An Information Retrieval Report

on

One-Stop News Buzz

Submitted by

Voruganti Sai Harini 14IT152
Divya Narayanan 14IT211
Jaju Krishna Bhagwan 14IT213

under the guidance of

Dr. Sowmya Kamath
Department of Information Technology, NITK Surathkal

in partial fulfillment for the award of the degree
of

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

At

Department of Information Technology
National Institute of Technology Karnataka, Surathkal.

CERTIFICATE

This is to certify that the project entitled "One-Stop News Buzz" is a bonafide work carried out as a part of the course INFORMATION RETRIEVAL(IT362), under my guidance by

1.Voruganti Sai Harini (14IT152)
2.Divya Narayanan(14IT211)
3.Jaju Krishna Bhagwan (14IT213)

Students of VI Sem B.Tech (IT) at the Department of Information Technology, National Institute of Technology Karnataka , Surathkal , during the academic semester of JAN-MAY 2017 in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology , at NITK Surathkal.

Place:
Date :

Signature of the Instructor

DECLARATION

We hereby declare that the project entitled "One-Stop News Buzz" submitted as part of the partial course requirements for the course Information Retrieval (IT362) for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal during the JAN-MAY 2017 semester has been carried out by us. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.Further, we declare that we will not share, re-submit or publish the code, idea, framework and any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the course Instruction.

Name and Signature of the Student :
1.Voruganti Sai Harini (14IT152)
2.Divya Narayanan(14IT211)
3.Jaju Krishna Bhagwan (14IT213)

PLACE :
DATE :

Abstract

Online news reading has become very popular as the web provides access to news articles from millions of sources around the world. The key challenge is to provide a wide coverage of news from all walks of life while keeping in mind not to overlook the highly important and trending ones. The application developed in this project provides user with quick and interesting walkthrough about the current affairs. The features provided are extremely friendly viz. Query based News Summarization, Summarization of single News article, Follow Up of News articles, Background of a particular News article, Breaking News based on Twitter, Sentimental Analysis of News articles and Recommendation of news articles.

Contents

List of Figures

1 Introduction

News reading has completely transformed with the advance of the Internet, from the traditional model of news consumption via physical newspaper subscription to access to thousands of sources via the Internet. A critical problem with news service websites is that the volumes of articles can be overwhelming to the users. The challenge is to help users find news articles that are interesting to read and provide a means to access the current affairs easily. This is what this application deals with. Multiple document summarization based on query is not something which is seen commonly. It also helps the user save a lot of time on reading lengthy news articles because the summaries obtained are pretty good and highlight all the important points from various relevant documents.

Also, sometimes the users may be interested in keeping himself updated on particular news articles, so the follow up of those news articles will be received by the user for few days. The user may also be interested in knowing about what could have happened with a particular news article in the past few days for which the background of the article is provided to the user.

Breaking news based on twitter is something that is never used in the past and is being implemented in this application. A simple sentimental analysis on news articles is also being done that will give the user an insight on whether the news talks about something that is positive or negative.

Recommendation of news articles is another area of interest. By extracting the tags of the user's starred documents, similar documents are recommended to the user.

2 Literature Survey

2.1 Background

Newspaper applications generally provide the users with latest news but these are lengthy to read and often time consuming for the user. They generally fail to provide summaries which just highlight the important points which can be used for quick glance. Also, the user may not have an option of getting to know about a trending topic from many websites. Breaking news is provided by websites but this is not based on a trending topic in a social networking website. News websites do not support an option that allows users to be updated with a particular topic of interest for upcoming days. The papers titled "A Hybrid News Recommendation Algorithm based on User's Browsing Path"[2], "Topic Retrieval and Articles Recommendation" [3], Personalized News Recommendation Based on Click Behavior" [4] , "A Contextual-Bandit Approach to Personalized News Article Recommendation"[5] , "WOOF: user defined news recommendation system"[6] were referred.

Reference [5] uses an approach in which a learning algorithm sequentially selects articles to serve users based on contextual information about the users and articles. [3] has analyzed that typical search results are large amount of articles, which match the keywords exactly but are on many different topics. It was understood that it is very time and effort consuming to read through the papers manually and select out desirable ones from them. In [6], there was simple classification of news. User is also provided with a short summary of the news.

But none of the above referred papers provided a way which in user could read newspaper articles in a time effective manner. Also there was no way in getting to know about a topic from multiple sources. News recommendation is being implemented in this project, but in addition to it, other functionalities are also being added.

2.2 Outcome of Literature Survey

The lack of many functionalities in most of the referred papers has motivated us to build an application to overcome these shortcomings and give a better experience to the user which is also time effective and informative. Breaking news based on Twitter is a new concept and provides a very quick way of being updated with the current news. Also query based summarization is another new feature not observed elsewhere. Sometimes, users may be interested in different type of news like positive or negative based on his mood, so these have also been added. A user maybe interested in being updated about a particular news article for a couple of days which was added to Follow Up. They may

also be interested in knowing what happened to a particular incident in the past couple of days which was included in background. These new additional features are expected to give a better experience to the user of the application. Some of the ideas that were used from [3] were query based retrieval of articles, which was used for multiple document summarization, follow up and recommendation. From [2] and [4] some insights obtained were that like in websites as clicks are recorded and used for recommending articles, similarly in this application, the starred and favorite documents of the user are used for retrieving documents that would be recommended to the user. Tags are extracted from them to perform this function. From [5], it was decided to use a ranking algorithm to perform keyword based matching of the query with the retrieved articles based on their frequency and consideration of the document length. As mentioned in [6], a short summary of the news article was provided to the user on demand, this includes both multiple and single document summarization. This would also reduce the time that user would spend on reading lengthy articles.

2.3 Problem Statement

To give an enriching News reading experience to the user by providing multiple features on a single platform. The main idea is to reduce the time that the users spends on being updated with current affairs by providing many easy to use functionalities all at a single platform many of which are not provided by other news applications .

2.4 Objectives

Following major objectives are recognized:

- Query based multiple document summarization,
- Summarization of single document
- Follow up of news articles
- Background of a particular news article
- Breaking news based on twitter
- Sentimental analysis of news articles
- Recommendation of news articles

3 Methodology

Syntactic Similarity

This is Best Matching model which a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It was created via many experiments on variations of the probabilistic model. It ranks a set of documents based on the query terms appearing in each document. One of the most prominent instantiations of the function is as follows. Given a query Q, containing keywords b q1,q2...qn, the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \tag{1}$$

In this application the idf factor is not considered as all the documents retrieved have the relevant terms, so considering it would reduce the retrieval of good documents.

Semantic Similarity

The nouns as well as the synonyms of all terms in the document are extracted and the respective queries are generated. Synonyms of terms in the retrieved documents are generated and semantic similarity is done between the original and retrieved documents in order to provide the most semantically matching article as the follow up.

Summarization

Sum basic algorithm is being used to summarize articles. It considers the probability of occurrence of the relevant terms in the document in order to calculate the weight of each sentence. The highest weighted sentences in the document are being displayed as the summary. The number of sentences selected for summarization are limited.

News article extraction

A python news API "webhoseio" is used in order to obtain the news articles based on query. An account registration is required to access token key which allows up to 1000 free query requests per month. Many filters like language of the retrieved news articles, specific website to crawl, posts crawled since "n" number of days (maximum 30 days) can be specified along with the query. A query requests returns specified number of links that redirects to the required news article. Hence we need to find the redirected link from this link. To extract the news content i.e. title and text; we use another python API "newspaper". This API extracts the main news contents effectively from most of the news websites.

3.1 Query based multiple document summarization

BM25 model was used for matching of the document with queries in order to obtain relevant documents. Webhoseio was used to scrape articles for a particular query but as the results obtained weren't that good, BM25 model was used to retrieve the top relevant documents and multi document summarization was applied to the top results by considering the frequencies of the words in sentences to find the important sentences from among the relevant documents which is displayed as summaries.

3.2 Summarization of single document

A document is given as input and this is summarized by considering the frequencies of the words in sentences to find only the important sentences of that document which is displayed as summary.

3.3 Follow up of news articles

WordNet is being used to generate queries for the starred documents by the use of synsets. These queries are then used to retrieve documents by the use of webhoseio . From the resulting documents, queries are generated for each document by using the same procedure as above. After obtaining queries for each document, JCN similarity measure is used between the query of the starred document and the newly created queries and the documents corresponding to the most semantically similar queries are provided as top results. This is triggered once in few hours to keep the user updated about those topics.

3.4 Background of a particular news article

Use of the webhoseio API to obtain the background of the article based on its timestamp. Basically, all the events related to an article that have occurred over the past month are retrieved and provided to the user as the background of that particular article.read more

3.5 Recommendation of news articles

The nouns and names of places are extracted as tags from the user's starred documents and used as queries to webhoseio to obtain relevant documents for each query using the BM25 model. Then all the extracted tags are combined into a single query and BM25 model is applied on this combined query with the retrieved documents out of which the top few are retrieved and are recommended to the user.

3.6 Sentimental analysis of news articles

A dictionary of all words along with their weights, which may be positive or negative, is stored in a file. This is used to categorize a given news article as being positive, negative or neutral. Sometimes, users may be interested in different type of news like positive or negative based on his mood, hence, the need of this feature.

3.7 Breaking News Based on Twitter

Twitter has evolved from a micro-blogging social network to become, among other things, an essential source of breaking news. In this project, we limit our coverage of breaking news to India by following two news-centric channels viz. @ANInews and @PTInews on Twitter. Based on the tweet which user wishes to know about, tags (specifically noun phrases) are extracted from this tweet. These tags are then queried using webhoseio to fetch related articles. These fetched articles are compared with the complete original query using BM25 algorithm resulting on three based articles related to the tweet. A summary of breaking tweet is displayed using summarization of multiple documents (three articles).

4 Implementation

4.1 Work Done

Experimental setup of the project is mentioned below.

Operating System	Windows 8.1
Processor	Intel(R) Core(TM) i5-4210U CPU @2.40GHz
RAM	8.00 GB
System Type	64-bit OS, x64-based processor
Tool	Python 3.5.1, JetBrains PyChram IDE, NetBeans IDE

An interactive news recommendation system is implemented in this project using Java GUI. A separate window for each of the features described in objectives, viz. Query based News Summarization, Summarization of single news article, Follow up of news article, Background of a news article, Breaking News based on Twitter, Sentimental Analysis and recommendation of news articles is displayed.

4.2 Results and Analysis

4.2.1 Results

Figure 1-8 displays all the objectives implemented in this project on a Java GUI window.

Figure 1: Query based summarization

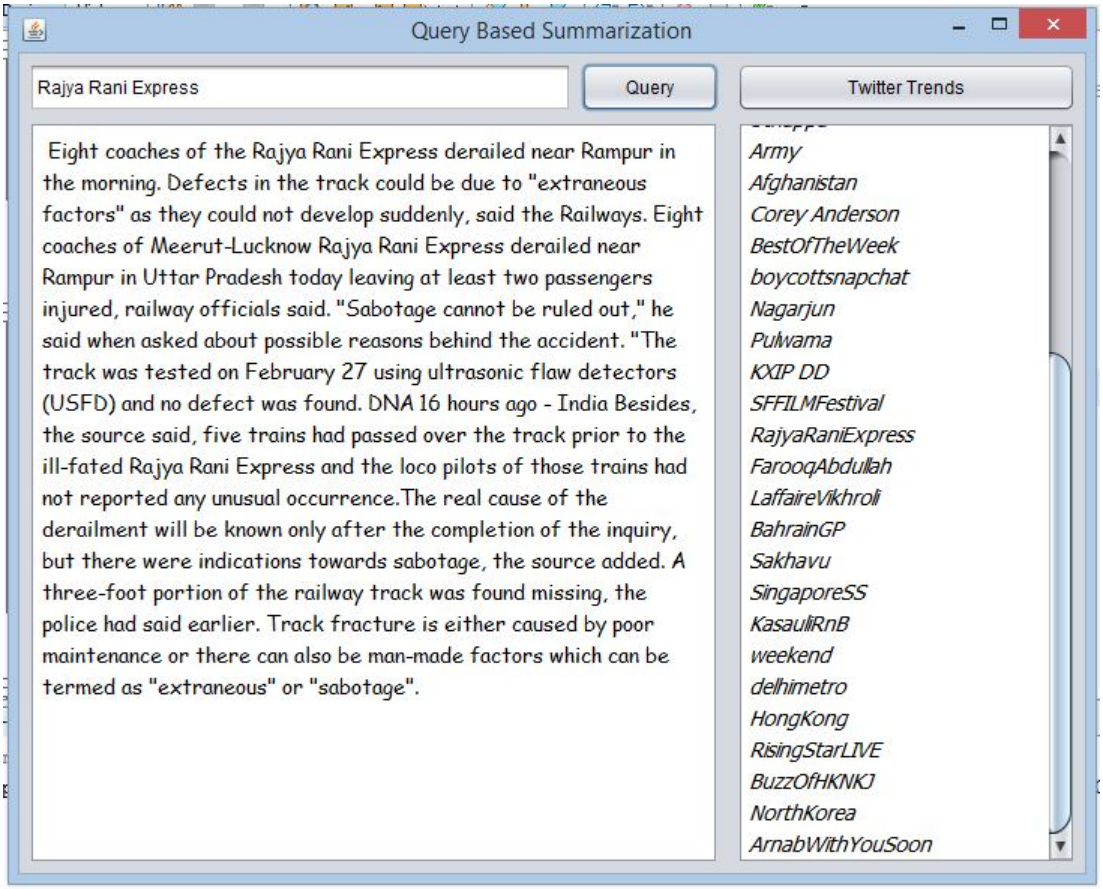


Figure 2: Query based summarization based on Twitter trends

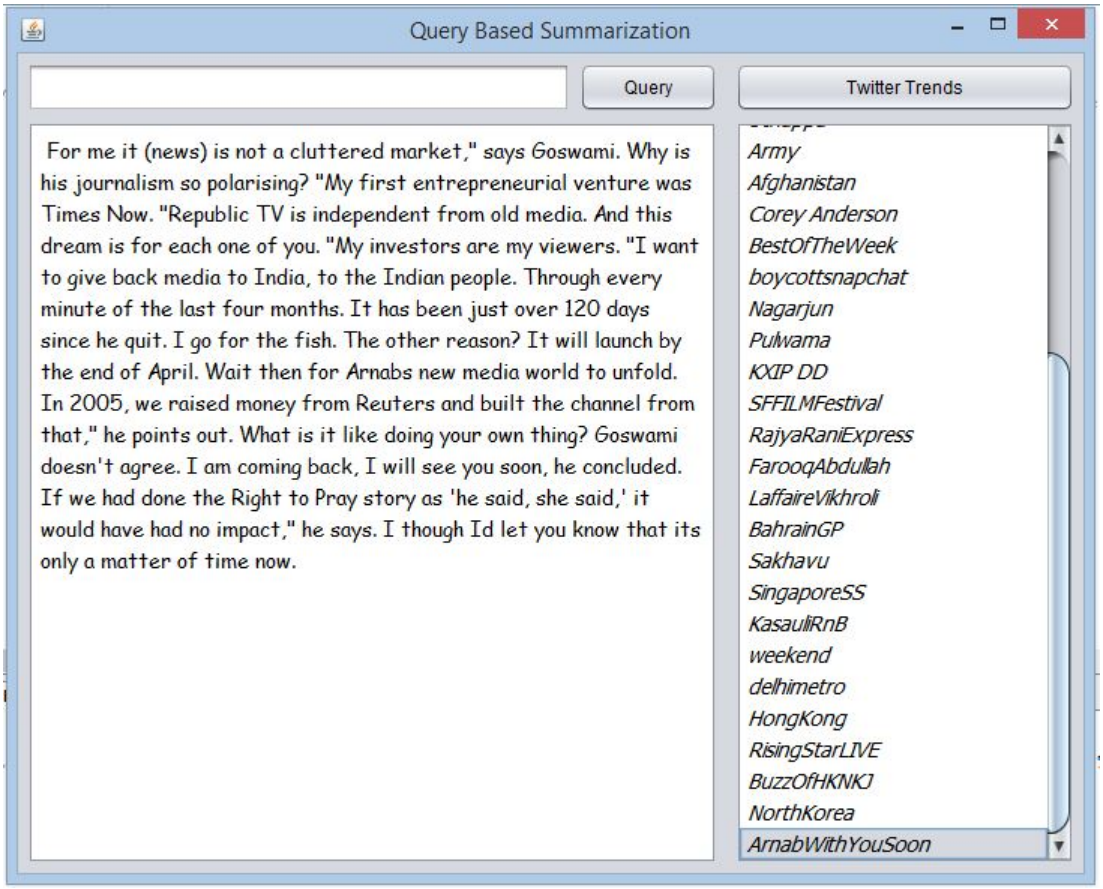


Figure 3: Single article summarization

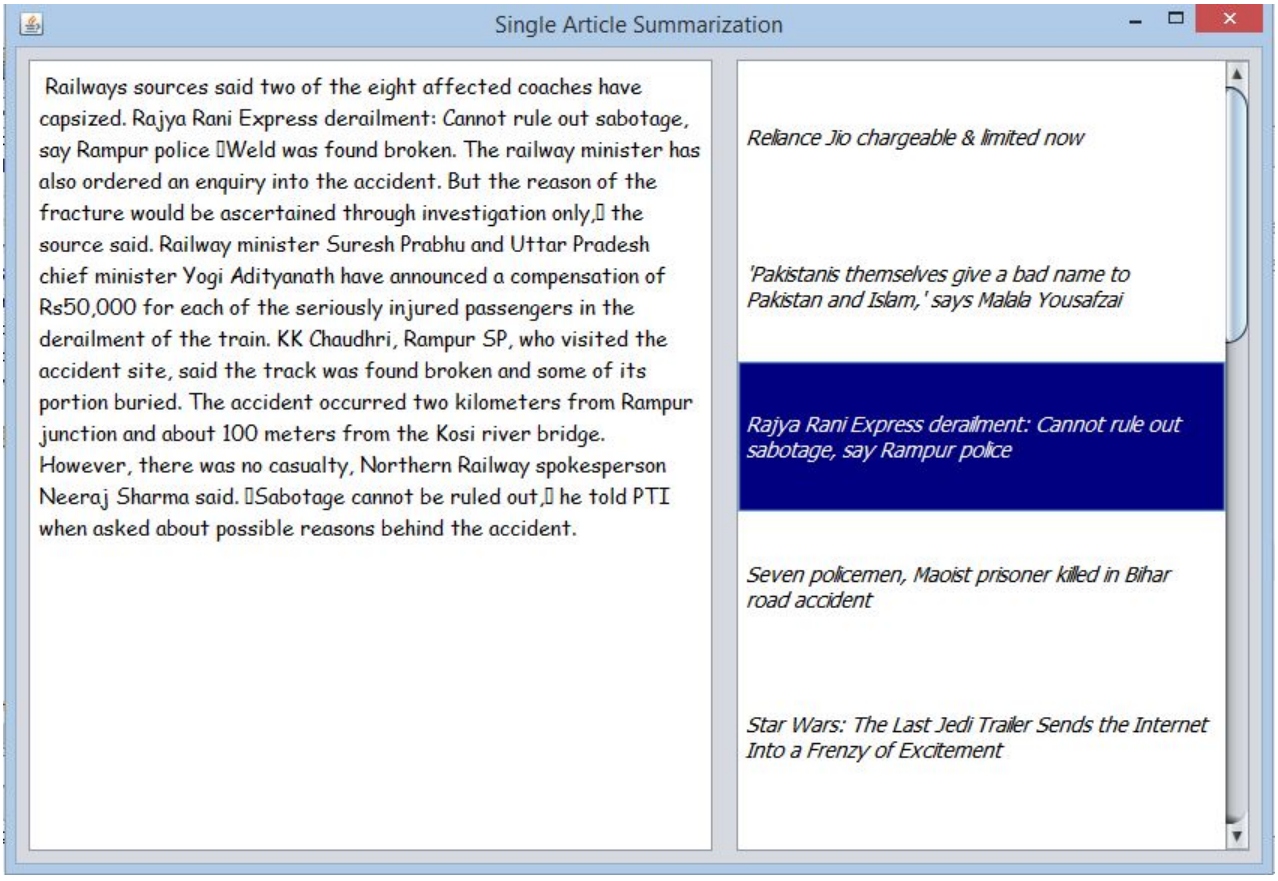


Figure 4: Article recommendation

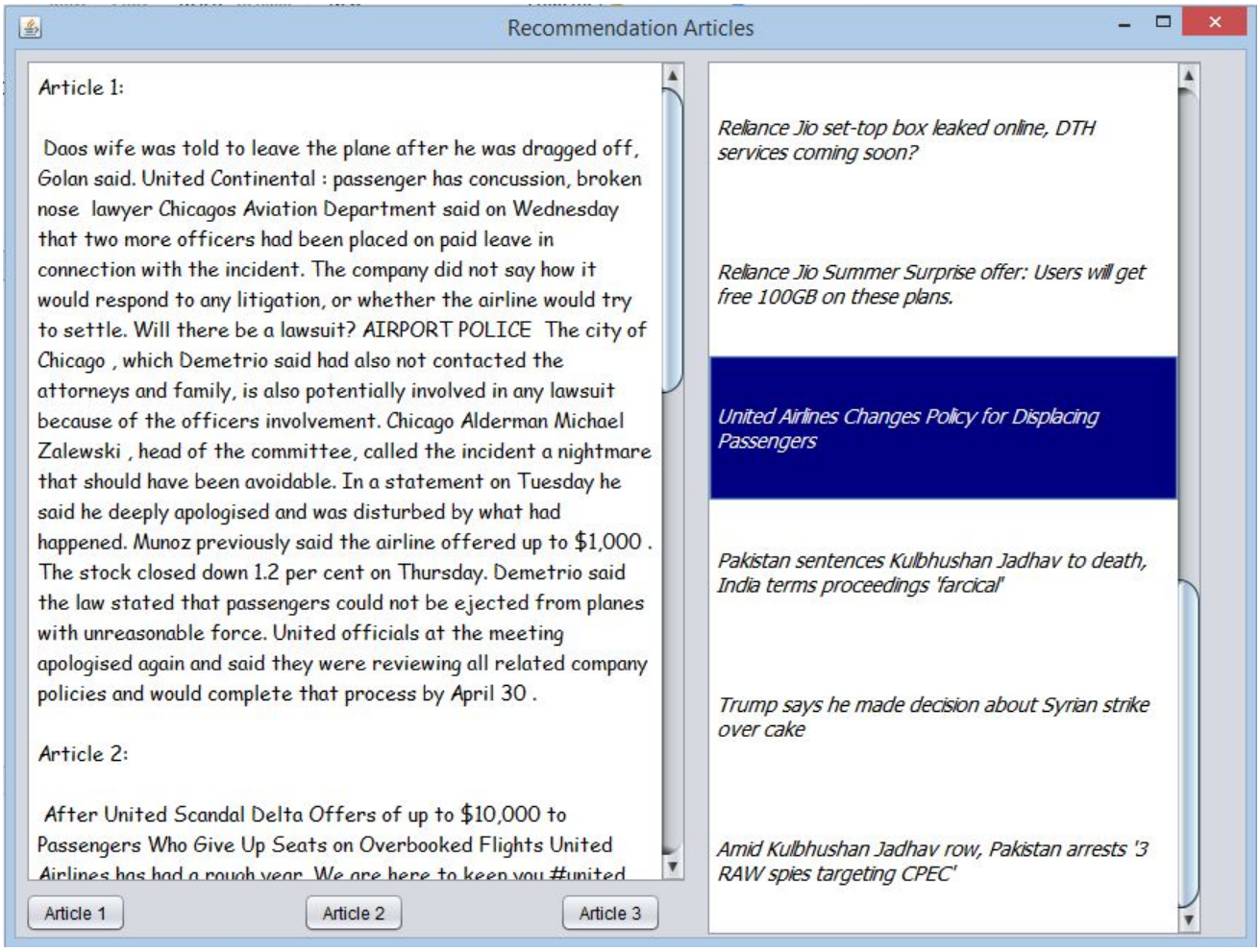


Figure 5: Simple sentiment analysis of articles

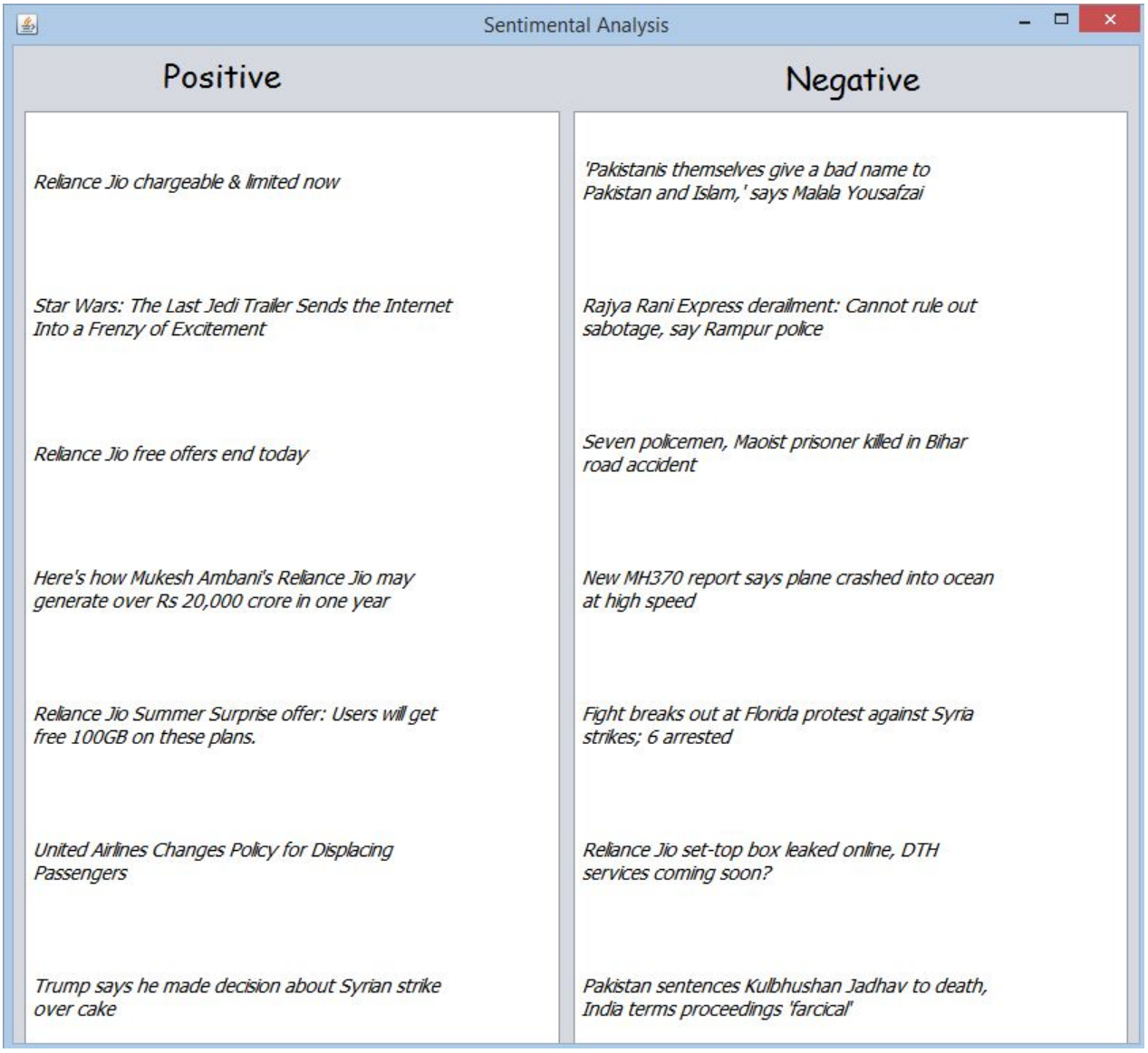


Figure 6: Background articles based on query

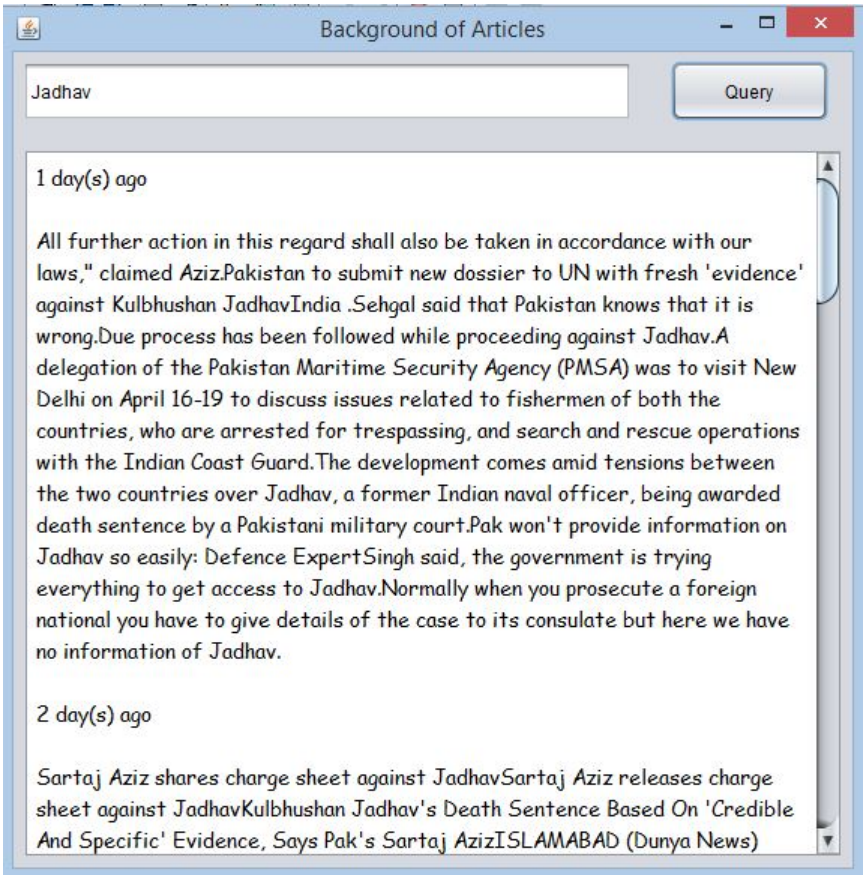


Figure 7: Breaking news based on Twitter

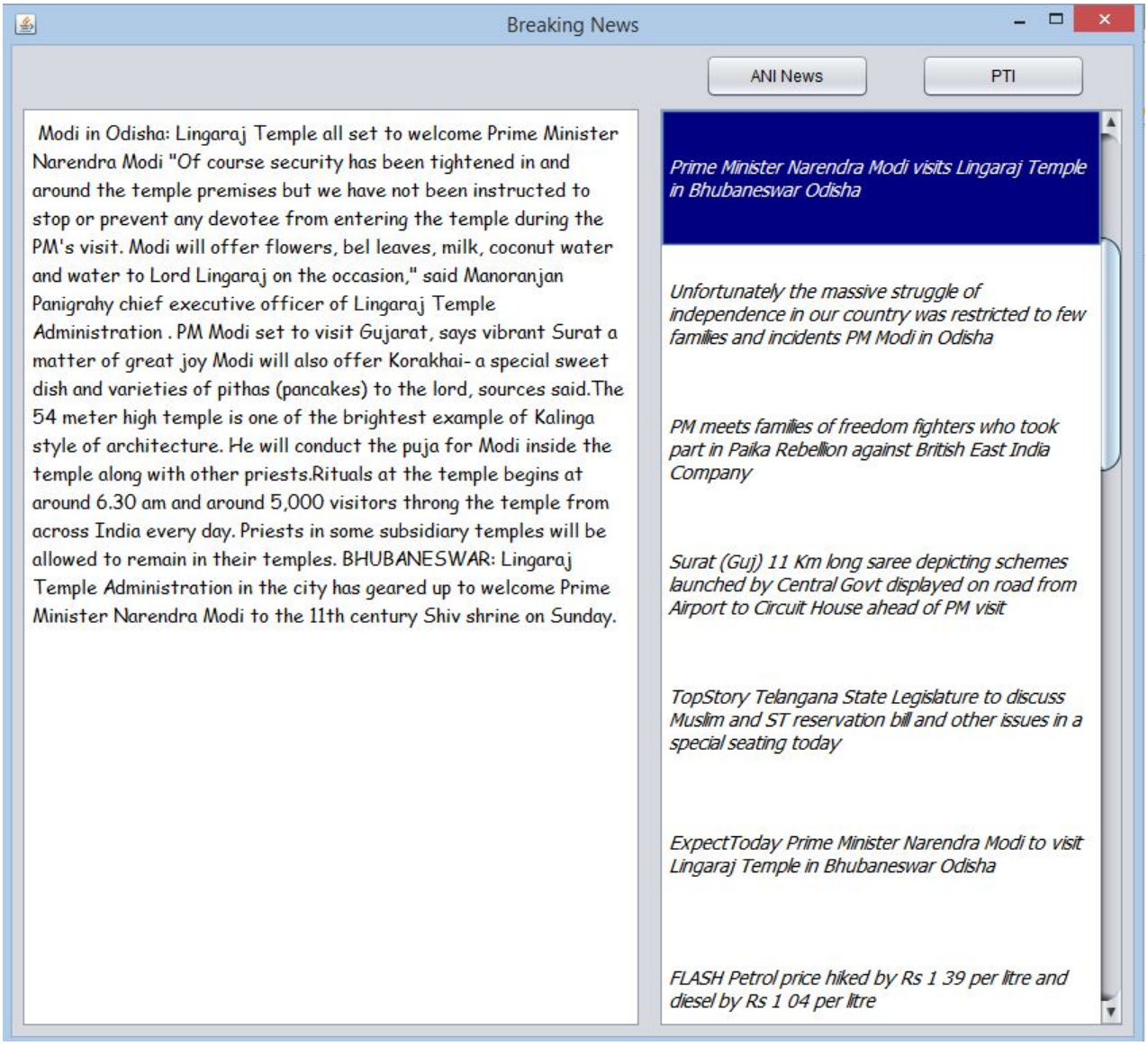
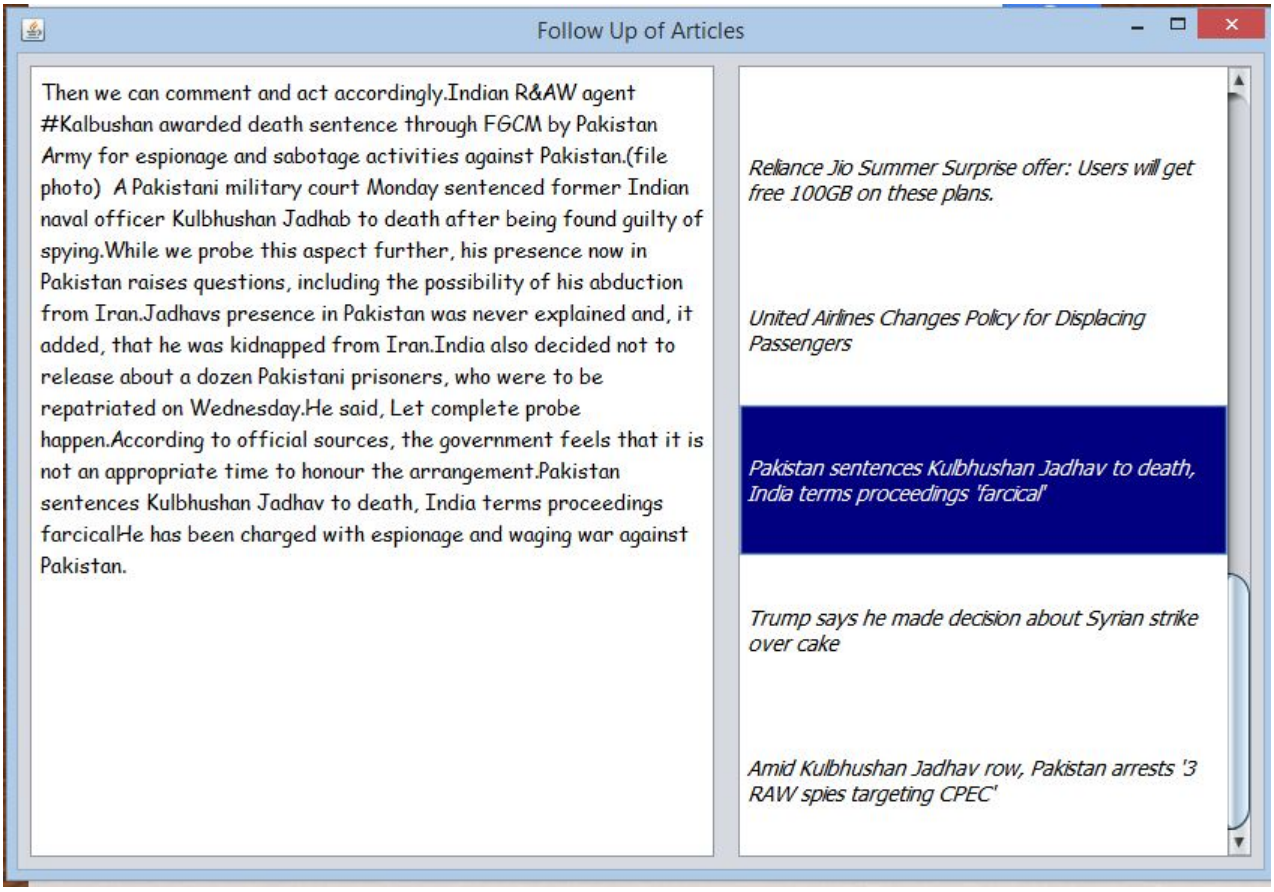


Figure 8: Follow up of an article



4.2.2 Analysis

We tested each of our objectives for different trending topics. Query based summarization works best when user is able to specify exact query related to trending topic. Also querying based on twitter trends is reliable only if the trend is an English word and makes complete sense in itself. Follow up and Background of an article is based on how dynamic the topic is. Recommendation of articles deals with recommending articles from different websites and providing a summary so the user doesn't have to read all the related articles. A simple sentimental analysis works well on the basis of positive and negative terms in the article. Breaking news based on Twitter completely relies on dynamicity of the followed twitter handles, viz. @ANInews and @PTInews. Tweets that are very recent might not have many related articles and hence our system might not give related news articles.

4.3 Individual Work Done

Figure 9: Individual contribution by Sai Harini

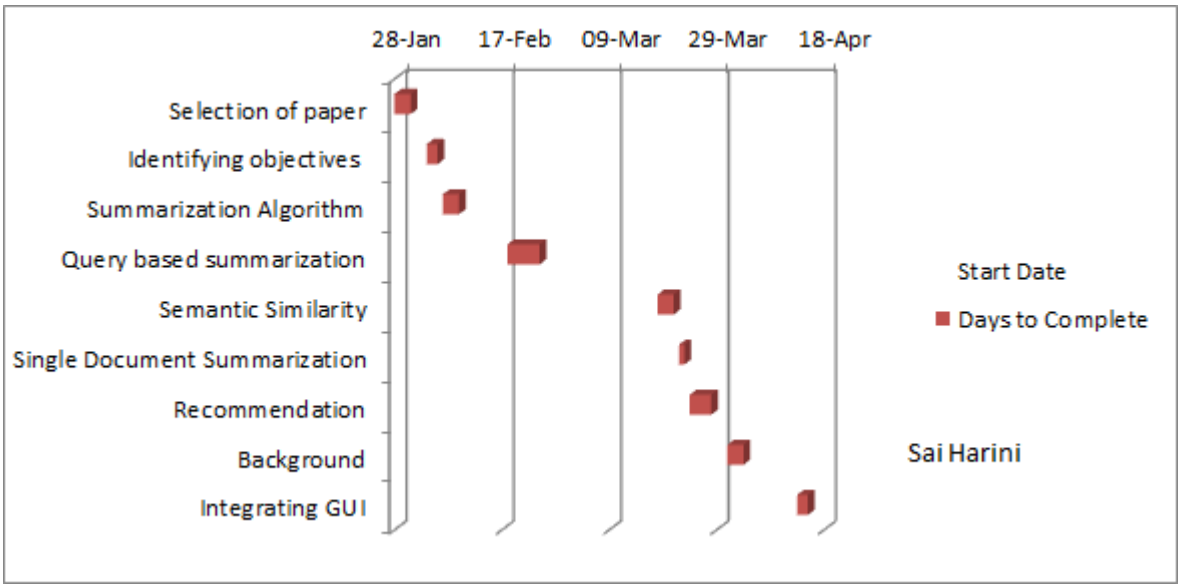


Figure 10: Individual contribution by Jaju Krishna Bhagwan

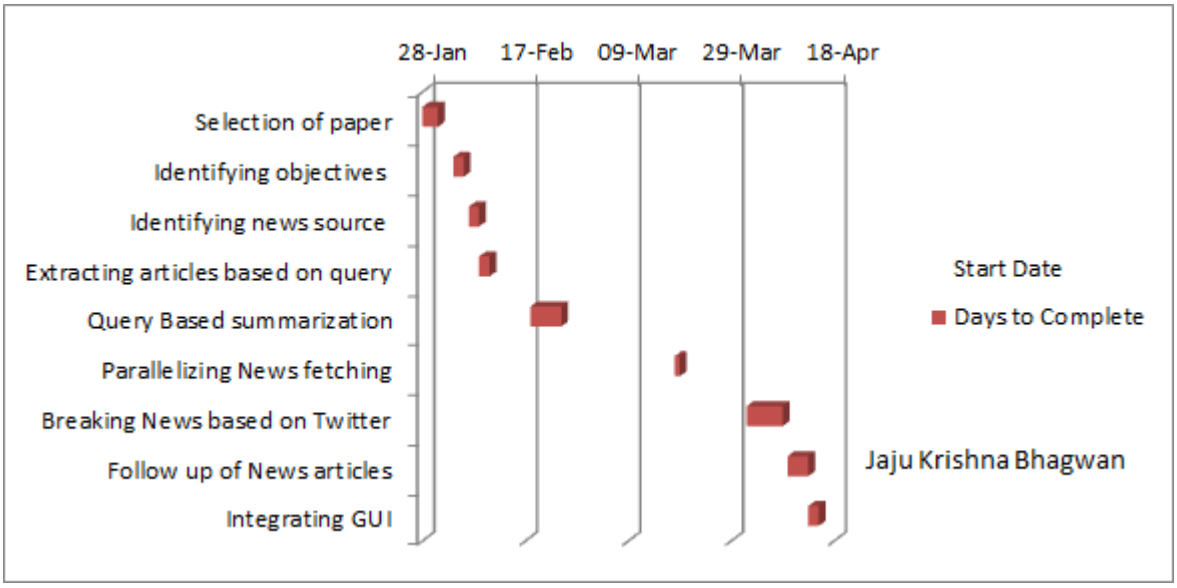
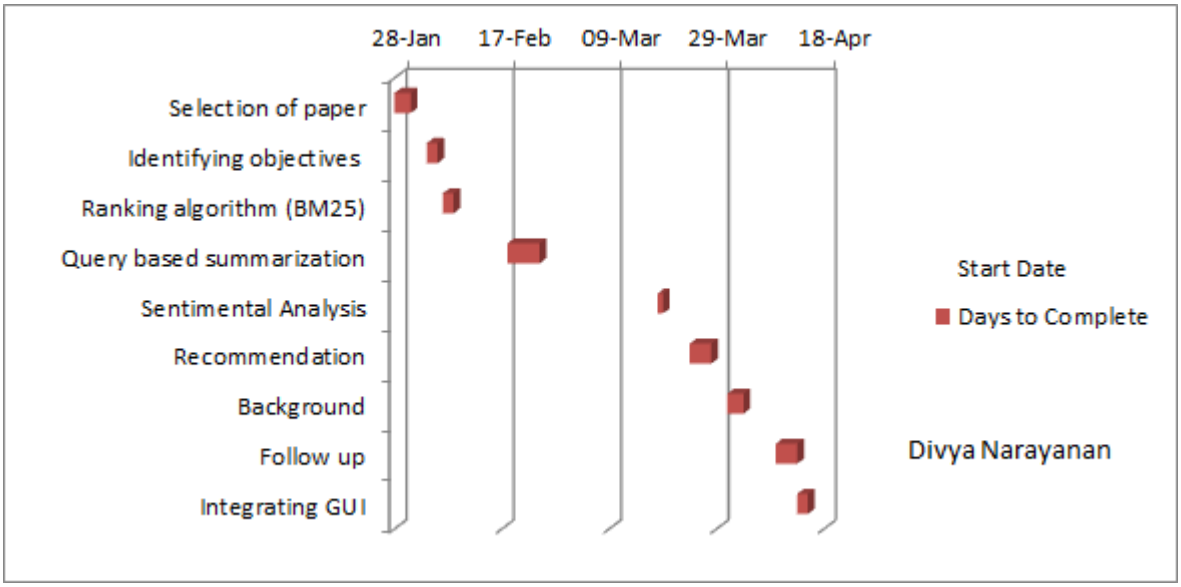


Figure 11: Individual contribution by Divya Narayanan



5 Conclusion and Future Work

5.1 Conclusion

This paper is aimed at providing an enriching News reading experience to the user by providing multiple features on the same platform.A unique feature viz. Query based News summarization feature is proposed. Other unique features like Follow up, Background, and Breaking News based on Twitter have also been implemented. Other features like recommendation of news articles, sentimental analysis, and single news article summarization are proposed using syntactic similarity and summarization algorithm. By providing these multiple features on a single platform, we provide a one-stop application for News readers.

5.2 Future Work

5.2.1 Android Application

The very next step in our project would be to develop an Android application instead of Java GUI that will interact with the python server and provide best user experience.

5.2.2 Using multiple servers

Since downloading an article, extracting its contents, extracting tags, finding similarities, ranking the articles, etc. takes time, multiple servers could be used that will work in a pipeline process.

5.2.3 Removal of duplicates

A near-duplicate removal algorithm to remove duplicated news articles could be used to reduce the working dataset.

5.2.4 Fact-check Tool

As “fake news” is becoming such a popular topic, a fact checking tool based on authenticity of the news websites that endorse the news article can be implemented.

References

[1] Webhoseio API for news articles extraction https://webhose.io/

[2] Chen Li, Zhengtao Jiang (2016). "A Hybrid News Recommendation Algorithm based on User's Browsing Path". Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7550917

[3] Jiahui Liu, Peter Dolan, Elin Rønby Pedersen. "Personalized News Recommendation Based on Click Behavior". Available: https://research.google.com/pubs/archive/35599.pdf

[4] Lihong Li, Wei Chu, John Langford, Robert E. Schapire. "A Contextual-Bandit Approach to Personalized News Article Recommendation". Available: https://www.cs.rutgers.edu/~lihong/pub/Li10Contextual.pdf

[5] Yu Wang, Xinyu Shen, Jinzhi Wang. "Topic Retrieval and Articles Recommendation". Available http://cs229.stanford.edu/proj2016spr/report/050.pdf

[6] Ruisheng Shi, Fan Yang, Lucas Qiu. "WOOF: user defined news recommendation system". Available: sifaka.cs.uiuc.edu/~wang296/Course/IR_Fall/docs/Projects/Samples/4-5.pdf