CS5560 KNOWLEDGE DISCOVERY AND MANAGEMENT

PROJECT REPORT-PHASE 2

SUMMARIZATION OF GOOGLE/TWITTER TRENDING NEWS

TEAM MATES

1. PRASANNA MUPPIDI(23)
2. SANTHOSH MOHAN MURARISHETTI(24)
3. ANUDEEP PANDIRI(29)
4. VAMSHI RAJARIKAM(36)

Index

1. Introduction:

1.1 Motivation:

Every social networking site provides data about the trending news but only to the users who have an existing account in that particular website. Displaying and letting every user know about the latest or the trending news is very important sometimes. For example, Earthquake. It’s very important for people to know about the status of the earthquake or status of people in particular locations. To let the users know about the latest or trending news with so much ease is the main motivation of our application.

1.2 Objective:

Our application takes information regarding all the latest or trending news from various social networking sites using their particular APIs and displays the summarized data to the users. Summarization is done using various NLPs. The unstructured data which is collected from the social networking sites is processed using TF-IDF and the output data is structured. This structured data is easily understandable by the users. It’s not required for the users to Login into the application and view the news. The news which is displayed is clear cut without any unnecessary information making it easy for the users. Collecting, Managing, Summarizing and Displaying all the trending news is the main objective of our application.

1.3 Expected Outcomes:

When we give an unstructured data file as an input, summarized and easily understandable data would be the outcome of our application.

2. Domain

- **Topic:** Summarization of trending topics in the world from social networks.
- **Technologies Used:**
  - ✓ Languages: Java, Scala
  - ✓ IDE: Intellij Idea
  - ✓ Frameworks: Spark
  - ✓ Libraries: CoreNLP, SparkNLP, JSON, WordNet, SparkMl
  - ✓ APIS: Google Trends, Guardian, NYtimes, USAtoday

3. Data Collection

There are many open source datasets available in the internet. There are mainly two types of data collection.

1) Static Data
2) Real Time Data using APIs

3.1 Static Data:

Static data is nothing but data which is immutable and not changes during time. It is a fixed data set. We collected static data from twitter using curl command

Example:

```
curl --get 'https://stream.twitter.com/1.1/statuses/sample.json' --header 'Authorization:
OAuth oauth_consumer_key="TjIDZ6XQX6TZOq64EZ49SatYb",
oauth_nonce="d54db403fb54cc9e5a10a92bb2741e6e",
oauth_signature="GqHWmZKgD8YO6rX5HgGKRuMFWGQ%3D",
oauth_signature_method="HMAC-SHA1", oauth_timestamp="1457754994",
oauth_token="453746488-vDRGN511Pk3g3tSvOhpgIdSRErFjXP5fClexkpWp",
oauth_version="1.0"' --verbose> tweet.txt
```

3.2 Real Time Data Using APIs:

Real time data is the data which changes dynamically with time. This data is provided by some open source APIs like Google trends, Hawt trends, USAToday, etc.,

Presently, we are collecting data using google trends API which gives JSON output. We are storing those data in text file which we will provide as input for future analysis.

API:

Google Trends:

https://www.google.com/trends/api/stories/latest?cat=m&fi=15&fs=15&geo=US&ri=300&rs=15&tz=300

Newyork Times:

https://api.nytimes.com/svc/search/v2/articlesearch.json?api-key=1069bc25bff24ebf8cf3dbae1133e000&q=tech&sort=newest&fl=lead_paragraph&page=0

4. Tasks

4.1 Rest API Service:

We are using Rest API service using HTTPURL connection in java. The API provides the top trending topics in the world with source name, article name and URL of the website.

Input: **”https://www.google.com/trends/api/stories/latest?cat=m&fi=15&fs=15&geo=US&ri=300&rs=15&tz=300”**

Output: **https://github.com/murarishetty/KDM-SM2016-TEAM-9/blob/master/src/gTrends**

The useful information is extracted from the response and stored in text file.

4.2 CoreNLP:

CoreNLP provides natural language processing tools which does grammatical analysis on the words.

Input: **https://github.com/murarishetty/KDM-SM2016-TEAM-9/blob/master/src/gTrends**

Output: **https://github.com/murarishetty/KDM-SM2016-TEAM-9/blob/master/documentation/CoreNLP_TF-IDF_output.docx**

4.3 TF-IDF:

TF-IDF of a word says how important a word is in a document. It is an important factor in determining the weightage of the word in the document or a collection of documents.

Input: **https://github.com/murarishetty/KDM-SM2016-TEAM-9/blob/master/src/gTrends**

Output: **https://github.com/murarishetty/KDM-SM2016-TEAM-9/blob/master/documentation/CoreNLP_TF-IDF_output.docx**

4.4 NGram and Word2Vec:

NGram is a sequence of N words from a given input text. NGram is a language model which is used to predict the next word in the corpus.

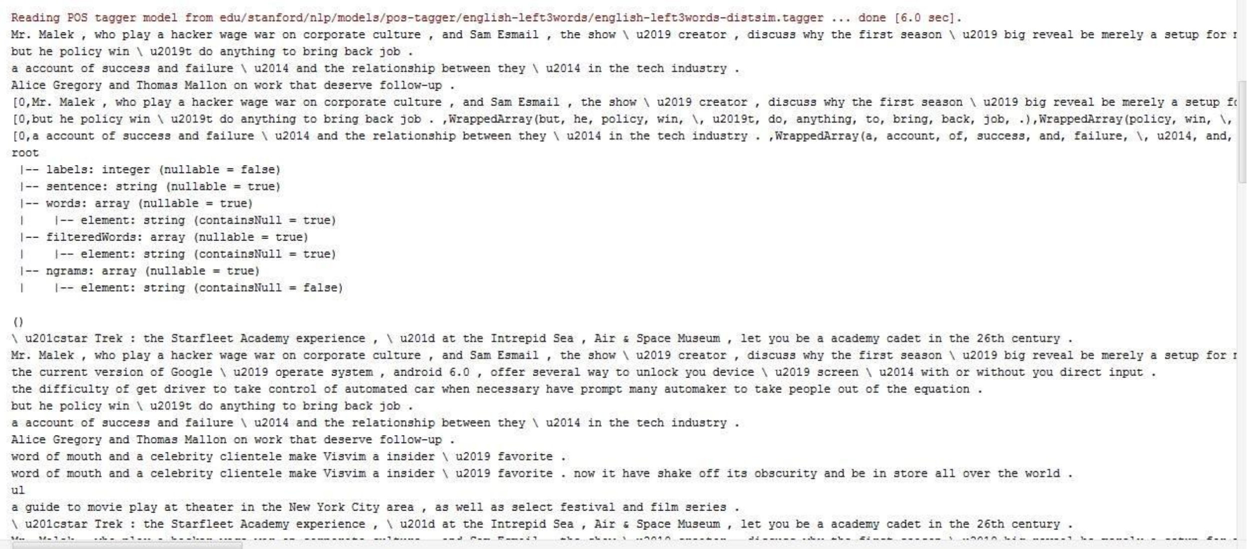
Word2Vec takes a large corpus of text data and gives a vector for each unique word as its output.

News Article 1: https://drive.google.com/file/d/0B4VHwW192C9HZ0s1QmU3dlp0VVE/view

News Article 2: https://drive.google.com/file/d/0B4VHwW192C9HZmQ2VTBCaHdsNTA/view

Output:https://drive.google.com/file/d/0B4VHwW192C9HaXAzeXpoelk1WWM/view?ts=57805ff1

Output Screenshot:



```
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done [6.0 sec].
Mr. Malek , who play a hacker wage war on corporate culture , and Sam Esmail , the show \ u2019 creator , discuss why the first season \ u2019 big reveal be merely a setup for
but he policy win \ u2019t do anything to bring back job .
a account of success and failure \ u2014 and the relationship between they \ u2014 in the tech industry .
Alice Gregory and Thomas Mallon on work that deserve follow-up .
[0,Mr. Malek , who play a hacker wage war on corporate culture , and Sam Esmail , the show \ u2019 creator , discuss why the first season \ u2019 big reveal be merely a setup fo
[0,but he policy win \ u2019t do anything to bring back job . ,WrappedArray(but, he, policy, win, \, u2019t, do, anything, to, bring, back, job, .),WrappedArray(policy, win, \,
[0,a account of success and failure \ u2014 and the relationship between they \ u2014 in the tech industry . ,WrappedArray(a, account, of, success, and, failure, \, u2014, and,
root
 |-- labels: integer (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)

()
\ u201cstar Trek : the Starfleet Academy experience , \ u201d at the Intrepid Sea , Air & Space Museum , let you be a academy cadet in the 26th century .
Mr. Malek , who play a hacker wage war on corporate culture , and Sam Esmail , the show \ u2019 creator , discuss why the first season \ u2019 big reveal be merely a setup for
the current version of Google \ u2019 operate system , android 6.0 , offer several way to unlock you device \ u2019 screen \ u2014 with or without you direct input .
the difficulty of get driver to take control of automated car when necessary have prompt many automaker to take people out of the equation .
but he policy win \ u2019t do anything to bring back job .
a account of success and failure \ u2014 and the relationship between they \ u2014 in the tech industry .
Alice Gregory and Thomas Mallon on work that deserve follow-up .
word of mouth and a celebrity clientele make Visvim a insider \ u2019 favorite .
word of mouth and a celebrity clientele make Visvim a insider \ u2019 favorite . now it have shake off its obscurity and be in store all over the world .
ul
a guide to movie play at theater in the New York City area , as well as select festival and film series .
\ u201cstar Trek : the Starfleet Academy experience , \ u201d at the Intrepid Sea , Air & Space Museum , let you be a academy cadet in the 26th century .
```

4.5 Name Entity Extraction:

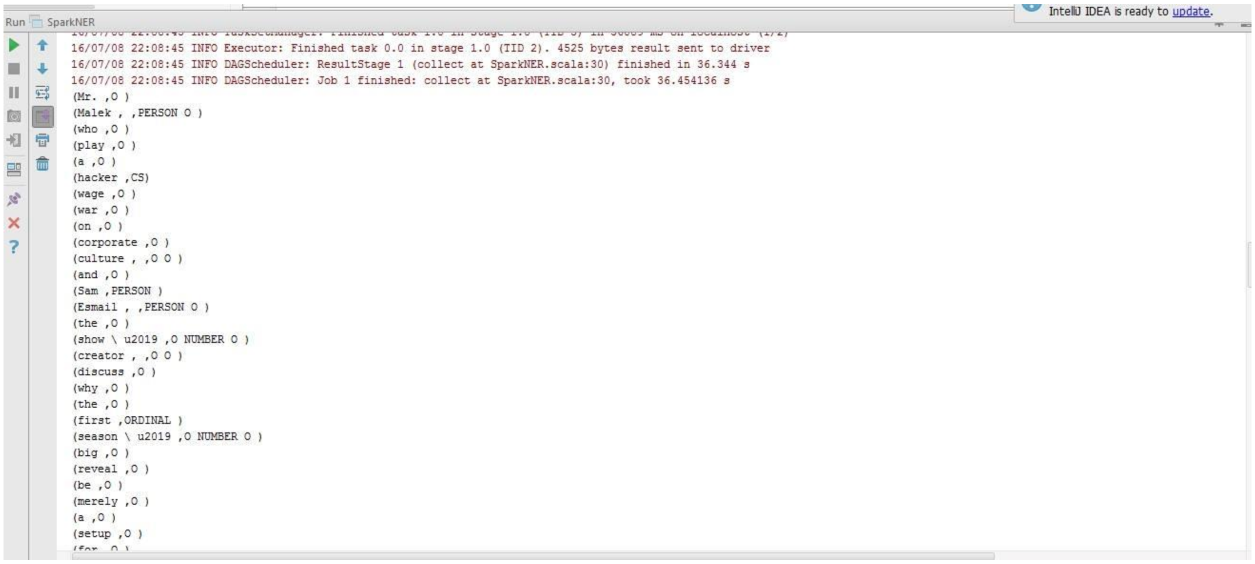
Name entity extraction specifies and tags words such as persons, places, organizations in the given input corpus.

News Article 1: https://drive.google.com/file/d/0B4VHwW192C9HZ0s1QmU3dlp0VVE/view

News Article 2: https://drive.google.com/file/d/0B4VHwW192C9HZmQ2VTBCaHdsNTA/view

Output:https://drive.google.com/file/d/0B4VHwW192C9HaXAzeXpoelk1WWM/view?ts=57805ff1

Output Screenshot:



```
16/07/08 22:08:45 INFO Executor: Finished task 0.0 in stage 1.0 (TID 2). 4525 bytes result sent to driver
16/07/08 22:08:45 INFO DAGScheduler: ResultStage 1 (collect at SparkNER.scala:30) finished in 36.344 s
16/07/08 22:08:45 INFO DAGScheduler: Job 1 finished: collect at SparkNER.scala:30, took 36.454136 s
(Mr. ,O )
(Malek , ,PERSON O )
(who ,O )
(play ,O )
(a ,O )
(hacker ,CS)
(wage ,O )
(war ,O )
(on ,O )
(corporate ,O )
(culture , ,O O )
(and ,O )
(Sam ,PERSON )
(Esmail , ,PERSON O )
(the ,O )
(show \ u2019 ,O NUMBER O )
(creator , ,O O )
(discuss ,O )
(why ,O )
(the ,O )
(first ,ORDINAL )
(season \ u2019 ,O NUMBER O )
(big ,O )
(reveal ,O )
(be ,O )
(merely ,O )
(a ,O )
(setup ,O )
(for ,O )
```

4.6 WordNet:

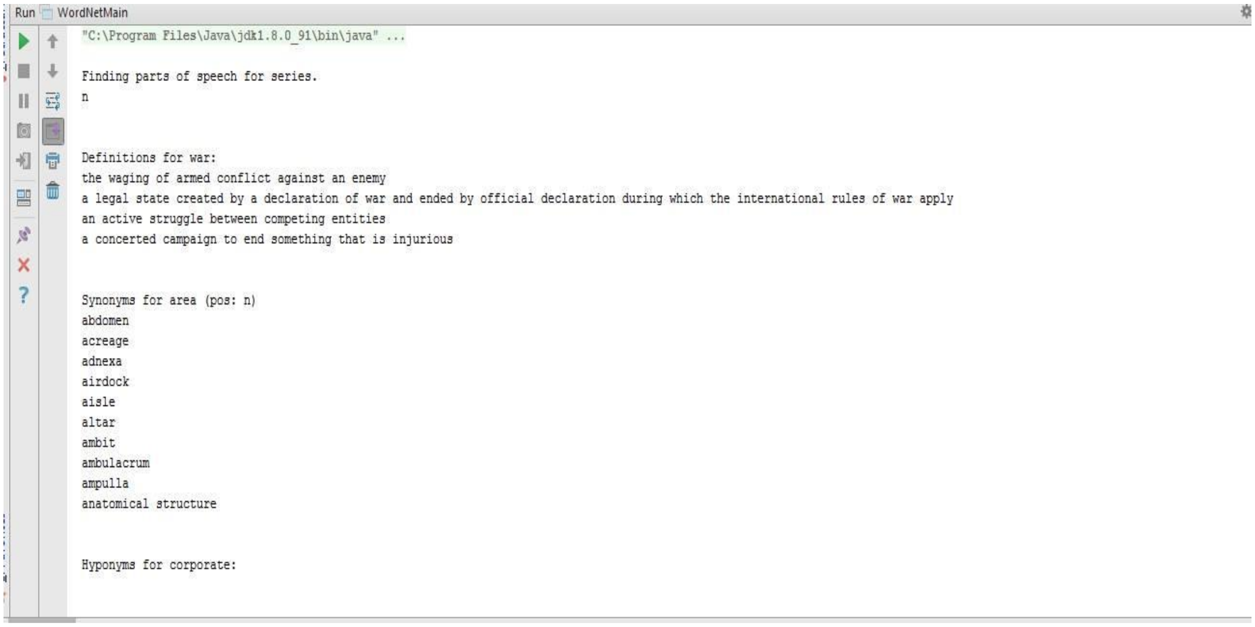
WordNet specifies the semantic relationship between different parts of the same object.

News Article 1: https://drive.google.com/file/d/0B4VHwW192C9HZ0s1QmU3dlp0VVE/view

News Article 2: https://drive.google.com/file/d/0B4VHwW192C9HZmQ2VTBCaHdsNTA/view

Output:
https://drive.google.com/file/d/0B4VHwW192C9HaXAzeXpoelk1WWM/view?ts=57805ff1

Output Screenshot:



```
"C:\Program Files\Java\jdk1.8.0_91\bin\java" ...

Finding parts of speech for series.
n


Definitions for war:
the waging of armed conflict against an enemy
a legal state created by a declaration of war and ended by official declaration during which the international rules of war apply
an active struggle between competing entities
a concerted campaign to end something that is injurious


Synonyms for area (pos: n)
abdomen
acreage
adnexa
airdock
aisle
altar
ambit
ambulacrum
ampulla
anatomical structure


Hyponyms for corporate:
```

4.7 SparkLDA:

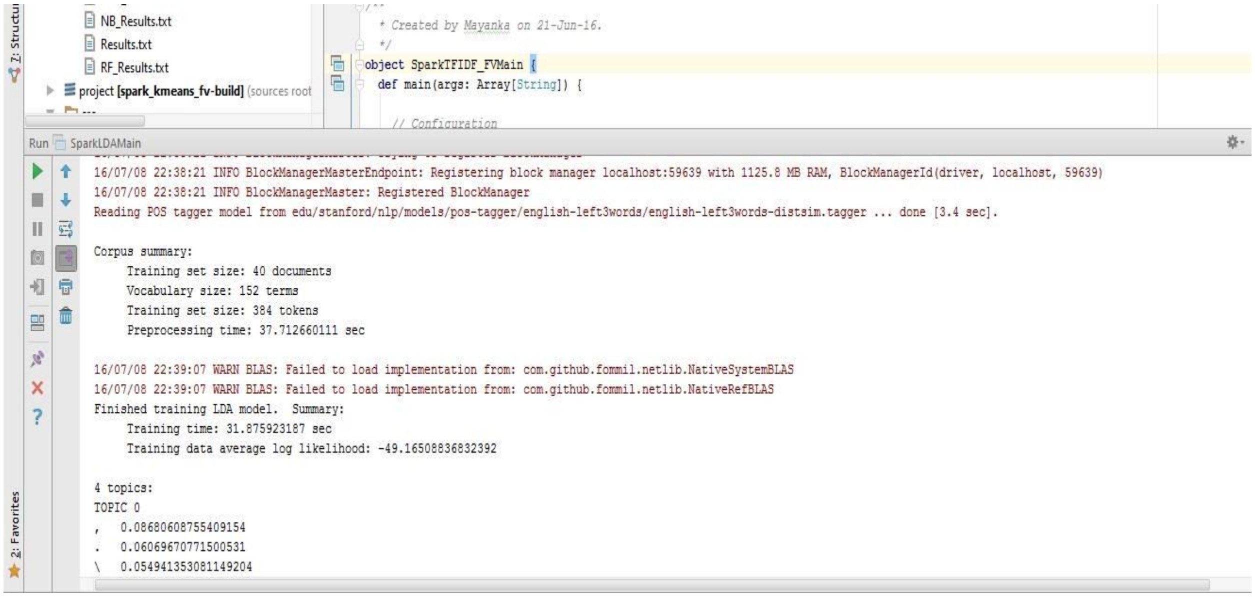
In SparkLDA, each word in each document is tagged under a specific topic.

News Article 1: https://drive.google.com/file/d/0B4VHwW192C9HZ0s1QmU3dlp0VVE/view

News Article 2: https://drive.google.com/file/d/0B4VHwW192C9HZmQ2VTBCaHdsNTA/view

Output:
https://drive.google.com/file/d/0B4VHwW192C9HaXAzeXpoelk1WWM/view?ts=57805ff1

Output Screenshot:



```
 * Created by Mayanka on 21-Jun-16.
 */
object SparkTFIDF_FVMain {
  def main(args: Array[String]) {

    // Configuration
```

```
16/07/08 22:38:21 INFO BlockManagerMasterEndpoint: Registering block manager localhost:59639 with 1125.8 MB RAM, BlockManagerId(driver, localhost, 59639)
16/07/08 22:38:21 INFO BlockManagerMaster: Registered BlockManager
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done [3.4 sec].

Corpus summary:
     Training set size: 40 documents
     Vocabulary size: 152 terms
     Training set size: 384 tokens
     Preprocessing time: 37.712660111 sec

16/07/08 22:39:07 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
16/07/08 22:39:07 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
Finished training LDA model.  Summary:
     Training time: 31.875923187 sec
     Training data average log likelihood: -49.16508836832392

4 topics:
TOPIC 0
,   0.08680608755409154
.   0.06069670771500531
\   0.054941353081149204
```

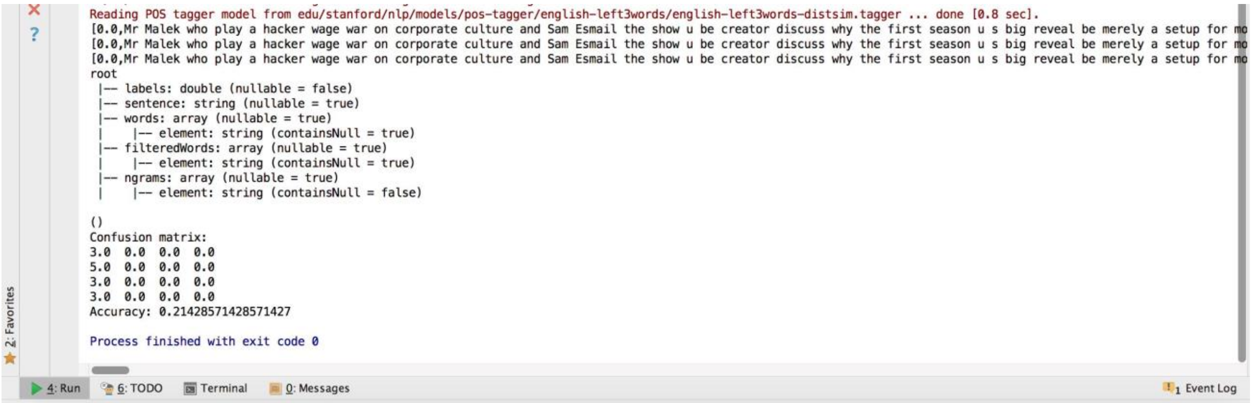
4.8 Feature Vector:

News Article 1: https://drive.google.com/file/d/0B4VHwW192C9HZ0s1QmU3dlp0VVE/view

News Article 2: https://drive.google.com/file/d/0B4VHwW192C9HZmQ2VTBCaHdsNTA/view

Output:
https://drive.google.com/file/d/0B4VHwW192C9HaXAzeXpoelk1WWM/view?ts=57805ff1

Output Screenshot:

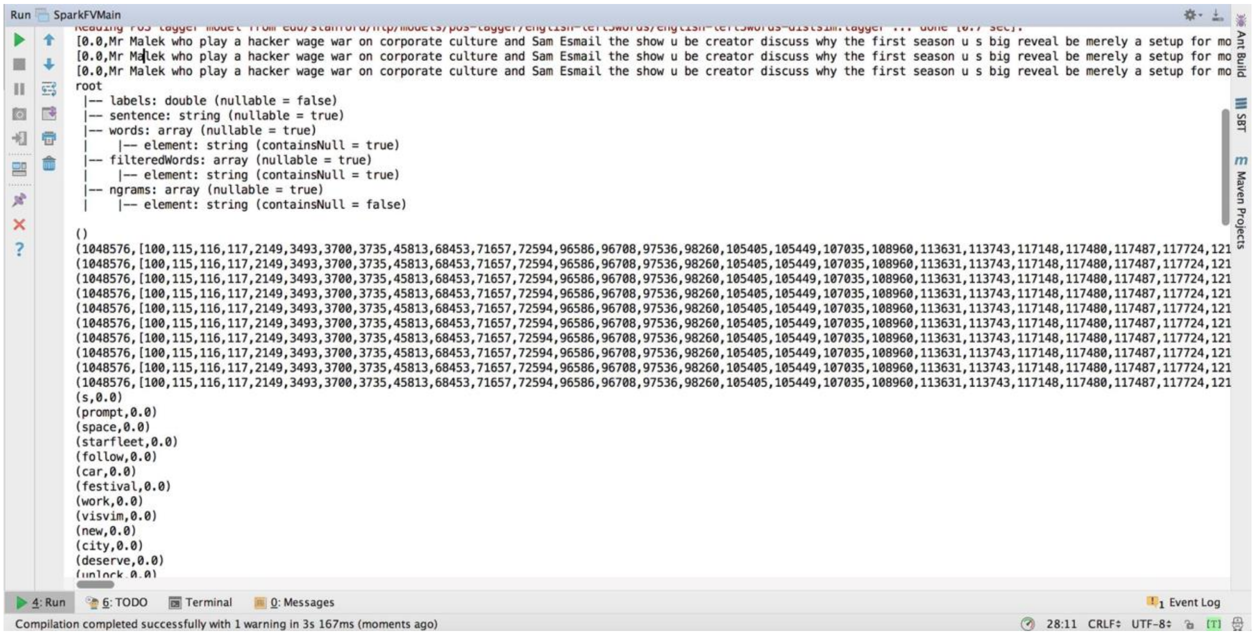


```
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done [0.8 sec].
[0.0,Mr Malek who play a hacker wage war on corporate culture and Sam Esmail the show u be creator discuss why the first season u s big reveal be merely a setup for mo
[0.0,Mr Malek who play a hacker wage war on corporate culture and Sam Esmail the show u be creator discuss why the first season u s big reveal be merely a setup for mo
[0.0,Mr Malek who play a hacker wage war on corporate culture and Sam Esmail the show u be creator discuss why the first season u s big reveal be merely a setup for mo
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:
3.0  0.0  0.0  0.0
5.0  0.0  0.0  0.0
3.0  0.0  0.0  0.0
3.0  0.0  0.0  0.0
Accuracy: 0.21428571428571427

Process finished with exit code 0
```

Feature Vector with TF-IDF:



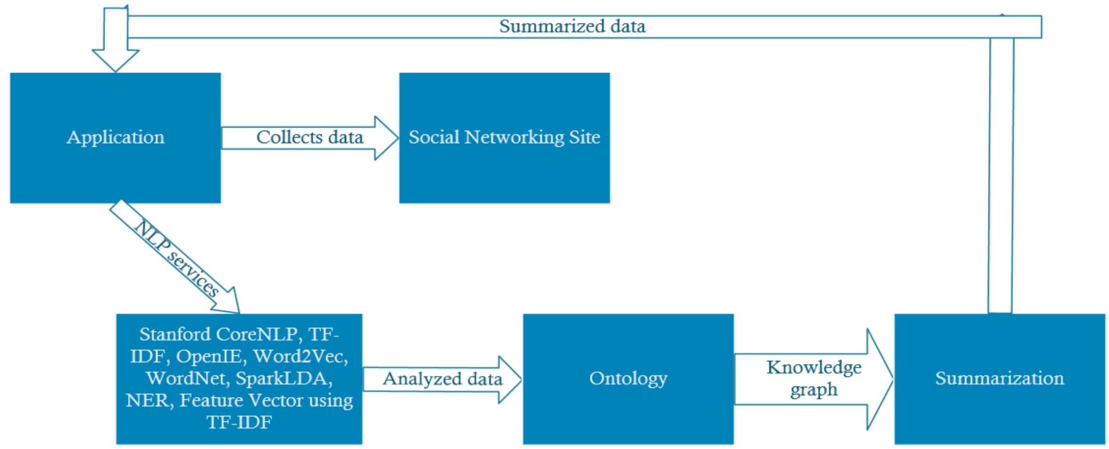
```
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done [0.7 sec].
[0.0,Mr Malek who play a hacker wage war on corporate culture and Sam Esmail the show u be creator discuss why the first season u s big reveal be merely a setup for mo
[0.0,Mr Malek who play a hacker wage war on corporate culture and Sam Esmail the show u be creator discuss why the first season u s big reveal be merely a setup for mo
[0.0,Mr Malek who play a hacker wage war on corporate culture and Sam Esmail the show u be creator discuss why the first season u s big reveal be merely a setup for mo
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)

()
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(1048576,[100,115,116,117,2149,3493,3700,3735,45813,68453,71657,72594,96586,96708,97536,98260,105405,105449,107035,108960,113631,113743,117148,117480,117487,117724,121
(s,0.0)
(prompt,0.0)
(space,0.0)
(starfleet,0.0)
(follow,0.0)
(car,0.0)
(festival,0.0)
(work,0.0)
(visvim,0.0)
(new,0.0)
(city,0.0)
(deserve,0.0)
(unlock,0.0)
```

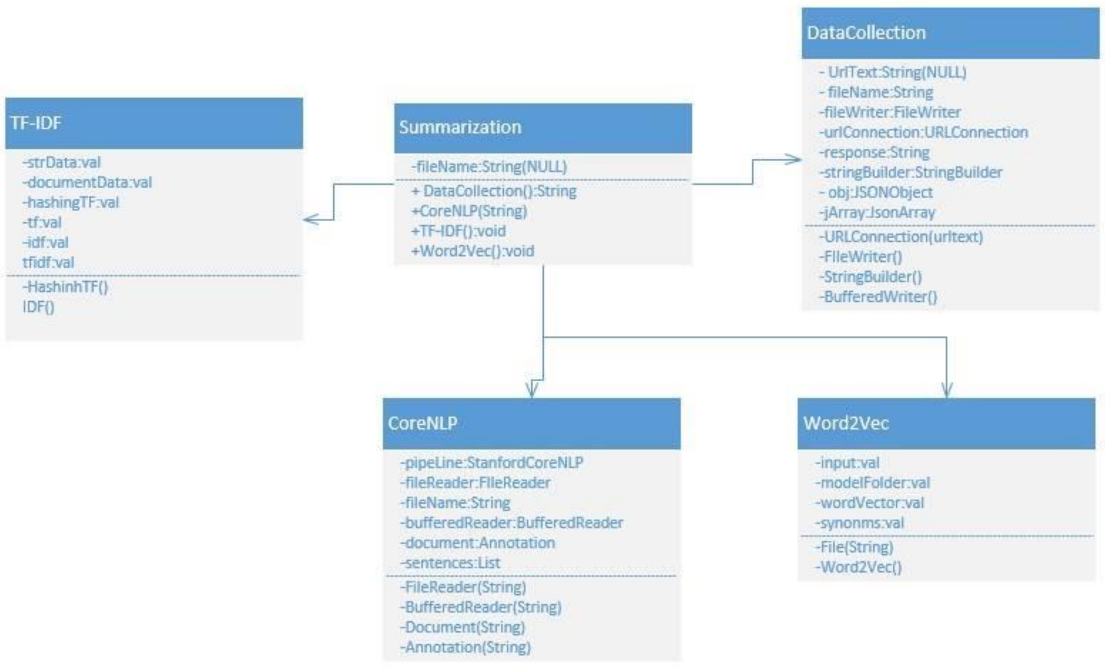
```
Compilation completed successfully with 1 warning in 3s 167ms (moments ago)
28:11  CRLF÷  UTF-8÷
```

5. Implementation Specification:

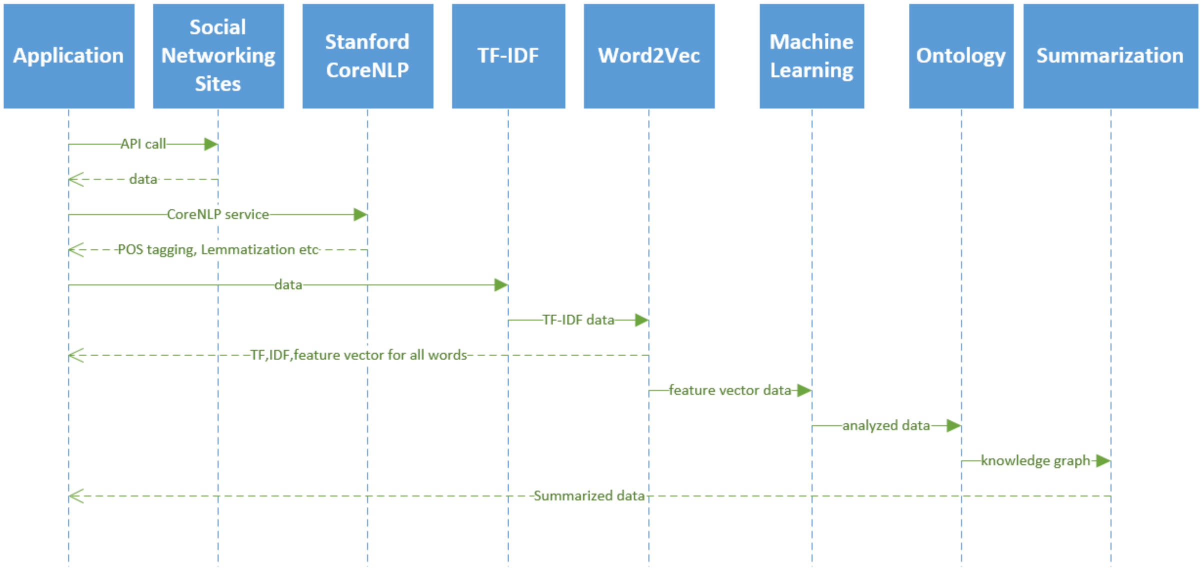
5.1 Architecture Diagram



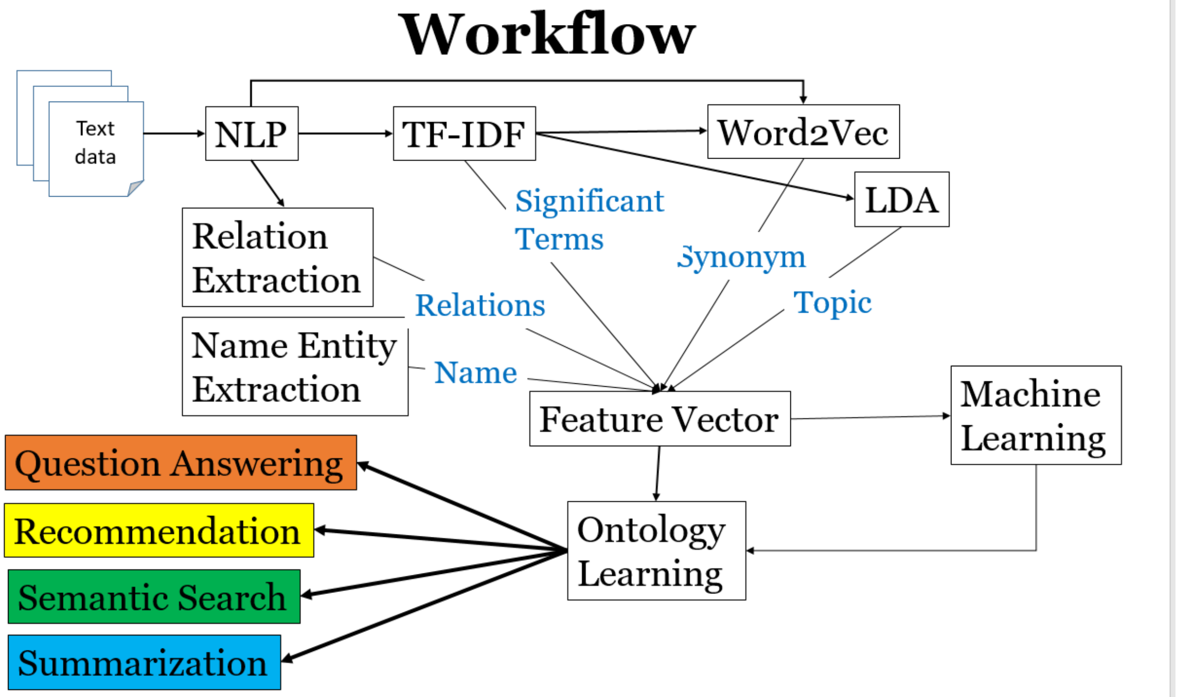
5.2 Class Diagram



5.3 Sequence Diagram



5.4 Workflow

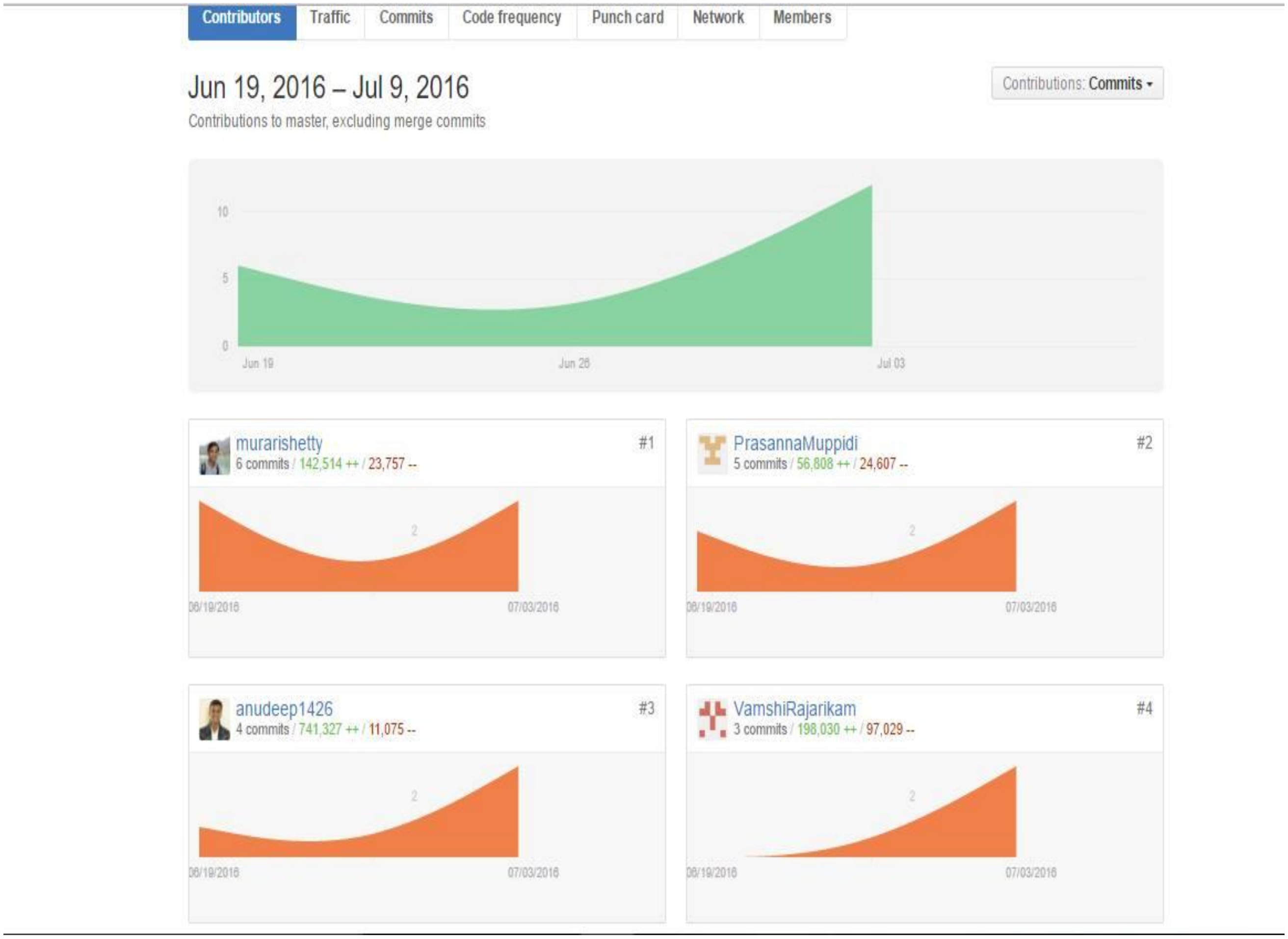


5.5 Existing Services Used

- Stanford CoreNLP
- SparkNLP(TF-IDF)
- OpenIE
- Word2Vec
- SparkLDA
- Name Entity recognition
- WordNet
- Feature Vector
- RestAPI

6. Project Management

6.1 Github



Jun 19, 2016 – Jul 9, 2016

Contributions to master, excluding merge commits

murarishetty
6 commits / 142,514 ++ / 23,757 --

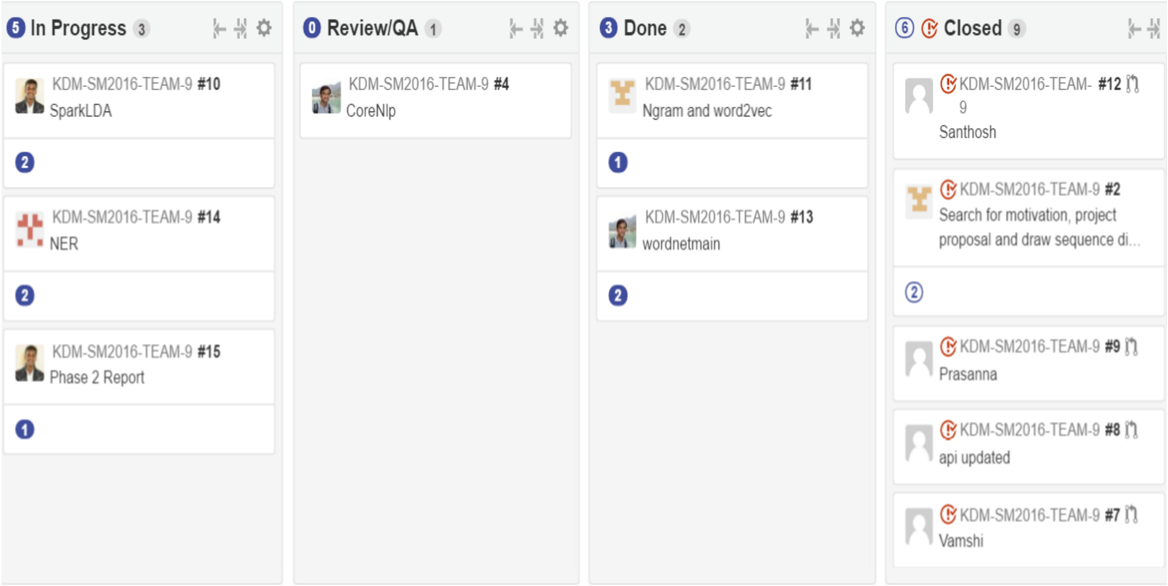
PrasannaMuppidi
5 commits / 56,808 ++ / 24,607 --

anudeep1426
4 commits / 741,327 ++ / 11,075 --

VamshiRajarikam
3 commits / 198,030 ++ / 97,029 --

Fig: Code Contribution

6.2 Zenhub



5 In Progress 3

KDM-SM2016-TEAM-9 **#10**
SparkLDA
2

KDM-SM2016-TEAM-9 **#14**
NER
2

KDM-SM2016-TEAM-9 **#15**
Phase 2 Report
1

0 Review/QA 1

KDM-SM2016-TEAM-9 **#4**
CoreNlp

3 Done 2

KDM-SM2016-TEAM-9 **#11**
Ngram and word2vec
1

KDM-SM2016-TEAM-9 **#13**
wordnetmain
2

6 Closed 9

KDM-SM2016-TEAM-9 **#12**
Santhosh

KDM-SM2016-TEAM-9 **#2**
Search for motivation, project proposal and draw sequence di...
2

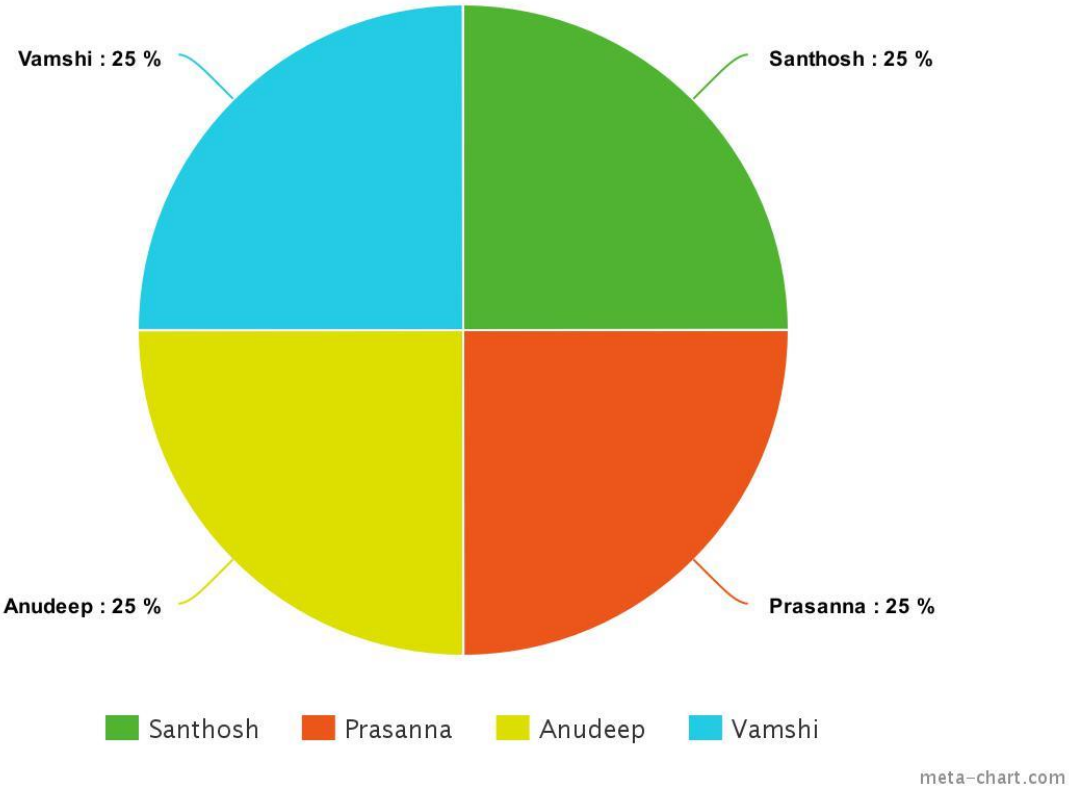
KDM-SM2016-TEAM-9 **#9**
Prasanna

KDM-SM2016-TEAM-9 **#8**
api updated

KDM-SM2016-TEAM-9 **#7**
Vamshi

6.3 Contribution

Name	Work
Santhosh Mohan Murarishetti	Data Collection using API (Real Time), Design of Application Workflow, WordNet, NER Implementation
Prasanna Muppidi	CoreNLP, NGram, Word2Vec Implementation
Anudeep Pandiri	TF-IDF, SparkLDA Implementation
Vamshi Rajarikam	Feature Vector implementation



7. References

- https://www.google.com/trends/
- https://www.quora.com/Does-Google-Trends-have-a-publicly-available-API
- http://spark.apache.org/documentation.html
- https://spark.apache.org/docs/1.1.0/mllib-feature-extraction.html
- http://stanfordnlp.github.io/CoreNLP/
- https://developer.usatoday.com/