# A REPORT

# ON

# GENERATING SHORT NEWS STORIES ON TRENDING TOPICS USING TWITTER AND RSS FEEDS

BY

SUGAM GARG 2014A7PS092P<br>
DIVISH DAYAL 2014A7PS132P<br>
A ESHWAR RAM 2014A7PS137P<br>
AMEY AGARWAL 2014A7PS148P

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE,PILANI**

**(NOV,2016)**

# A REPORT
# ON

# GENERATING SHORT NEWS STORIES ON TRENDING TOPICS USING TWITTER AND RSS FEEDS

**Prepared by**

| | | |
|---|---|---|
| **Sugam Garg** | **2014A7PS092P** | **BE(Hons.) Computer Science** |
| **Divish Dayal** | **2014A7PS132P** | **BE(Hons.) Computer Science** |
| **A.Eshwar Ram** | **2014A7PS137P** | **BE(Hons.) Computer Science** |
| **Amey Agarwal** | **2014A7PS148P** | **BE(Hons.) Computer Science** |

**Prepared in fulfilment of**

**Information Retrieval Project**

## AT

## BIRLA INSTITUTE OF TECHNOLOGY

## & SCIENCE, PILANI

**(NOVEMBER, 2016)**

# ABSTRACT

The purpose of this report is to present the project we did under the Information Retrieval Course. The application lets the user set new categories of interest like politics, sports and tech as input. The back-end monitors RSS feeds for a selected topic and ranks the articles using twitter. If an article trends above a certain threshold, the user is sent a summary of story built from the news source article. The need of the project, the methodology used and the final outcome of the project is presented in this report.

# CONTENTS

# 1. PROBLEM STATEMENT

The purpose of this application is to give a list of trending news topics on internet at one place and simultaneously providing users with the links to detailed articles on the topic from different news sources like Hindustan Times, Times of India, NDTV, etc.

# 2. BACKGROUND OF THE PROBLEM

This project is largely made possible by the emergence of Twitter as a standard source for the popularisation of real-time key events. In this particular project, our source of news articles - RSS feeds of various news agencies - gives us a source of continuously updated list of news articles. We now need a way to select articles related to topics trending at a given point of time. For this purpose, we use Twitter trends api to identify the trending topics on the social media platform among the people and with a matching algorithm, we identify and publish the trending news articles on our gui-website.

## 2.1. Motivation of the problem

The avid news-readers who are characteristically passive(most of the population) would need a a) filter for the top trending news topics as well as b) a short summary of the news articles to save time and improve readability/save time. This application, using the concepts of information retrieval, delivers upon these ideas and gives the user the ability to quickly go through the trending topics at any point of time.

## 2.2. Technical issues included in the work

The project uses python/JavaScript coding to implement different modules. The project uses the concept of vectorisation of each doc(news articles/tweets) by creating clusters of news articles as well as tweets(related to a trending topic) and comparing them using cosine similarity of vectors. This way, we get clusters of news articles that are trending and hence report these articles to the gui. The project uses MongoDB database to store data and node(a JavaScript framework) for the GUI.

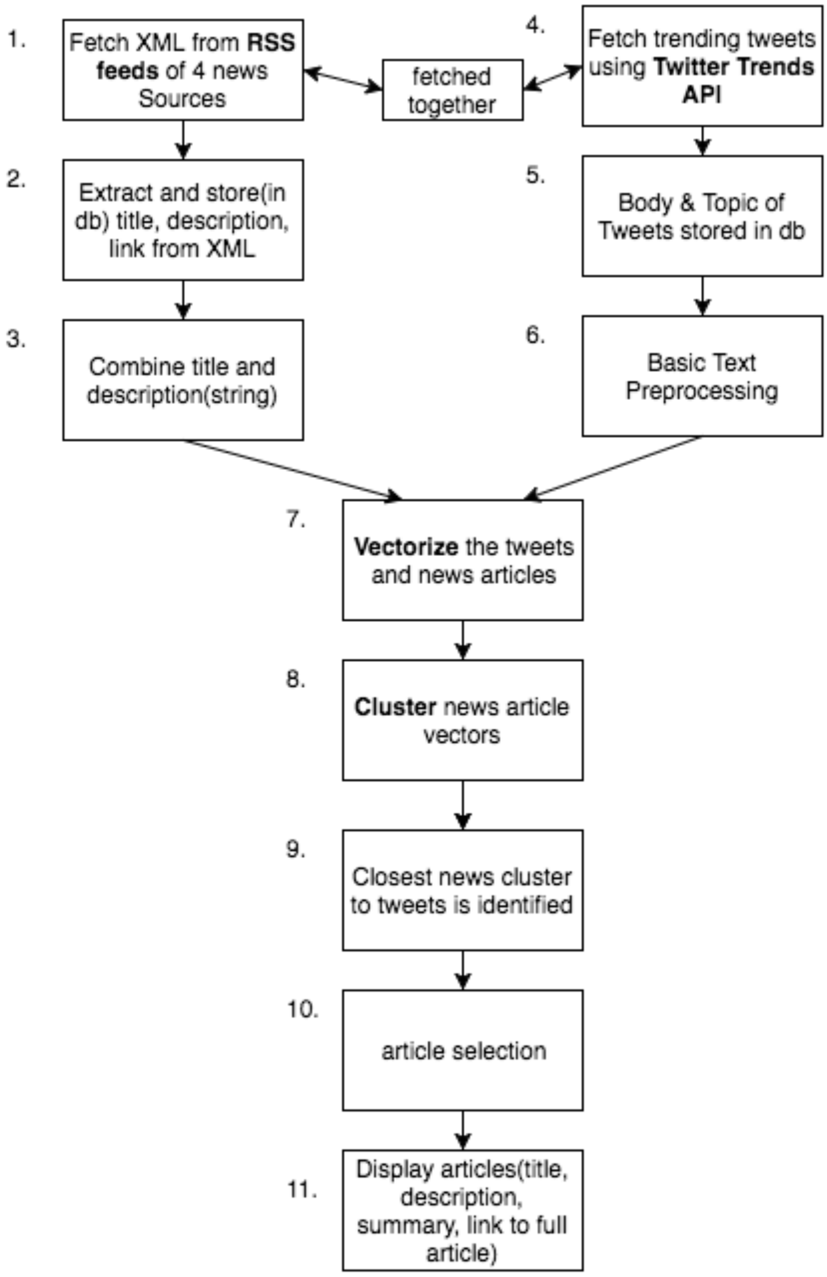
# 3. Literature Survey

## 3.1. Relevant research papers

We first look at existing work on the use of Twitter as a real-time event identification domain. [1] , [2]  and [3](attached in references in the end) give us a rich body of work in the validation of using Twitter for this purpose. For instance [2] (Li, Lei, Khadiwala and Chang) argue for the usage of Twitter as a resource for detecting and analyzing events based on the fact that tweets are created in real time and that tweets have a broad coverage of events with millions of users of all kinds tweeting about information they possess. This enables twitter to have a large volume of content related to a wide range of topics. More specifically, we also draw upon existing work that draws parallels between twitter and conventional news reports - the work in [3] (Subasic and Berendt) analyses whether Twitter users create news or re-report news from professional news outlets. They compare the content in twitter feeds and from articles in professional news outlets, Reuters and Associated Press and conclude that their results suggest that twitter users extend news by commenting upon it rather than create news or re-report it. Importantly, their work also establishes the similarity of tweets to the headlines of news articles- a correlation that is central to the principle correctness of our methodology as we do this exact comparison to filter news articles. We also draw upon the work in [4]  which looks at whether Twitter can be a substitute for newswire - although we are not interested in their central question, their results relating to the time lag between twitter and news article content being updated is of interest to us. By comparing feed from Twitter Streaming API and various news sources- BBC, CNN, Google News, New York Times,Guardian, Reuters, The Register, and Wired - for a time period of 77 days, they conclude that there is no significant lead either source has in content getting updated in context of major news events. With this result in hand , we need not worry about a lag in information getting updated to either source and hence are justified in using Twitter as a event identification source to select news articles.

# 4.System Description

## Block Diagram:


1.
Fetch XML from **RSS feeds** of 4 news Sources
fetched together
4.
Fetch trending tweets using **Twitter Trends API**
2.
Extract and store(in db) title, description, link from XML
5.
Body & Topic of Tweets stored in db
3.
Combine title and description(string)
6.
Basic Text Preprocessing
7.
**Vectorize** the tweets and news articles
8.
**Cluster** news article vectors
9.
Closest news cluster to tweets is identified
10.
article selection
11.
Display articles(title, description, summary, link to full article)

## Block Description:

4.1. This module fetches XML from the RSS feeds of the four news source websites we are using in this project -  NDTV, Hindustan Times, Times of India and Rediff. The fetching procedure “newsfetcher” is defined in the file “news.js” in folder “js”. We use the NPM package parse-rss for this procedure.

4.2. In this module, we extract the title, description, link and created-on-date for each news article from the XML and store these attributes in our database. Throughout the project, we use MongoDB databases to store the results of each step of the process. The schema for news article objects is specified in the “news.js” file in the models folder. The link to the common MongoDB database is specified in the driver file “index.js” in folder “js”.

4.3. In this module, we combine the title and description strings to form the main comparison string for each article fetched from the MongoDB database.

4.4. This module extracts trends and tweets using the Twitter trends API . The procedures for extraction are defined in “tweets.js” and “trends.js” in file “js” .

4.5. This module stores tweets and trends extracted using the procedures of the previous module in the MongoDB databse using the logical schema for trends and tweets for MongoDB stored in “tweets.js” and “trends.js” in file “js/models”. The link to the common MongoDB database is specified in the driver file “index.js” in folder “js”.

NOTE: The driver file “index.js” in folder “js” calls the fetching procedures defined in modules 1 and 4 and the storing procedure in modules 2 and 5. The fetching and storing is done in parallel by the javascript modules.

4.6. This module defines the function for the preprocessing on the body of all tweets fetched from the MongoDB database. URLs and emoticons are removed from the body of tweets while hashtags and twitter handles are split based on camel casing. Words in the tweet bodies are also spell checked and corrected whenever not found in the English Dictionary. To prevent the wrong “correction” of proper nouns that are relevant to a trend, we maintain a frequency table of all words in our collection of tweets and allow spell correction of words only if they occur below a certain frequency. The tweet bodies of each tweets is replaced with the processed content and updated in the MongoDB database.

This function is defined in the file “prepro.py” in folder “py”. The module tweet-preprocessor is used to remove URLs and emoticons (alternatively also dealt with by the spell check function which replaces all emoticons with the same alphabet). The module regex is called to split hashtags and twitter handles based on a camelcasing splitting regular expression. We also define a spell checking and correcting function “replace” that uses the enchant module of python (which itself uses Aspell) and edit distance from the python module nltk. The module nltk is also used for tokenization and to build the frequency table of all words.

4.7. This module stems and vectorized the news article descriptors , tweet bodies and trend aggregates and stores them back in the MongoDB database. This is the done in the TF-IDF vector space. The stemming is done based on Porter’s Algorithm[5].

This module is implemented in the python file “vectorizer.py” in folder “py”- the TF-IDF Vectorizing Function is imported from the scikit-learn module of python. All news articles fetched from the MongoDB databases have their descriptors (from Module 3) collected together. These descriptors are stemmed using the Porter’s Algorithm based stemmer imported from nltk. All tweets are fetched from the MongoDB database and have their bodies ( preprocessed in module 6) collected and stemmed using the nltk stemmer. These two collections are then vectorized in the TF-IDF space together. These vectors are updated in the MongoDB database.

For each trend, the vectors of all the tweets pertaining in that trend are aggregated to give a trends vector . These vectors are also updated in the MongoDB database.

4.8. In this module, we cluster news article objects based on the vectors obtained from the previous module. This is done to group news articles on similar topics together for ease of identification by comparing to trends from twitter. We use cosine distances of vectors for the clustering. The clustering algorithm is the K- Means Clustering algorithm. The results of the clustering – that is the cluster ID of each article is updated and stored back in the MongoDB database.
This module is implemented in the python file “clusture.py” in the folder “py”- the function is “clustre_news”.The K-Means Clustering algorithm is imported from the scikit-learn module of python.

4.9. This module identifies the closest cluster of news articles for each trend aggregate vector fetched from the MongoDB database. For each trend, the tweet aggregate vector is compared to each cluster's representative vector and the cluster with best cosine similarity is chosen.
This module is implemented in the function “associate_tweet” of the python file “clusture.py” in folder “py”.

4.10. Once the closest cluster is detected by the previous module, all article's are compared using their vector's cosine similarity to the trend aggregate vector. If the similarity is above a certain threshold (set to 0.7) , we note the URL of the article. All the URLs of the selected story are stored in the MongoDB database in a new collection called “stories”.

This module is implemented in the function “associate_tweet” of the python file “clusture.py” in folder “py”.

4.11. The final module scrapes the entire article of selected stories from the previous module , summarized them and outputs to the GUI.

The scraper module is “scraper.js” in folder “js” (called in driver file “index.js”) and the summarizer module is “summarizer.py” in folder “py”. The scraper module uses the node.js module “scraperjs” for scraping while the summarizing module uses the python module “sumy” for summarization.

The GUI file is “index.html”(stored in static folder) which has to be run through node js server. The GUI dynamically updates the stories as and when new trending stories come up through AJAX requests to the server. Each story consists of a title and the brief summary of the story accummulated by various news sources.

NOTE: All the python based modules are called in the driver file “index.py” and run sequentially-

1. Preproc()- Preprocessing Tweets
2. Vectorizer()- Vectorizing News Article descriptors, tweets and trends
3. Clustre()- Clustering News Articles
4. Summarizer() - Summarizing selected stories

NOTE: As a MongoDB database is used as an intermediate storage space for each step, we use the drivers PyMongo for python and Mongoose for Javascript.

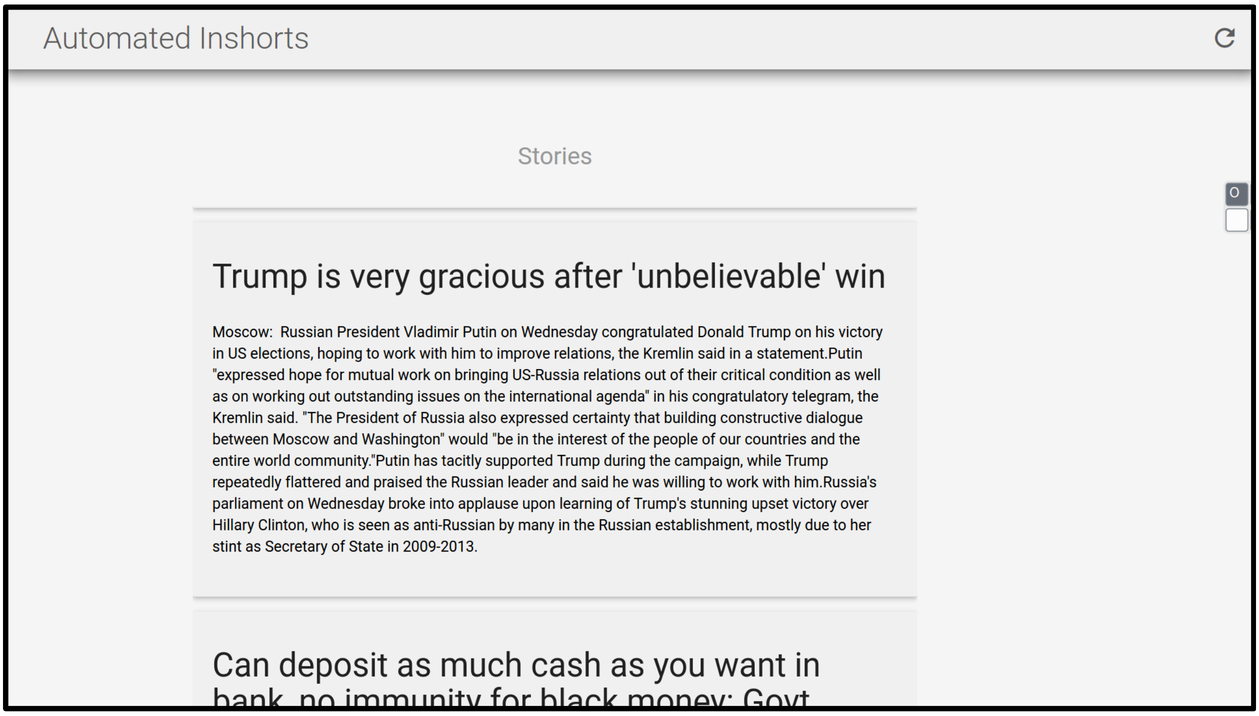
# 5.Evaluation Strategy

The project works on popular trends ongoing on twitter at any given time and the news articles matching with the trend strings. We can tentatively evaluate what the trending topics are at any point of time and then check if the results are related to it. So this makes the evaluation strategy to be mostly subjective and observation based rather than clearly demarcated strategy.

# 6.Experimental Results and Evaluation

The application is ideated in a way that it doesn't require a query to retrieve results from the system. We are giving user trending news which does not require input from the user. The system will give different results based on the news trending at that time.

Screenshots from stories received on 9/11/2016 3:54pm IST :

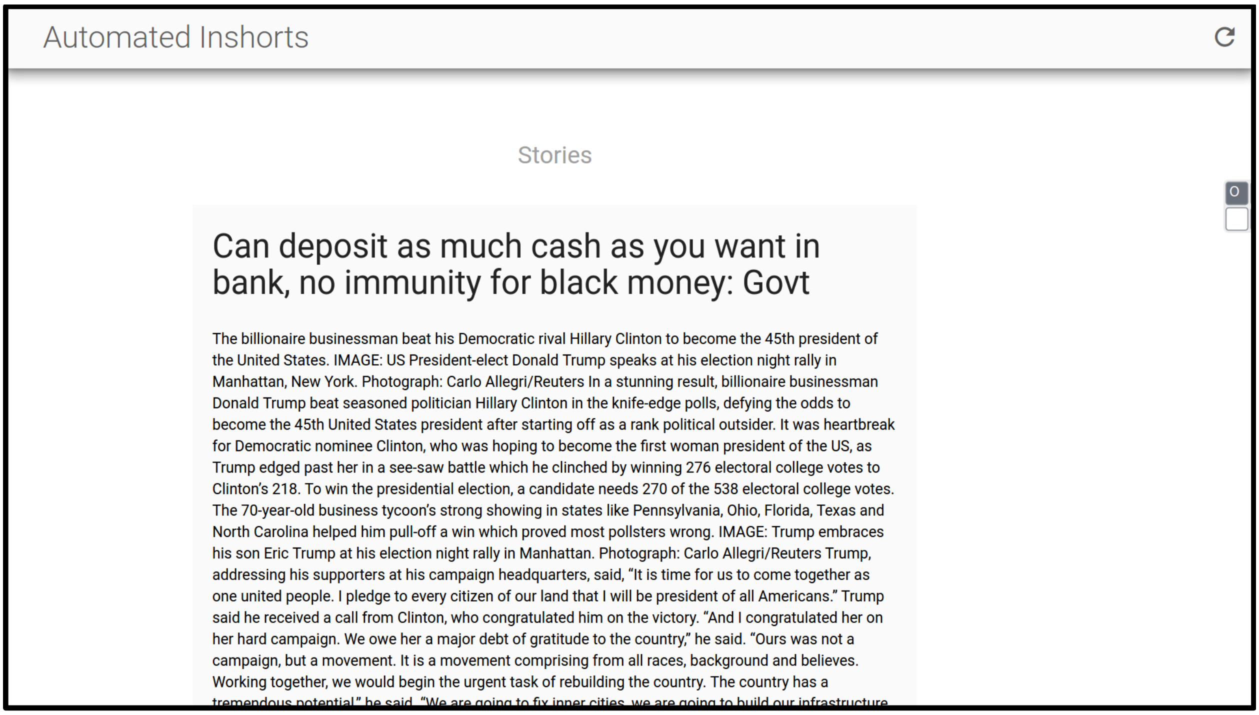


Trump is very gracious after 'unbelievable' win

Moscow:  Russian President Vladimir Putin on Wednesday congratulated Donald Trump on his victory in US elections, hoping to work with him to improve relations, the Kremlin said in a statement.Putin "expressed hope for mutual work on bringing US-Russia relations out of their critical condition as well as on working out outstanding issues on the international agenda" in his congratulatory telegram, the Kremlin said. "The President of Russia also expressed certainty that building constructive dialogue between Moscow and Washington" would "be in the interest of the people of our countries and the entire world community."Putin has tacitly supported Trump during the campaign, while Trump repeatedly flattered and praised the Russian leader and said he was willing to work with him.Russia's parliament on Wednesday broke into applause upon learning of Trump's stunning upset victory over Hillary Clinton, who is seen as anti-Russian by many in the Russian establishment, mostly due to her stint as Secretary of State in 2009-2013.

Can deposit as much cash as you want in bank, no immunity for black money: Govt

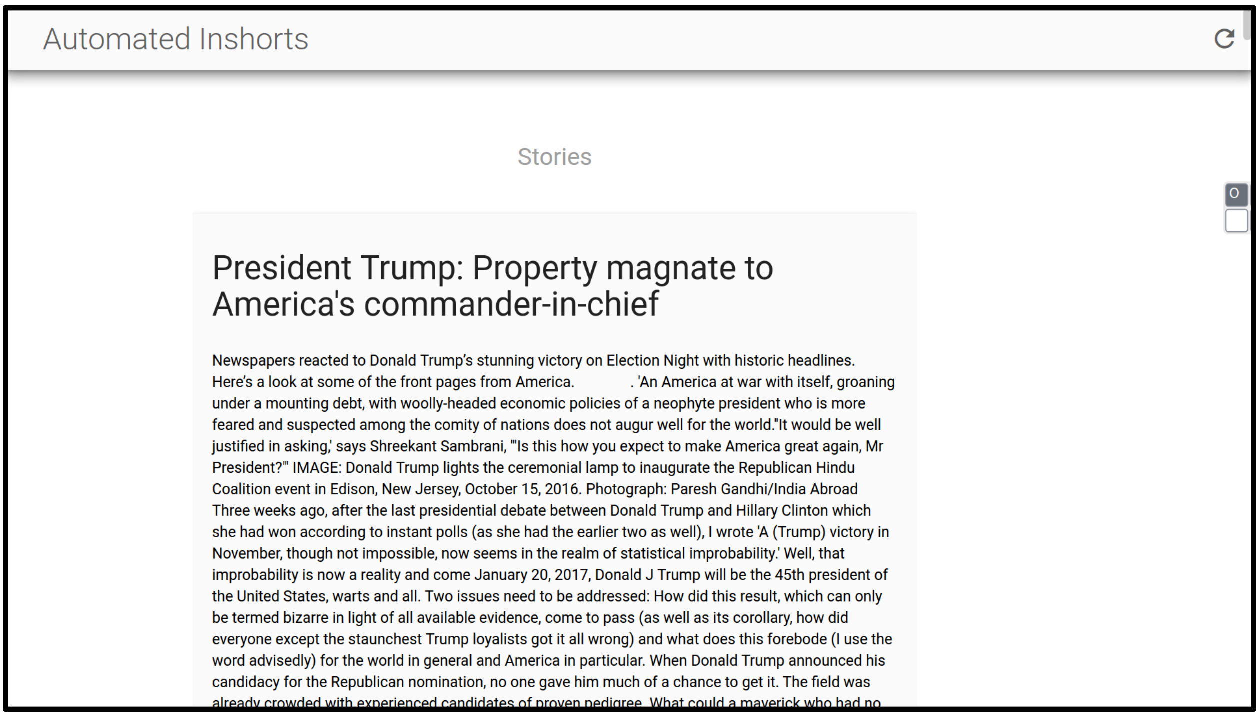
Figure1. Screenshot of the first story retrieved.



Can deposit as much cash as you want in bank, no immunity for black money: Govt

The billionaire businessman beat his Democratic rival Hillary Clinton to become the 45th president of the United States. IMAGE: US President-elect Donald Trump speaks at his election night rally in Manhattan, New York. Photograph: Carlo Allegri/Reuters In a stunning result, billionaire businessman Donald Trump beat seasoned politician Hillary Clinton in the knife-edge polls, defying the odds to become the 45th United States president after starting off as a rank political outsider. It was heartbreak for Democratic nominee Clinton, who was hoping to become the first woman president of the US, as Trump edged past her in a see-saw battle which he clinched by winning 276 electoral college votes to Clinton's 218. To win the presidential election, a candidate needs 270 of the 538 electoral college votes. The 70-year-old business tycoon's strong showing in states like Pennsylvania, Ohio, Florida, Texas and North Carolina helped him pull-off a win which proved most pollsters wrong. IMAGE: Trump embraces his son Eric Trump at his election night rally in Manhattan. Photograph: Carlo Allegri/Reuters Trump, addressing his supporters at his campaign headquarters, said, "It is time for us to come together as one united people. I pledge to every citizen of our land that I will be president of all Americans." Trump said he received a call from Clinton, who congratulated him on the victory. "And I congratulated her on her hard campaign. We owe her a major debt of gratitude to the country," he said. "Ours was not a campaign, but a movement. It is a movement comprising from all races, background and believes. Working together, we would begin the urgent task of rebuilding the country. The country has a tremendous potential," he said. "We are going to fix inner cities, we are going to build our infrastructure

Figure2. Screenshot of the second story retrieved.



President Trump: Property magnate to America's commander-in-chief

Newspapers reacted to Donald Trump’s stunning victory on Election Night with historic headlines. Here’s a look at some of the front pages from America. . 'An America at war with itself, groaning under a mounting debt, with woolly-headed economic policies of a neophyte president who is more feared and suspected among the comity of nations does not augur well for the world."It would be well justified in asking,' says Shreekant Sambrani, "'Is this how you expect to make America great again, Mr President?'" IMAGE: Donald Trump lights the ceremonial lamp to inaugurate the Republican Hindu Coalition event in Edison, New Jersey, October 15, 2016. Photograph: Paresh Gandhi/India Abroad Three weeks ago, after the last presidential debate between Donald Trump and Hillary Clinton which she had won according to instant polls (as she had the earlier two as well), I wrote 'A (Trump) victory in November, though not impossible, now seems in the realm of statistical improbability.' Well, that improbability is now a reality and come January 20, 2017, Donald J Trump will be the 45th president of the United States, warts and all. Two issues need to be addressed: How did this result, which can only be termed bizarre in light of all available evidence, come to pass (as well as its corollary, how did everyone except the staunchest Trump loyalists got it all wrong) and what does this forebode (I use the word advisedly) for the world in general and America in particular. When Donald Trump announced his candidacy for the Republican nomination, no one gave him much of a chance to get it. The field was already crowded with experienced candidates of proven pedigree. What could a maverick who had no

Figure3. Screenshot of the first story retrieved.

The validity of these results can only be subjective since we can’t surely comment on the news trends. From the news trends of 9/11/16 , we can say that Donald Trump’s win/US election and the ban on currency notes were the two most trending news. Our results show these two stories as the most trending news.

Evaluation of the retrieved results :

1. Silhouette Coefficient : The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

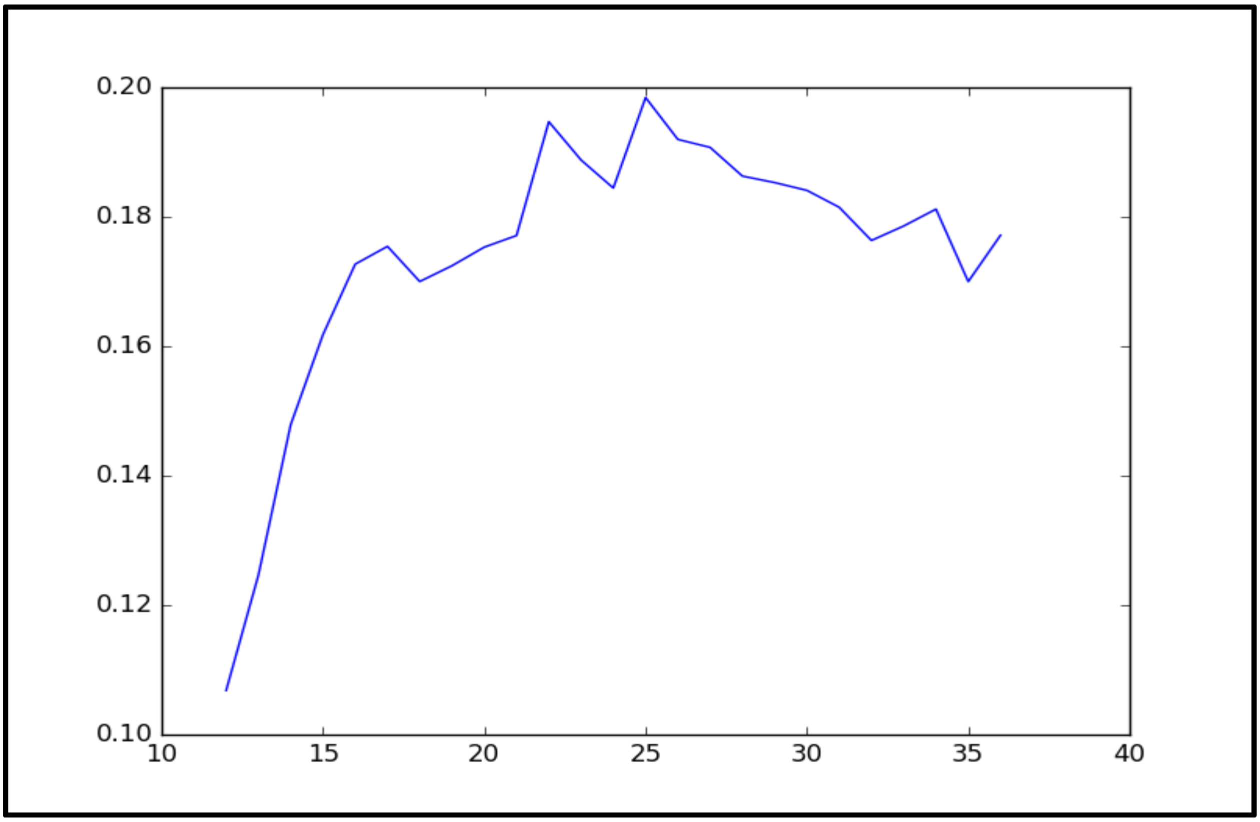


Figure4. The figure shows the graph of the Silhouette Coefficient.

# 7. Conclusion and Future Work

This app will filter the huge chunks of news and display only the news which is trending currently. This will help everyone who can’t find time to read all the news but still want to be updated with the current popular ongoings. At one place, people will be able to find news that everyone is talking about. This app also also summarizes the news saving valuable time of people.

In future, the app can consist of a notification generator, where in people will be updated each time a new news article comes up. The app can also consist of personalized filters where in the app will retrieve news relevant to user’s specific demographic and/or other filters. This app can be made for android and iOS too as the user base increases. We can also include a share option so that user can share the news he is reading with his/her social networking groups.

# References

## Research Papers:

1. Becker, H.; Naaman, M.; and Gravano, L. Beyond trending topics: Real-world event identification on Twitter. In Proc. of WSM , 2011.

2. Li, R.; Lei, K. H.; Khadiwala, R.; and Chang, K. C.-C. TEDAS: A Twitter-based event detection and analysis system.

3.Subaˇsi´c, I., and Berendt, B. Peddling or creating? Investigating the role of Twitter in news reporting. Inf. Retrieval, 2011

4.Petrovic, S., Osborne, M., McReadie, R., MacDonald, C., Ounis , I., Shrimpton, L. :Can Twitter replace Newswire for breaking news? In: ICWSM (2013)

5.An algorithm for suffix stripping, M.F.Porter