A REPORT

ON

PREDICTION OF TRENDING TOPICS IN ONLINE SOCIAL NETWORKS LIKE TWITTER

BY

RAGHAVENDER SAHDEV (2011A7PS257H)
PRANAV KABRA (2011A7PS382H)

UNDER THE SUPERVISION OF

Dr.G GeethaKumari
Computer Science Engineering Department.

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
HYDERABAD CAMPUS

November, 2013.

A REPORT

ON

PREDICTION OF TRENDING TOPICS IN ONLINE SOCIAL NETWORKS LIKE TWITTER

Submitted in partial fulfillment of the
Design Project CS F376

BY

RAGHAVENDER SAHDEV (2011A7PS257H)
PRANAV KABRA (2011A7PS382H)

UNDER THE SUPERVISION OF

Prof. G Geethakumari
Computer Science Engineering Department.

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
HYDERABAD CAMPUS

November, 2013.

CERTIFICATE

This is to certify that the report entitled '**Prediction of trending topics in online social networks like Twitter**' is submitted by **RAGHAVENDER SAHDEV (2011A7PS257H)** and **PRANAV KABRA(2011A7PS382H)** in partial fulfillment of the requirements of CS F376 (Design Project) embodies the work done by them under my supervision.

Signature of the supervisor

Name: G GEETHAKUMARI

Designation: Assistant Professor

Date: 29/11/2013.

ACKNOWLEDGEMENT

A report always requires the goodwill, encouragement, guidance and support of many people. The all-round aspect thinking that an engineer has to have can hardly be gained through books and classes. The exposure to industries, learning and fulfilling their requirements make us feel more confident about our knowledge, and such a learning process is very motivating to keep learning more. Right from the design aspects, protection schemes, calibration and maintenance to troubleshooting are part of our knowledge.

All this wouldn't have been possible without the constant help and support of my supervisor Dr.G Geethakumari, Department of Computer Science Engineering throughout the project. I am deeply indebted to her for giving me the opportunity to work under her supervision.

I would also like to thank Mr. Krishna Kumar and Ms. Agrima Srivastavafor their valuable inputs they kept pouring and seminars held during the entire period of the project. It was their continuous guidance that encouraged us to proceed with the report.

I would also like to thank my collegues Ms. Anupriya Gagneja, Sonakshi Gupta, Aishwarya Srivastava and Srilaxmi for their valuable suggestions on the project..

ABSTRACT

Twitter has become one of the world's favorite social networking hotspots in the past few years. However Twitter, unlike other social media, is generally utilized by people for expressing their opinions/feelings about various topics. The word limit of 140 characters prevents people from unwanted banter. However, it is this word limit that generates an extraordinary potential for information extraction and subsequent analysis.

People talk about various topics on Twitter; ranging from important ones like politics, sports, religion, public policies to the most trivial ones like the ones that fill the gossip columns. This generates vast amounts of readily available data that encompasses public opinion and using the power of a social network, it can be easily tapped to produce meaningful results.

With a greater buzz about a certain topic on Twitter, that topic is likely to become a TRENDING TOPIC. Twitter itself continuously monitors its data and is able to predict what topics are likely to be trending sometime in the future. This is a daunting task, but one worthy of the effort. With the amount of Twitter traffic it may become a massive advertising ground and the same information can be used for various other purposes. Thus the real task is to be able to predict trending topics on Twitter.

The underlying purpose of the project is to solve the above stated problem. Over the years a lot of research has gone into it and different approaches have been used. Twitter itself uses a proprietary algorithm for trending topic detection. However, our approach is novel as we have utilized two different domains to reach the conclusions.

Having extracted plenty of data from Twitter, we firstly utilized the human behavior to help our prediction. The social network graph enables us to determine the behavior of an individual by judging the behavior the persons that he/she may be connected to. The social network finally can be viewed as communities and we have utilized the interaction patterns to be a deciding factor in our prediction.

Also, we have used a non-parametric approach to solve the problem by utilizing the timestamps of the tweets. The volumes of tweets at different time intervals enables us to generate signals and their further pattern matching with sample trending and non-trending is a deciding factor. We have used an efficient mix of the above two approaches to develop a fairly good accuracy for our results.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 GENERAL

In the world today, social media has become one of the most pervasive forms of spread of information. News is everywhere. It affects our lives at home, at work and at play. One of the most fundamental natures of humans is to have an opinion on possibly every matter. The entire essence of a Social Web space like twitter is to capture the sentiments of the people and enable them to post short updates.

1.2 NEED FOR PRESENT STUDY

In recent years, there has been an explosion in the availability of data related to virtually every human endeavour — data that demands to be analysed and turnedinto valuable insights. Twitter is one the hottest platforms for social networking such that it has shown unprecedented growth in the past two to three years. The twitter traffic is so enormous that researchers have mined through the data and extracted extremely crucial and practically relevant data. Such large quantities of data present both opportunities and challenges. On the one hand, enough data can reveal the hidden underlying structure in a process of interest. On the other hand, making computations over so much data at scale is a challenge. Fortunately, in recent years, advances in distributed computing have made it easier than ever to exploit the structure in large amounts of data to do inference at scale. Twitter identified the potential of the buzzing topics and developed a proprietary algorithm that enables them to predict topics that are likely to trend.

Detection, classification, and prediction of events in temporal streams of information are ubiquitous problems in science, engineering and society. From detecting malfunctions in a production plant, to predicting an imminent market crash, to revealing emerging popular topics in a social network, extracting useful information from time-varying data is fundamental for understanding the processes around us and making decisions. This will enable us to predict if any news that is falsely broken would spread out of bounds or if any topic is being forcibly pushed to become a trend due to some spurious activity.

.

1.3 OBJECTIVES

1. To understand the importance of social media in general and review the trends that could spread through suitable literature review.
2. Implementation of test bed to collect real time data.
3. To understand the various algorithms used for analysis of twitter data.
4. To finalize and come up with an algorithm which could possibly predict whether a given topic would become a trend or not.
5. To design and develop a working web based application which could predict trending topics in twitter.

2. LITERATURE REVIEW

One of the most important parts of this project was research papers of highly qualified professors/people who had already done extensive research in the field of social network analysis, machine learning, data mining, etc. The research papers that we had read for a better understanding about the approach of our project can be found in the reference section of this report.

The study of the literature revolving around Twitter began during our semester break. Literature Review basically involved study of Online Social Networks, especially Twitter, to develop a deep understanding about the topic. An online course on Social Network Analytics came in handy. It helped in understanding the vast potential that Twitter offers, not only in terms of the trending topics but also the impact of the inferences that can be drawn from its dynamics. Research papers were the most important in understanding the dynamics of Twitter. We realized that a lot of research effort has been put in to identify avenues where the power of social network can be tapped. Be it sentiment analysis, prediction of trending topics, global news, detection of misinformation; each of the above has been attempted using Twitter traffic.

Following are some of the papers worth mentioning:

Towards More Systematic Twitter Analysis: Metrics for tweeting activities- This paper's major focus was on the power of hash tags. It described how the author had used hash tags as a means for prediction of trends in twitter.

From Obscurity to prominence in minutes – The basic take away from this paper was exact distinction and significance of a retweet, reply, mentions and a simple tweet. Also it throws light on the importance of real time twitter data for search engines.

Measuring User Influence in twitter the million follower fallacy – More number of followers DOES NOT imply that the tweeter will have greater impact can clearly be observed after reading this paper. Influence is neither gained suddenly nor accidently.

Twitter Under crisis – can we trust RT? - Tweets with RT need not always be authentic due to spam/forced RT tweets. Thus the quantum of re tweets for a topic cannot be used as the only metric for predicting trending topics.

Integrating Web based Intelligence retrieval and decision making from the Twitter Trends Knowledge Base- Various techniques have been identified as to why a trending topic becomes so. All were experimental parameters. Method of backtracking is used for prediction analysis

Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series- This research paper was a major breakthrough for us to identify the time based algorithm that we have used for our analysis.

Reading the above mentioned research papers we realized that predicting trending topics cannot be restricted to a single metric. Thus the next task that lay ahead was to identify some appropriate metrics to be able to categorize topics as trending or non-trending.

3. DATA EXTRACTION

Data extraction is one of the most integral parts of the project. There are various methods to extract data from social networking sites. It is required for project that we extract large amount of data and then use this data to predict trends.

Data can be extracted in real time using the Streaming API, which gives us continuous data over a period of days.

In this project we have extracted data from TAGS (Twitter Archiving Google Spreadsheet). When extracting data from this medium, we get a continuous data from twitter in the form of a csv (comma separated file) file, which can be viewed in excel. It has various columns such as from, to, tweet, id, date-time, geo-location, etc. We can then arrange the data in the form of increasing order of the date and time.

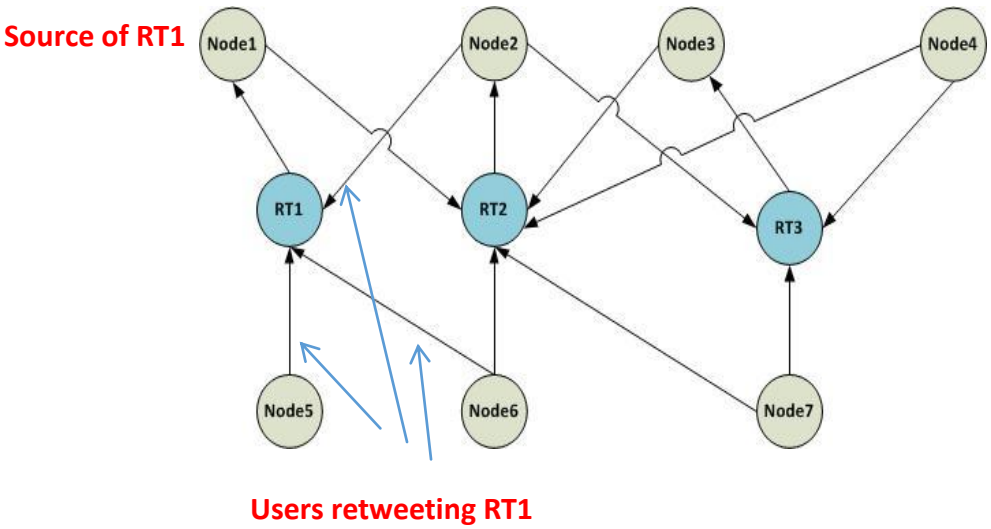
We have extracted data of about 20 topics in the form of csv files. We have then used these files' data as an input for our algorithm and then come up with a hypothesis which could possible predict on a new file whether it would become a trend or not. We have used these files data as a training set for our algorithm.

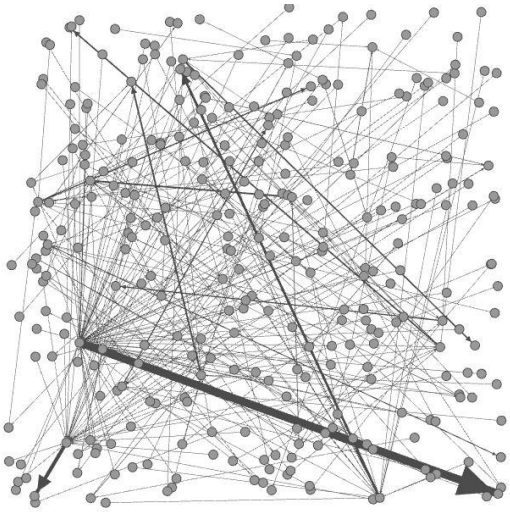
4. ALGORITHMS USED

1. COMMUNITY DETECTION ALGORITHM

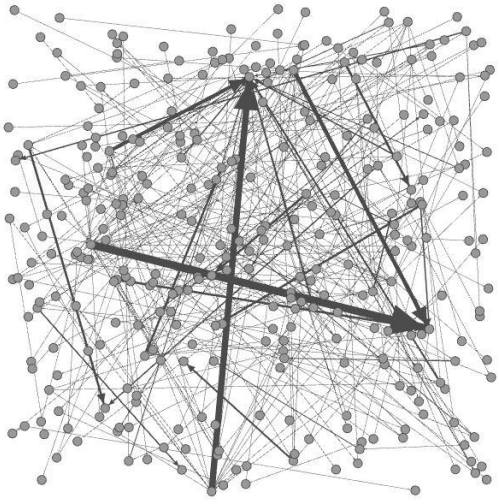
This algorithm is based entirely on the behavior of the users and their tweeting activity and patterns. It is observed that users tend to exhibit limited patterns and thus the variety of their tweets is limited. Rarely do we find people tweeting about a wide variety of topics. This is where we utilized the power of the social network graph to identify and isolate the users that exhibit similar behavior and tastes. Depending on the people the user follows and is himself being followed, we are able to identify groups that can be represented in a way of exhibiting a common behavior and can be viewed as a single community. This greatly reduces the complexity of our analysis by simplifying our graphical networks.

From the data we collected, we used the retweets as a means to identify the required patterns. We extracted the unique retweets and the original user that posted the tweet. Also for each particular retweet, we identified the users that actually retweeted. Using these three different components we devised a graph that essentially depicted the relationship among the original tweeters and the rewteeters through the tweet itself as the node. The same has been depicted in the graph below.



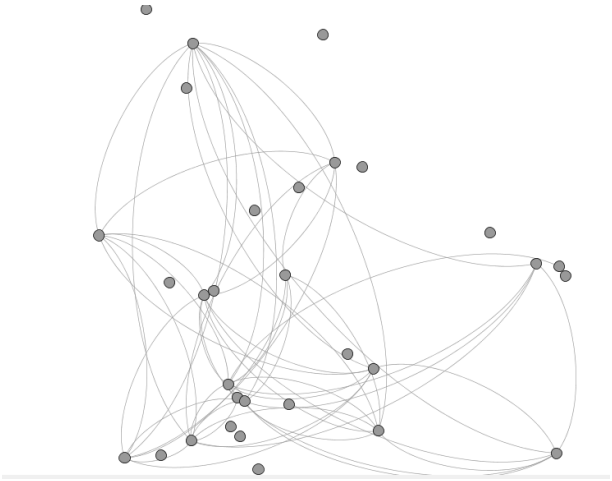


Graph of Trending Topic

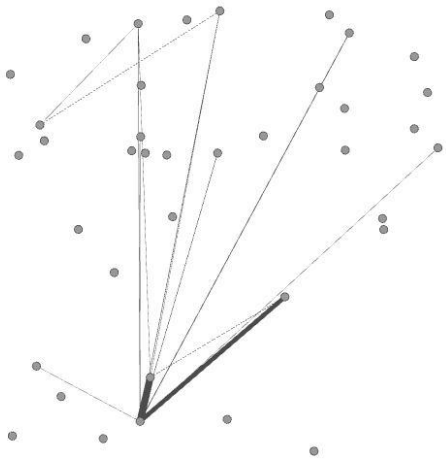


Graph of a non-Trending topic

Once this has graph has been plotted, we observe that it leads to a very complicated pattern even for a small data set. Thus we proceed by simplifying the graph to identify communities. This leads to a much simpler graph such that all the nodes belonging to the community are represented as a single node and the multiple edges are collapsed as a single edge. An edge in this final graph indicates that there is interaction among the two communities.



Community Graph corresponding trending topic

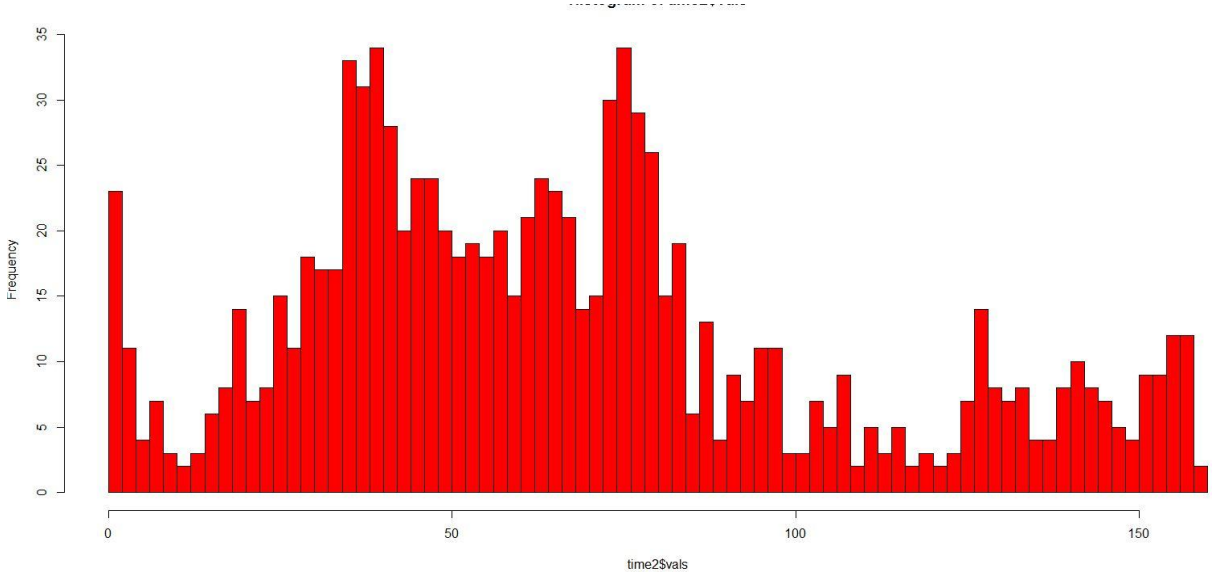


Community graph for the corresponding non-Trending graph

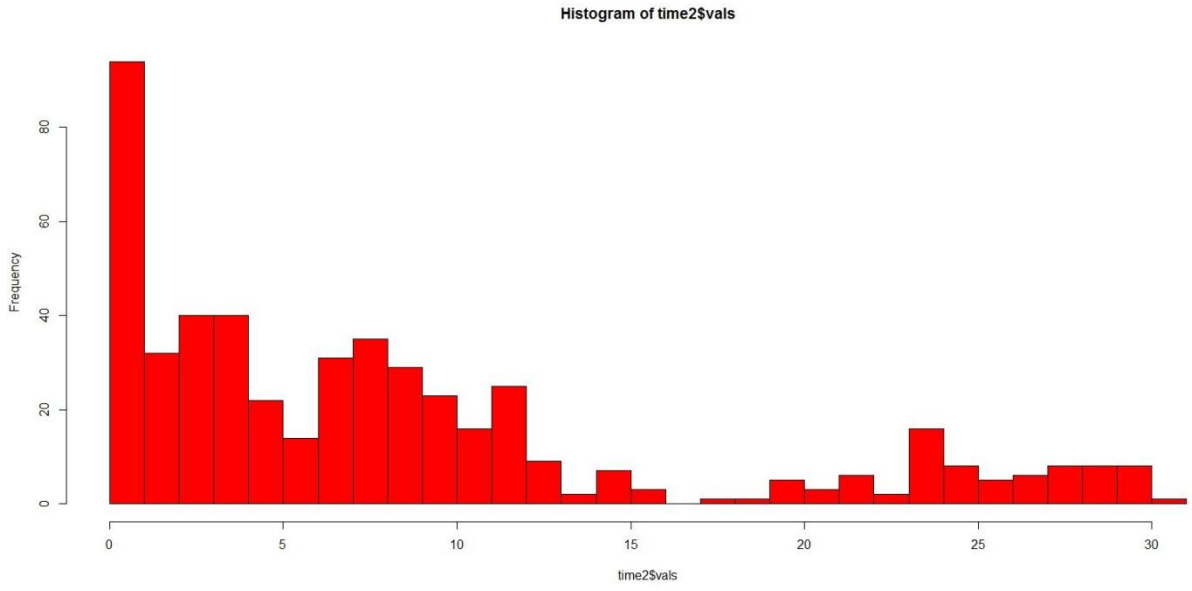
By intuition as well, we can claim that greater the chatter of topic across communities, greater the chances of it spreading across the twitter users and higher the chances of the topic to trend. Due to the lack of any fixed threshold to adjudge a topic as trending, we tested this algorithm against many data sets having previously known the trending/non-trending nature and compared the ratio of the edges to the nodes in the community graphs. It was observed that trending topics generally had a ratio of greater than 1.2 while the non-trending topics had ratios less than 0.25. Results varying in between the two extremes used the results from the second algorithm to ascertain the prediction.

2. TIME-STAMP SIGNAL BASED ALGORITHM

This algorithm has been highly inspired by a research work at Massachusetts Institute of Technology. They have used a novel non-parametric approach to address the issue. They claim that no explicit parameters are required to adjudge the topic as trending or non-trending. We have adopted a similar approach and used the time-stamps of the tweets to find the volumes of tweets for different time intervals. Just like in the MIT research paper, we have created signal of the twitter traffic for trending and non-trending topics. We calculated the volumes of tweets for 2 minute time bins and plotted a continuous graph for it. With the research, we concluded the presence of peaks is extremely essential for a topic to trend. Intuitively, it's obvious that a topic will trend if there is a sharp increase in the discussion. A sort of flat curve will not trend because its activity is sort of stagnant or static. Thus the major goal is to identify peaks in the graphs. However peaks are not simply enough to judge that. We have plotted the Twitter traffic graphs for both trending and non-trending topics. We utilized them as our sample reference signals : +ve signals are the ones which trended while –ve are the ones that didn't. For any given observed signal, we use a sort of regression approach to calculate the distances between the signals at various points and thus judge the proximity of the signal to our reference signal. If observed that our observed signal is closer to that of the positive signal at greater no of points, we conclude that it is likely to be trending. Otherwise its adjudged as a non-trending



Time v/s frequency(2 mins) for trending topic



Time v/s frequency(2 mins) for non-trending topic

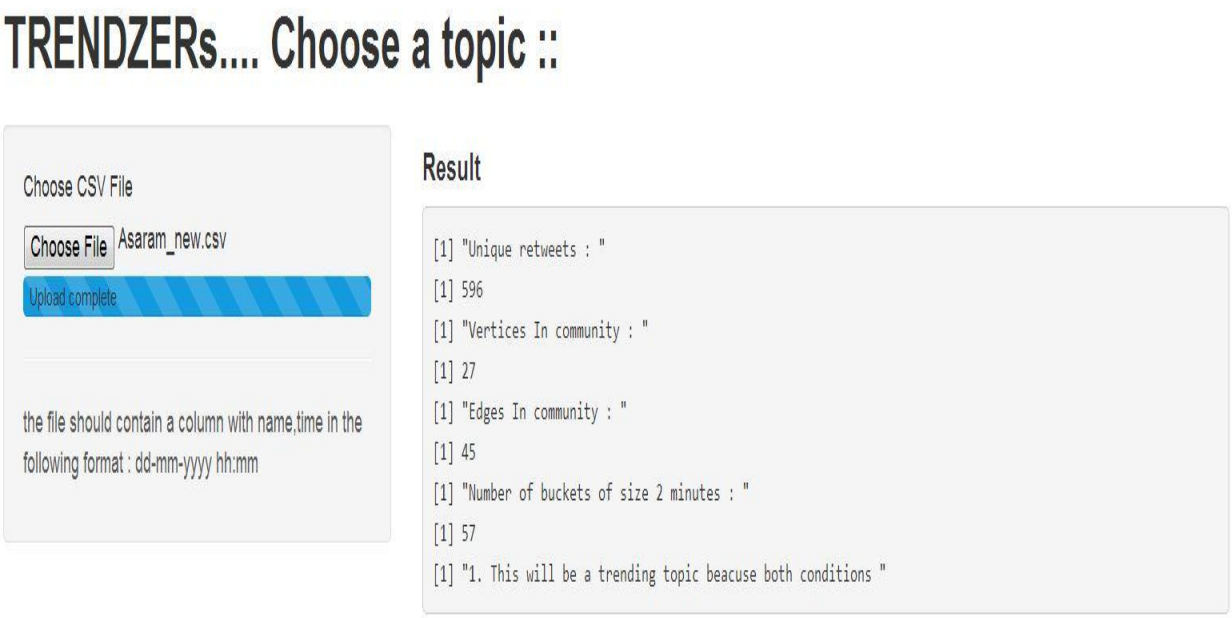
6. FRONT-END AND RESULTS

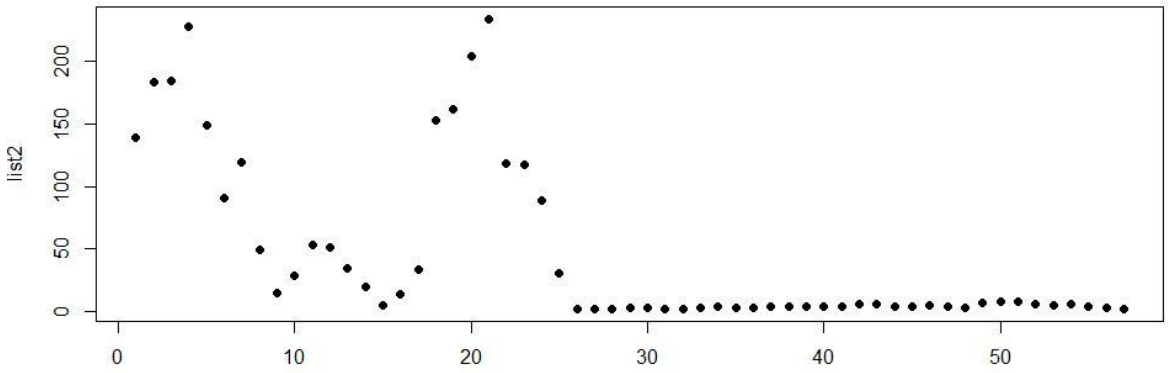
FRONT-END

We have used R's shiny package to create the frontend. Shiny package enables the user to create web applications in R easily. Shiny makes it super simple for R users like you to turn analyses into interactive web applications that anyone can use. Let your users choose input parameters using friendly controls like sliders, drop-downs, and text fields. Easily incorporate any number of outputs like plots, tables, and summaries.

No HTML or JavaScript knowledge is necessary. If one has some experience with R, he is just minutes away from combining the statistical power of R with the simplicity of a web page.

Some of the screenshots of our applications are shown below:





RESULTS

We tested our algorithm on 8 different data sets and were able to classify 6 of them correctly as to whether a topic would become a trend or not.

7. CHALLENGES AND IMPROVEMENTS

- The biggest challenge that we faced is the lack of any fixed metric or threshold values for classification
- Owing to the highly dynamic and volatile nature of twitter traffic, our results will keep on changing if we stick to a fixed metric
- Thus there is need to incorporate NLPs and Machine Learning algorithms to expect better results.
- Also the application can be made a little more dynamic so as to make it easy for the user to use.

REFERENCES

1. Towards More Systematic Twitter Analysis: Metrics for Tweeting Activities - Axel Bruns (Queensland University, Brisbane) , Stefan Stieglitz (University of Munster)
2. Credibility ranking of tweets during high impact events - Aditi Gupta, IIIT Delhi
3. Trend or No Trend: A Novel Nonparametric Method for classifying Time Series by StanislavNikolov Massachusetts Institute of Technology 2011
4. Integrating Web-based Intelligence Retrieval and decision-making from the Twitter Trends Knowledge Base- Marc Cheong, Vincent Lee 2009
5. Measuring User Influence in Twitter: The Million Follower Fallacy- Meeyoung Cha, HamedHaddadiy, Fabr´ıcioBenevenutoz, Krishna P. Gummadi,2010
6. Retweets;but Not Just Retweets: Quantifying and Predicting Influence on Twitter- Evan T.R. Rosenman,2012
7. R and Data Mining: Examples and Case Studies- Yanchang Zhao