stopped SparkContext
16/07/08 21:58:36 INFO ShutdownHookManager: Shutdown hook called
16/07/08 21:58:36 INFO ShutdownHookManager: Deleting directory C:\Users\vilas\AppData\Local\Temp\spark-d3b84157-3504-4f3e-ac9c-54f198249045

Process finished with exit code 0

```

- 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)

```

String[] poss = wordnet.getPos(word);
for (int j = 0; j < poss.length; j++) {
    System.out.println("\n\nSynonyms for " + word + " (pos: " + poss[j] + ")");
    String[] synonyms = wordnet.getAllSynonyms(word, poss[j], 10);
    for (int i = 0; i < synonyms.length; i++) {
        System.out.println(synonyms[i]);
    }
}
    
```

Run WordNetMain

```

C:\Program Files\Java\jdk1.8.0_91\bin\java" ...
Finding parts of speech for win.
v
n
Finding parts of speech for loss.
v
n
Definitions for loss:
something that is lost
gradual decline in amount or activity
the act of losing someone or something
the disadvantage that results from losing something
the experience of losing a loved one
the amount by which the cost of a business exceeds its revenue
military personnel lost by death or capture
euphemistic expressions for death
Synonyms for win (pos: v)
accept
accomplish
achieve
    
```

```

Corpus summary:
Training set size: 10704 documents
Vocabulary size: 11485 terms
Training set size: 60775 tokens
Preprocessing time: 36.617139417 sec

Finished training LDA model. Summary:
Training time: 231.321711026 sec
Training data average log likelihood: -51.39299144407459

4 topics:
TOPIC_0::;0.06898491214163481
TOPIC_0;rt;0.05230523781160823
TOPIC_0;.;0.0443182452233433
    
```

Run SparkLDAMain

```

accessory 4.330187058220508E-7
engine 4.3300706399910715E-7
founder 4.3299449089844737E-7
steakhouse 4.329928297333121E-7
brokerage 4.32980000717234E-7
coffee 4.3296174743087827E-7
menswear 4.3295446272151215E-7
drinks 4.329446783631709E-7
leasing 4.329246106191E-7
external 4.329217171459645E-7
144 4.32877349955435E-7
annex 4.328732236055471E-7
books 4.3286129582288156E-7
chocolate 4.3285034999841574E-7
catering 4.3283666341396286E-7
145 4.3282839027119965E-7
parlor 4.327825690362709E-7
presentation 4.3278136670538397E-7
450,000 4.3275898474339983E-7
    
```

Process finished with exit code 0

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

```

val model = word2Vec.fit(ngramDataFrame)

val output_df = model.getVectors.rdd
  .map { case Row(word: String, vector: Vector) => (word, vector) }

val featureVector = output_df.join(outputRDD).map(f => {
  new LabeledPoint(f._2._2, f._2._1)
})
  
```

Run SparkFVMain

```

(16/07/08 12:37:19 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
16/07/08 12:37:19 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
(3.0,[0.06986328214406967,0.08610797673463821,-0.06651458144187927,0.0050154319033026695,0.048778362572193146,-0.01627318747341633,0.054114941507577896,-0.026237327605485916,0.
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

Process finished with exit code 0
  
```

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_out.put.pdf

The screenshot shows an IDE with a project named 'Spark_KMeans_FV'. The code in the main editor is as follows:

```
// Read the file into RDD[String]
val input = sc.wholeTextFiles("data/20_twitter_wiki/**", 500).map(line => {
  val location_array = line._1.split("/")
  val class_name = location_array(location_array.length - 2)
  var ff = line._2.replaceAll("[^a-zA-Z\\s:]", " ")
  ff = ff.replaceAll(":", " ")
  //Getting Lemmatized Form of the word using CoreNLP
  val lemma = CoreNLP.returnLemma(ff)
  (mapping.value.get(class_name).get.toDouble, lemma)
})
```

The output window shows the following logs:

```
16/07/08 12:42:25 INFO BlockManagerMaster: Trying to register BlockManager
16/07/08 12:42:25 INFO BlockManagerMasterEndpoint: Registering block manager localhost:51092 with 2.4 GB RAM, BlockManagerId(driver, localhost, 51092)
16/07/08 12:42:25 INFO BlockManagerMaster: Registered BlockManager
16/07/08 12:42:27 WARN : Your hostname, DESKTOP-0CQAA2F resolves to a loopback/non-reachable address: fe80:0:0:0:9dbe:f363:315a:a773%wlan1, but we couldn't find any external IP
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done [0.6 sec].
[0.0,obama schedule he campaign appearance with hillaryclinton before comey speak action speak louder than word he kne BloodyDifficultWoman HillaryClinton http t co hih srn zz
[0.0,e don t know but Russia probably do Clinton rt gamma ray if this doesn t demand realdonaldtrump vote NOTHING will traitorousfbi realdonaldtrump do you answer the tweet or
[0.0,we don t know but Russia probably do Clinton rt creativene jrjohm pastormarkburn HillaryClinton FBI she walk free for ignorance to where we would be in prison for yr c rt
root
|-- Labels: double (nullable = false)
|-- sentence: string (nullable = true)
|-- words: array (nullable = true)
|   |-- element: string (containsNull = true)
|-- filteredWords: array (nullable = true)
|   |-- element: string (containsNull = true)
|-- ngrams: array (nullable = true)
|   |-- element: string (containsNull = false)

()
Confusion matrix:
2.0  0.0  1.0  0.0 |
4.0  0.0  0.0  1.0
0.0  1.0  1.0  1.0
1.0  1.0  0.0  1.0
Accuracy: 0.2857142857142857

Process finished with exit code 0
```

■ 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

```
val spark = SparkSession
    .builder
    .appName("SparkW2VML")
    .getOrCreate()

try persuade,0.9999271295394773]
[stevecasull @neilturner_,0.9999223192014983]

* * :
[tired even,0.9999997252937245]
[! https://t.co/hamkx17m,0.999993512123526]
[today leftist,0.9998832436736813]

announce chris,0.9999928400911934]
[golf buddy,0.9999879088650858]
[still voting,0.9999849287155952]

clinton buy :
[#neara 16,0.9999991268815702]
[rt @occinnoc,0.9999945532953244]
[realdonaldtrump https://t.co/f6qc7votnc,0.9999847580361189]

fuck fuck :
[8nypost @speakeerryan,0.9999545161843504]
[phone eat,0.999940231394208]
[even #Zbi,0.9999075607160088]

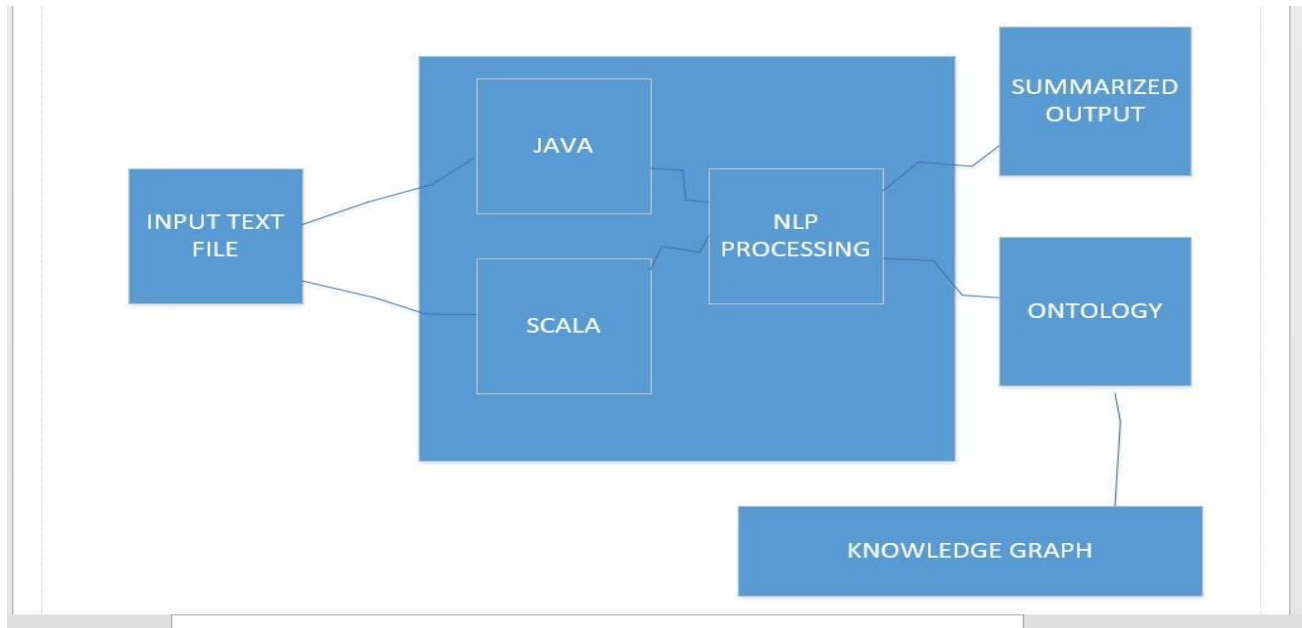
buy fbi :
[#corrupt @obamaclintoninc,0.9999960428397687]
[event,,@realdonaldtrump @tuckerleon,0.9999948101587859]
[nice ribbon,0.9999915503316831]

Process finished with exit code 0
```

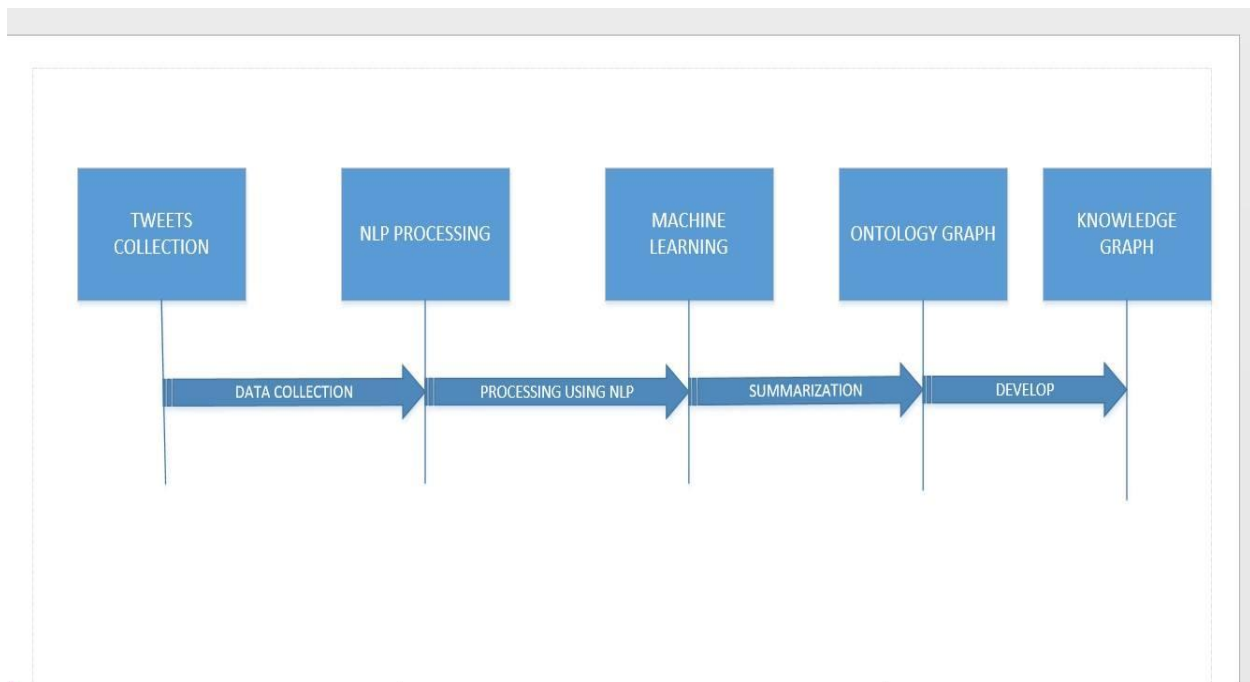

TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

5. Implementation specification:

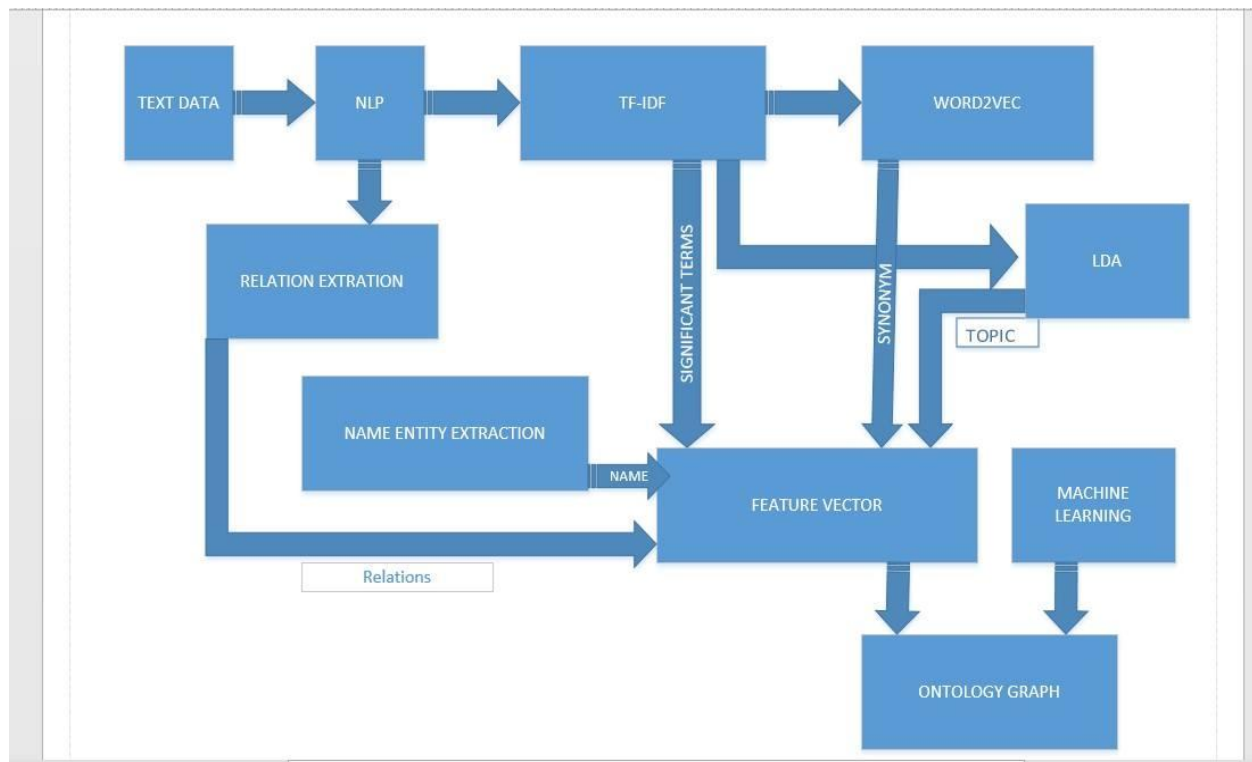
Software Architecture :



SEQUENCE DIAGRAM :



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



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

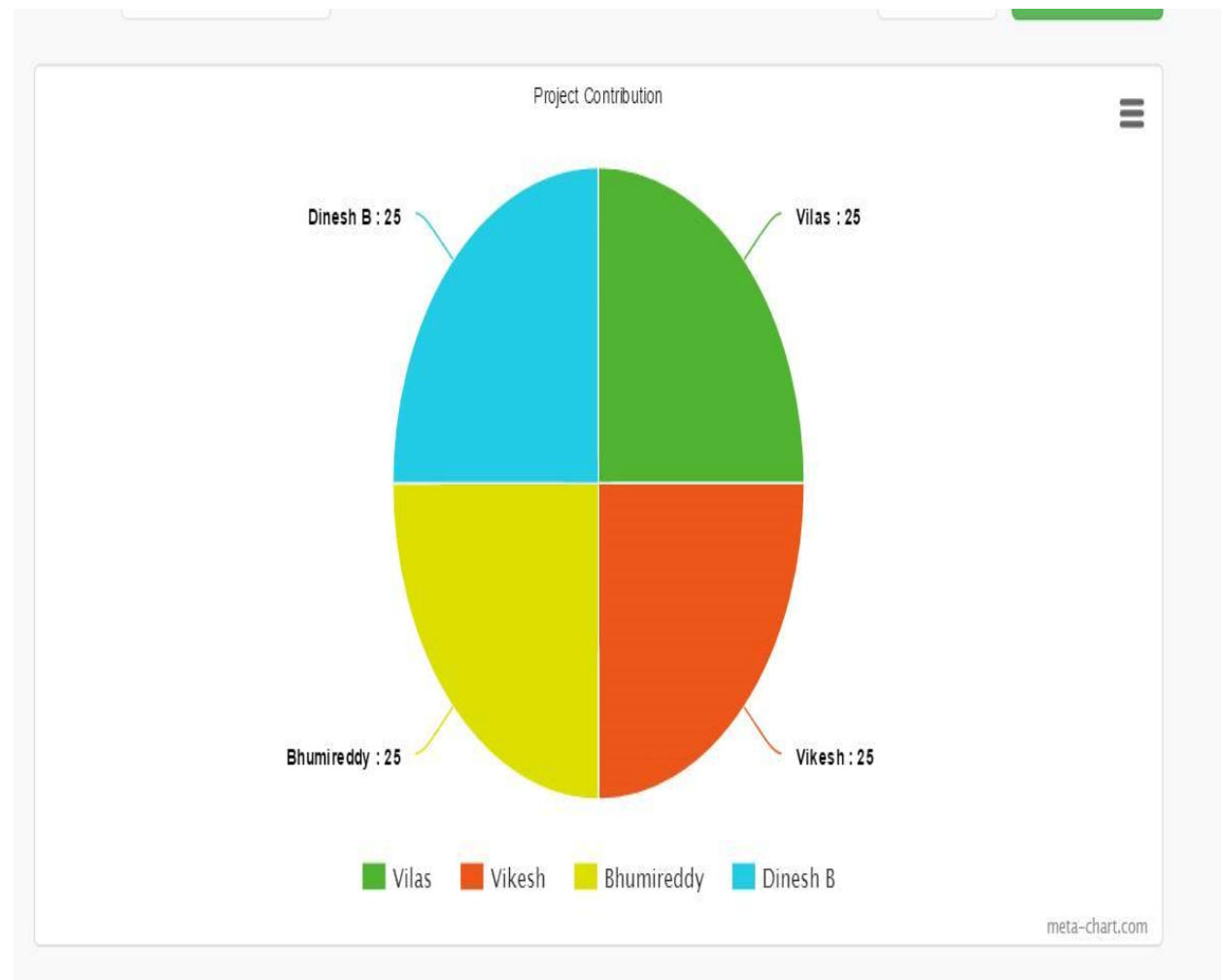
New Services:

Tweet collection using Java Code.

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

6. 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:

vilasmamidyala / KDM_SM16_SM

Unwatch 4 Star 0 Fork 2

Code Issues 0 Pull requests 0 Boards Burndown Wiki Pulse Graphs Settings

Filters is:issue is:closed Labels Milestones New issue

Clear current search query, filters, and sorts

	0 Open	✓ 14 Closed	Author	Labels	Milestones	Assignee	Sort
documentation 1							1
#14 opened 12 hours ago by vilasmamidyala Documentation cha... New Issues							
Generate Feature Vector for Machine Learning (Tutorial 8) 2							1
#13 opened 12 hours ago by vilasmamidyala Generate Feature V... New Issues							
Conduct Topic Discovery (Tutorial 8) 1							1
#12 opened 13 hours ago by vilasmamidyala Conduct Topic Disco... In Progress							
Conduct Name Entity Extraction/Relation Extraction 1							1
#11 opened 13 hours ago by vilasmamidyala Conduct Name Entit... New Issues							
Conduct NGram and Word2Vec (Tutorial 6) 1							1
#10 opened 13 hours ago by vilasmamidyala Conduct NGram and... New Issues							
Try to use WordNet (Tutorial 7) 1							1
#9 opened 13 hours ago by vilasmamidyala Try to use WordNet (... New Issues							








TEAM 8 : VILAS MAMIDYALA (18)

VIKESH PADARTHI (27)

DINESH KUMAR BANDAM (2)

BHUMIREDDY RANJITHA REDDY (4)

Milestones:

Documenattion par2 Closed 2 minutes ago ⌚ Last updated less than a minute ago continue editing the document 2	 100% complete 0 open 1 closed Edit Reopen Delete
Try to use WordNet (Tutorial 7) Closed 2 minutes ago ⌚ Last updated less than a minute ago Please complete this task at the possible earliest and once you are... (more)	 100% complete 0 open 1 closed Edit Reopen Delete
Conduct NGram and Word2Vec Closed 2 minutes ago ⌚ Last updated less than a minute ago Please complete this task at the possible earliest and once you are... (more)	 100% complete 0 open 1 closed Edit Reopen Delete
Conduct Name Entity Extraction/Relation Extraction Closed 2 minutes ago ⌚ Last updated less than a minute ago Please complete this task at the possible earliest and once you are... (more)	 100% complete 0 open 1 closed Edit Reopen Delete
Conduct Topic Discovery (Tutorial 8) Closed 2 minutes ago ⌚ Last updated less than a minute ago Please complete this task at the possible earliest and once you are... (more)	 100% complete 0 open 1 closed Edit Reopen Delete
Generate Feature Vector for Machine Learning (Tutorial 8) Closed 2 minutes ago ⌚ Last updated less than a minute ago Please complete this task at the possible earliest and once you are... (more)	 100% complete 0 open 1 closed Edit Reopen Delete
Documentation changes Closed 2 minutes ago ⌚ Last updated less than a minute ago Please complete this task at the possible earliest and once you are... (more)	 100% complete 0 open 1 closed Edit Reopen Delete

COMP-SCI 5560 (SUMMER 2016) – KNOWLEDGE DISCOVERY AND MANAGEMENT

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

vilasmamidyala / KDM_SM16_SM

Unwatch 3 Star 0 Fork 1

Code Issues 3 Pull requests 0 Boards Burndown Wiki Pulse Graphs Settings

Repos (1/1) show one Labels Milestones Assignees

Search (/) New issue

0 New Issues

0 Icebox

1 Backlog

- KDM_SM16_SM #5 Document part 3

2 In Progress

- KDM_SM16_SM #1 architecture diagram and sequence diagram
- KDM_SM16_SM #7 collect twitter data

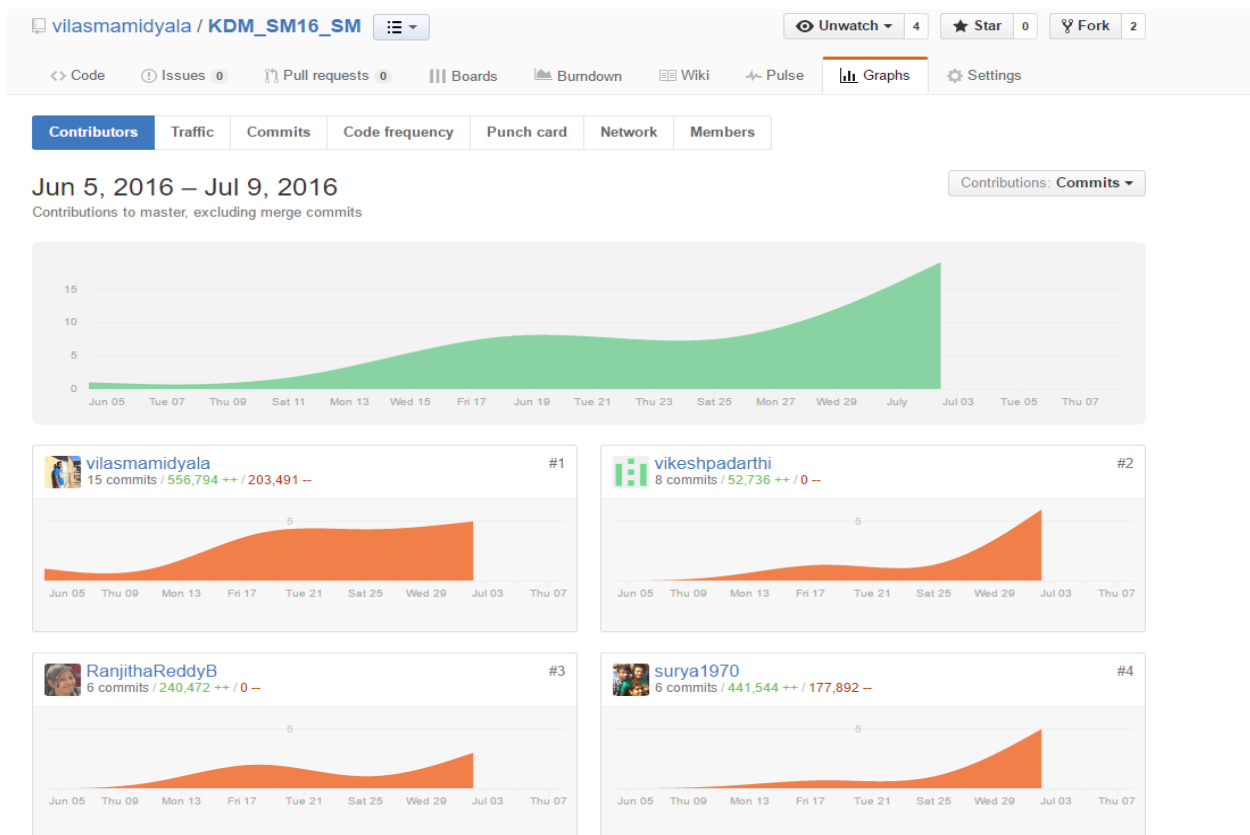
0 Review/QA

0 Done

4 Closed

- KDM_SM16_SM #6 task for wordcount
- KDM_SM16_SM #3 Try NLP processing
- KDM_SM16_SM #4 Try Information Extraction/Retrieval technologies
- KDM_SM16_SM #2 documentation-part1

Contribution of source code github:



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 and 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 sources 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.