CS5560 KNOWLEDGE DISCOVERY AND MANAGEMENT

PROJECT REPORT-PHASE 1

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
  - ✓ APIS: Google Trends

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:

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

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:

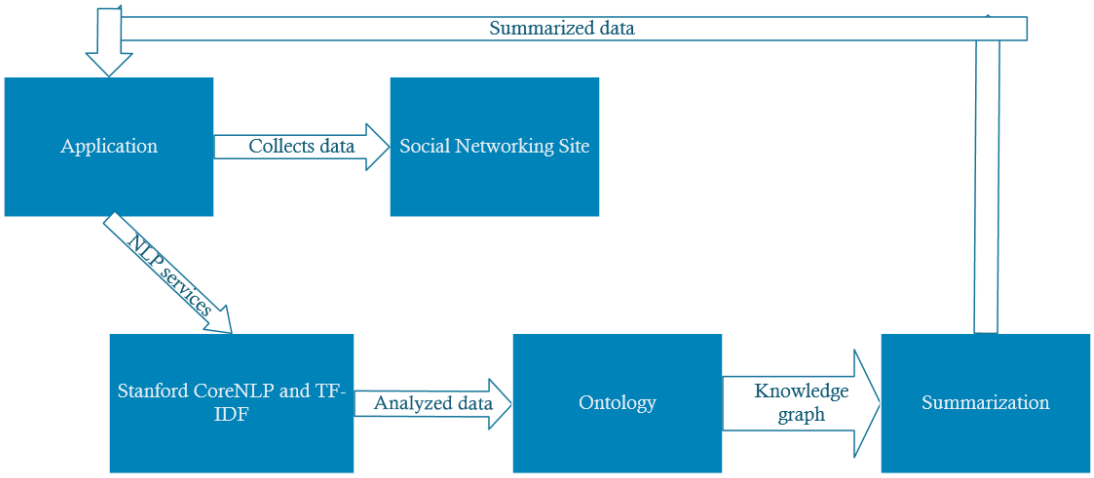
TF-IDF of a word says how important a word is in a document. It is an important factor in determining the weightage of the word in the document or a collection of documents.

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

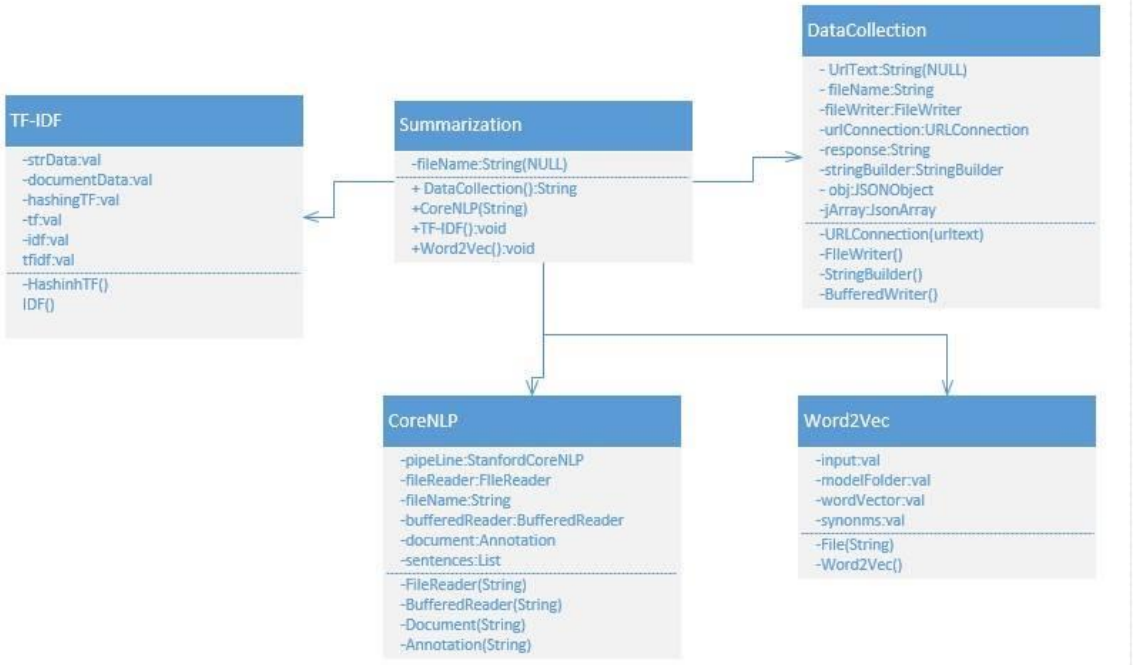
Output: https://github.com/murarishetty/KDM-SM2016-TEAM-9/blob/master/documentation/CoreNLP_TF-IDF_output.docx

5. Implementation Specification:

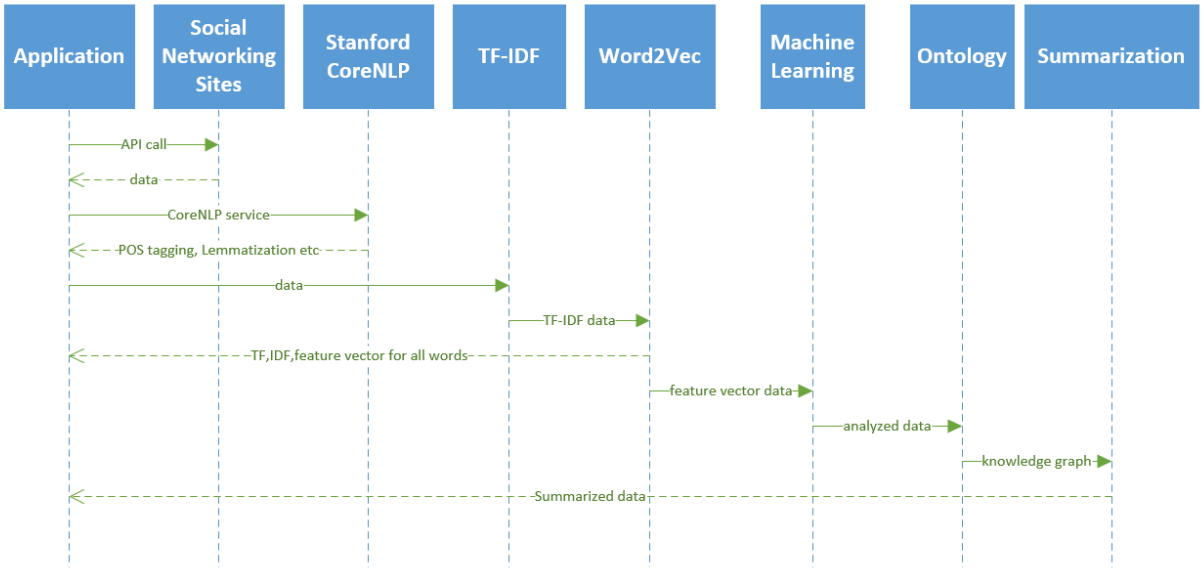
5.1 Architecture Diagram



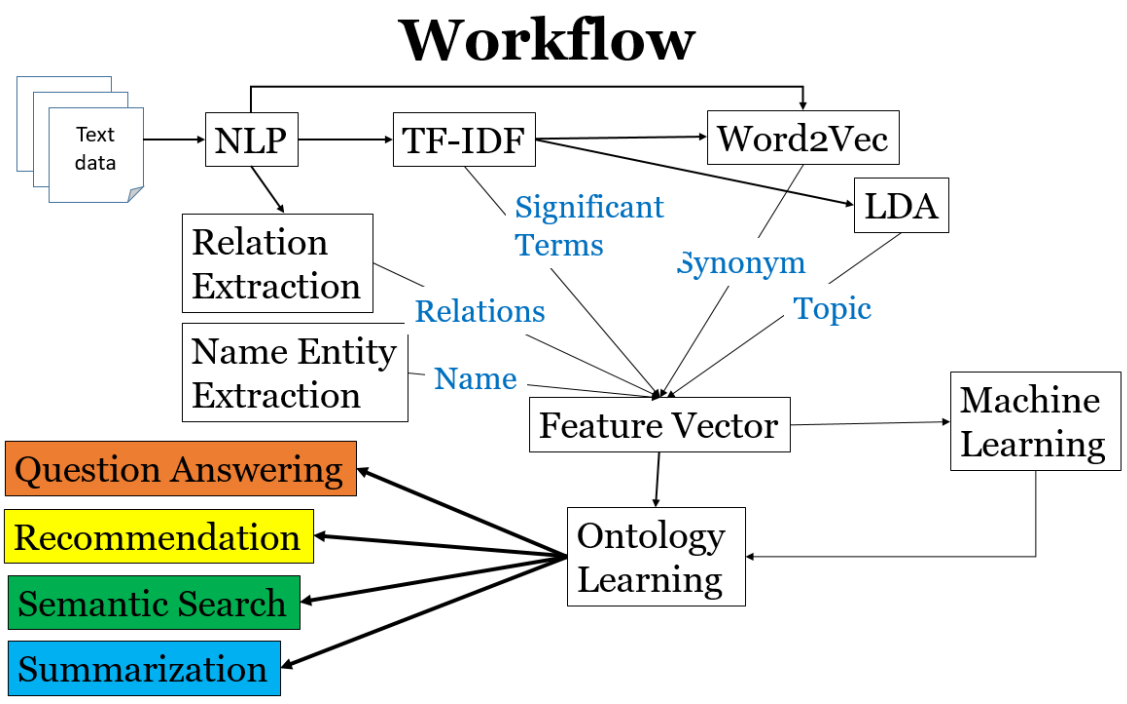
5.2 Class Diagram



5.3 Sequence Diagram



5.4 Workflow



5.5 Existing Services Used

- Stanford CoreNLP
- SparkNLP(TF-IDF)
- RestAPI

6. Project Management

6.1 Github

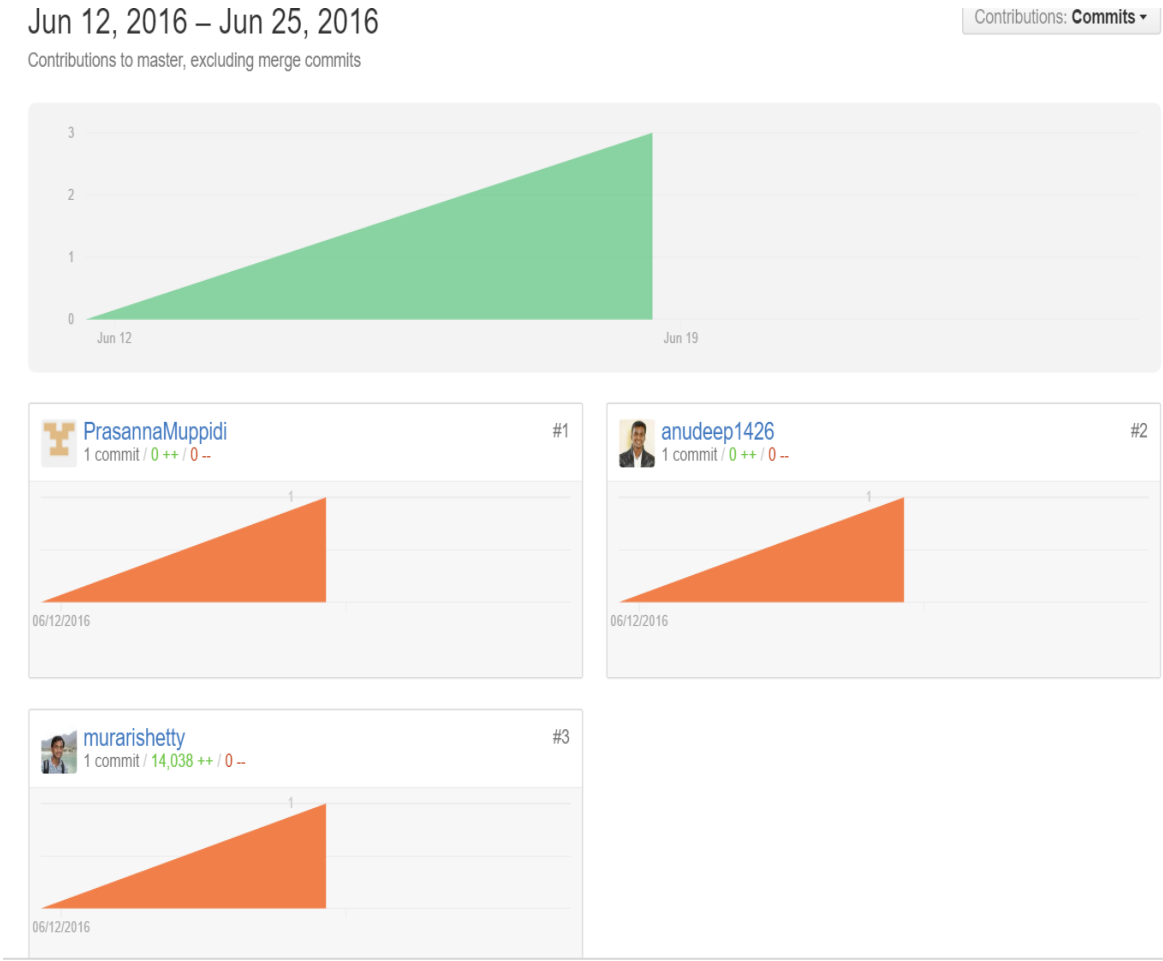
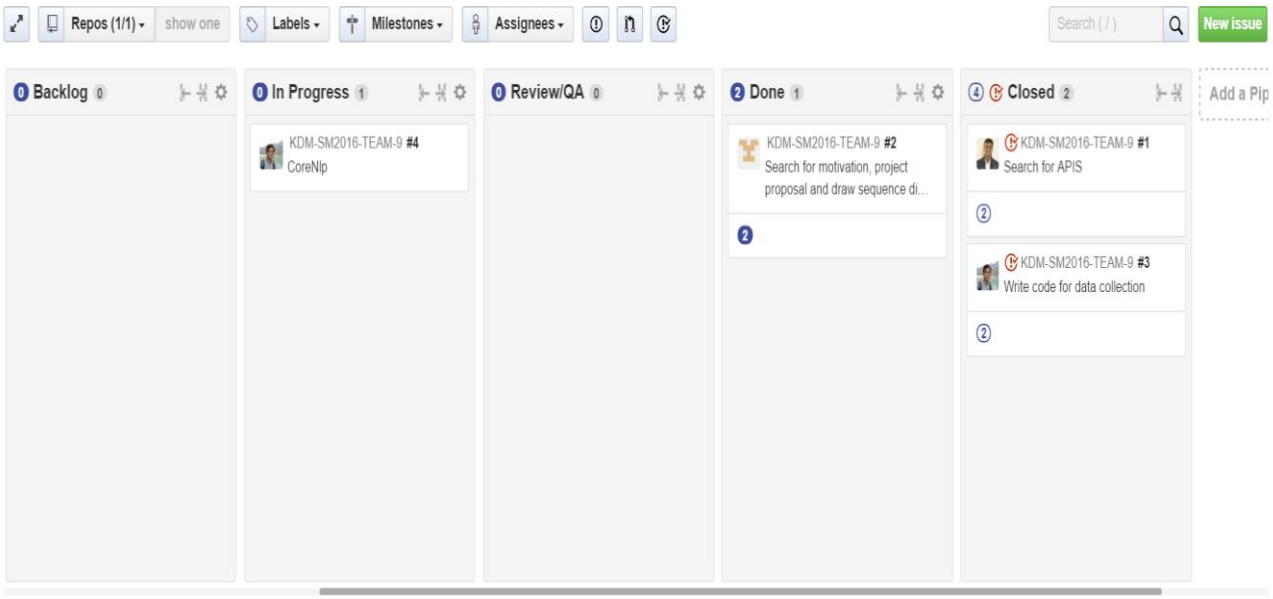


Fig: Code Contribution

6.2 Zenhub



6.3 Contribution

Name	Work
Santhosh Mohan Murarishetti	Data Collection using API (Real Time), Design of Application Workflow
Prasanna Muppidi	CoreNLP
Anudeep Pandiri	TF-IDF
Vamshi Rajarikam	Documentation

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/