DIC

PROJECT 1 : PROBLEM 5

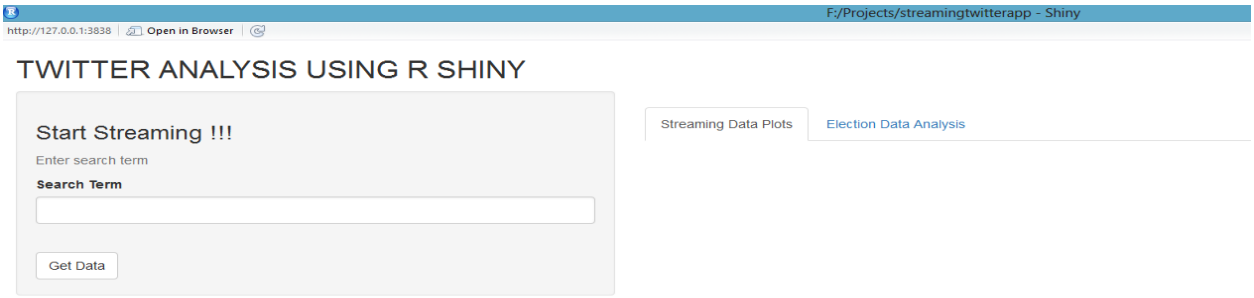
TWITTER DATA ANALYSIS USING R SHINY

NAME : ANIKET THIGALE
UB NO : 50168090

PROBLEM DESCRIPTION

Write a R script to process and analyze streaming Twitter data about ongoing elections using R Shiny. Also collect tweets for a week from various regions of the country and plot the daily and weekly trend and summarize the stats.

UI



The UI consists of a search field box for entering the term that you want to query twitter for, an action button when you want to start streaming and tabPanels for displaying some real time plots. It also has a tabPanel for displaying summary and some plots like trends for the data that we have initially collected for a week.

There are two components of this Problem:

1. Election Data Analysis
2. Streaming Data to R Shiny

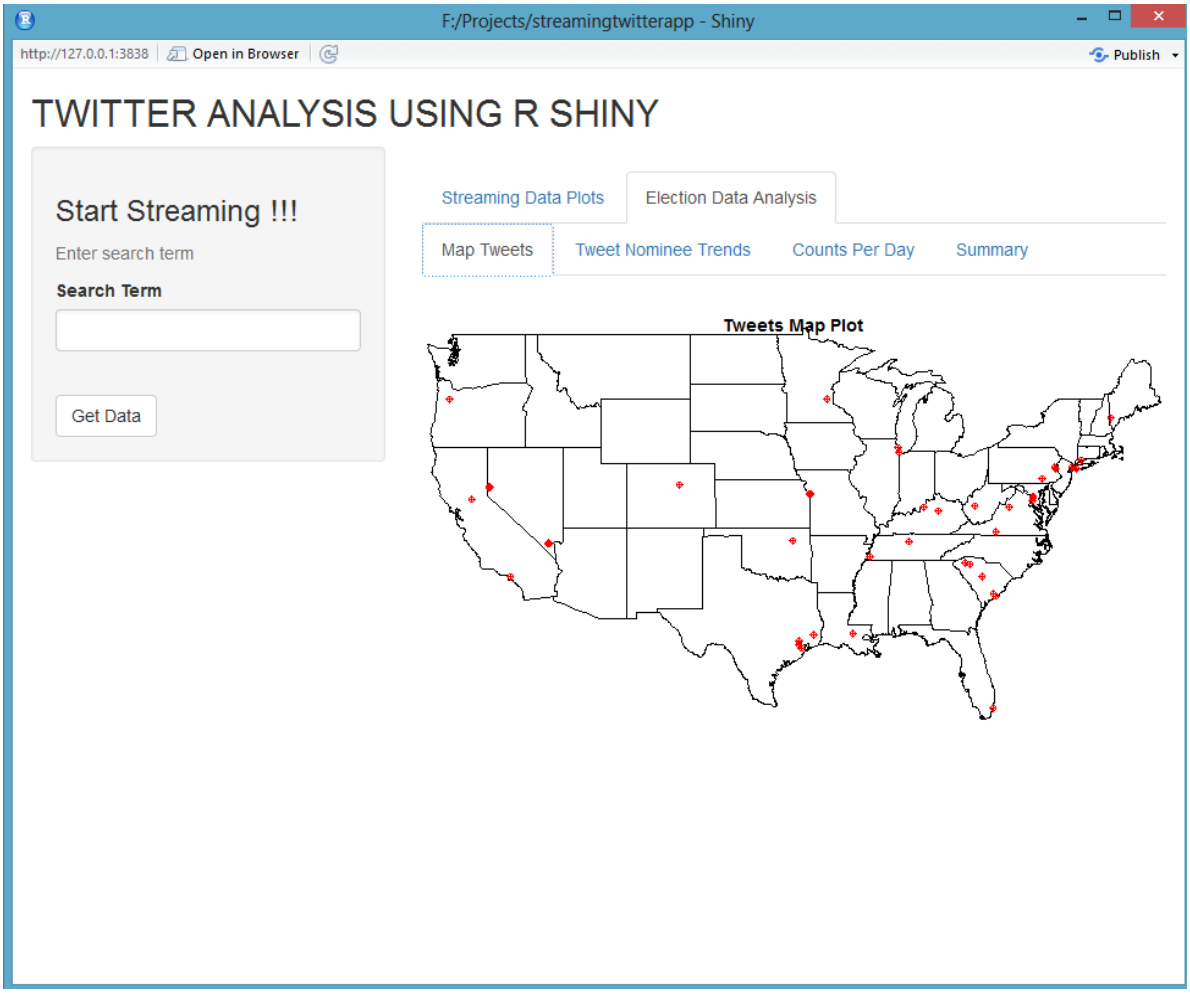
ELECTION DATA ANALYSIS

DATA

We use the election twitter data collected in problem 1 for analysis. This data has been collected over a week from February 20th 2016 to February 27th 2016 using various keywords like election2016, clinton, trump, cruz, bush etc.

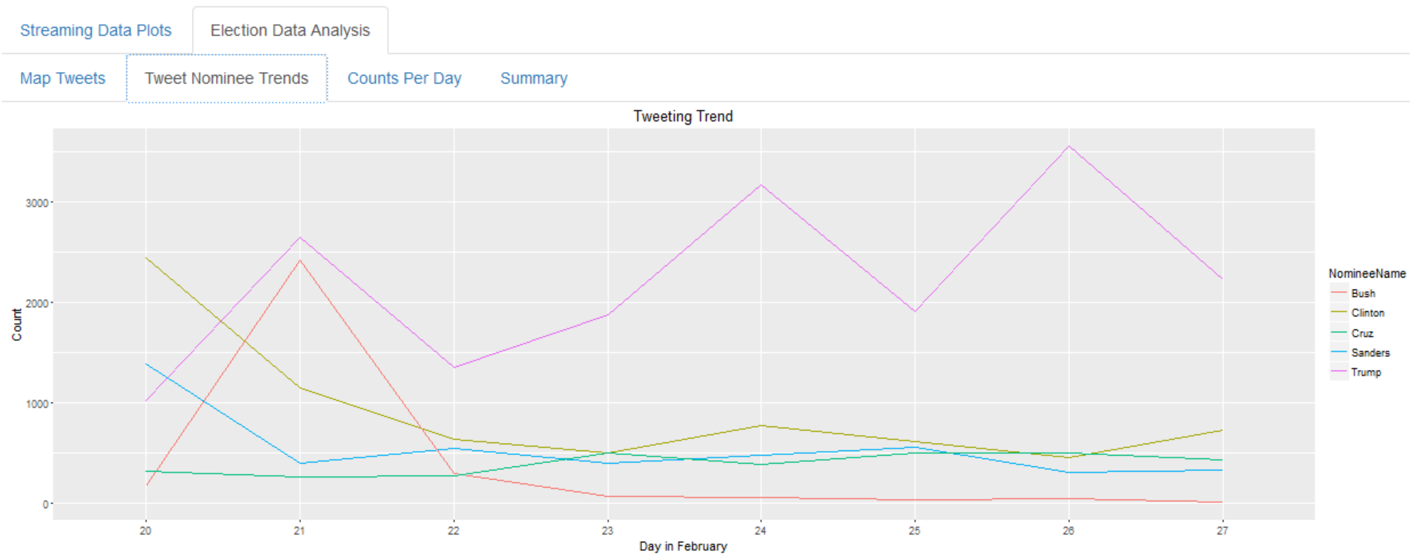
PLOTS

1. MAP



We filter tweets which have latitudes and longitudes to display them on the map. According to this, there are more tweets coming from the East Coast, but since only a small number of tweets have latitude and longitude, this is very inaccurate.

2. TWEET TRENDS



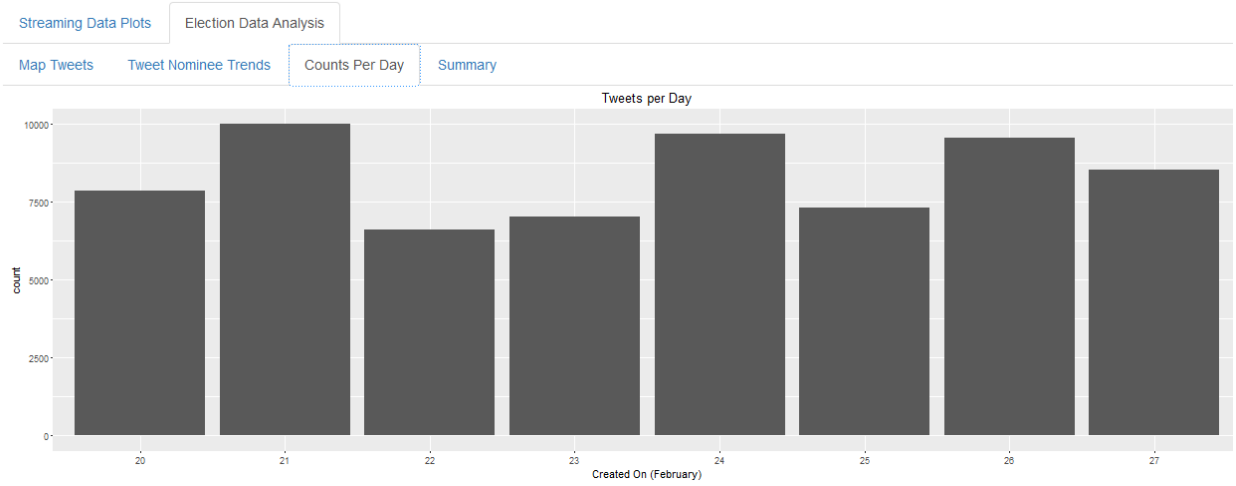
Here I have plotted the tweets count for every presidential nominee across the days.
We can see that Trump tweets topped the chart everyday, because of multiple factors including some of the controversial statements that he made in that period as well as his wins in some regions of the country.
Hillary Clinton won the Nevada caucus and Donald Trump won South Carolina on the 21st of February which results in the sharp spike of tweets on that day.

3. WORDCLOUD (not present on dashboard but my code is included)

I created a wordcloud for all tweets in which Trump was mentioned by extracting text, cleaning it by removing punctuations, stop words and certain other text, stemming words, creating a corpus, constructing Term Document Frequency Matrix using the **tm, SnowballC** and **wordcloud** package available in R.

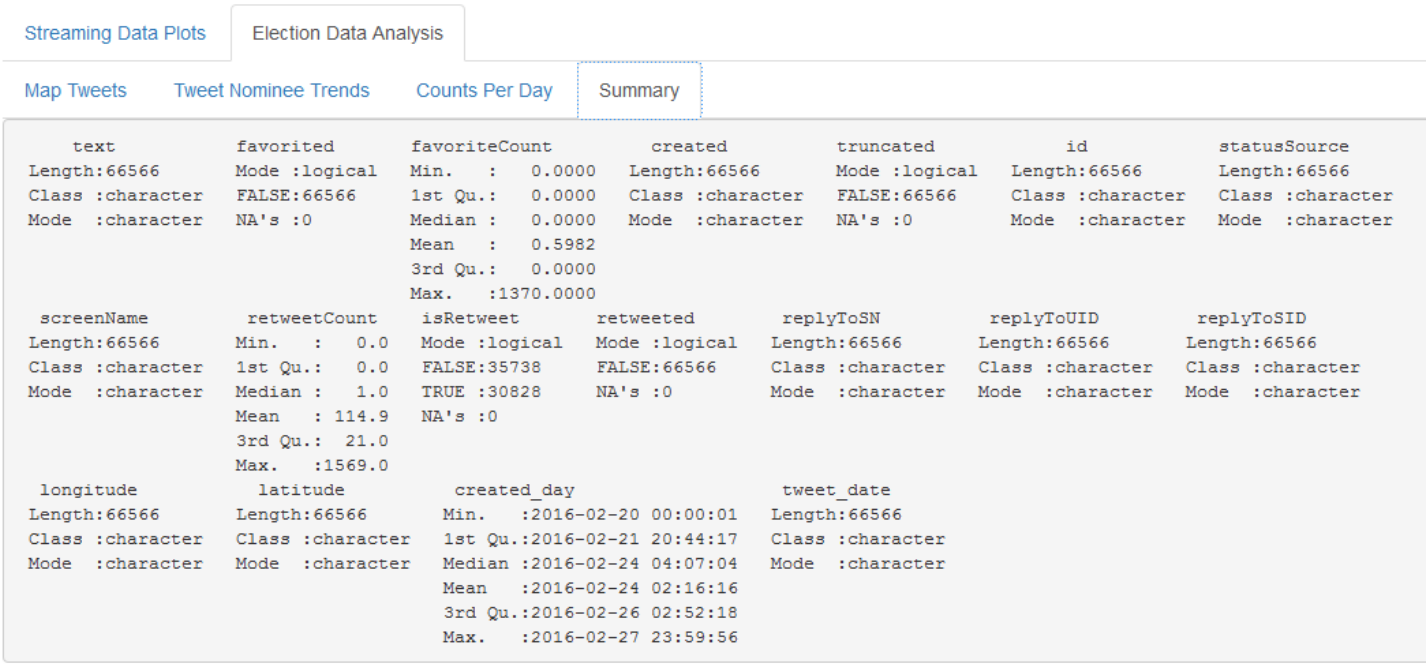
There were a large number of tweets about Trump's win in the South Carolina and Nevada elections and hence we can see the words "win" , "southprimari" , "nevada" as some of the top ones. Other keywords associated with Trump include "nevertrump", a hashtag which was trending, "racism" , "absent" which refers to the GOP debate he didn't attend and so on.

4. TWEET VOLUME



Just gives the volume of tweets recorded for every day in the range of 20th February till 27th February.

5. SUMMARY



Gives us some more information about the tweet fields like count, mean and type of field.

STREAMING DATA

We use the **streamR** package available for R to continuously stream tweets and write an observer block to detect changes.

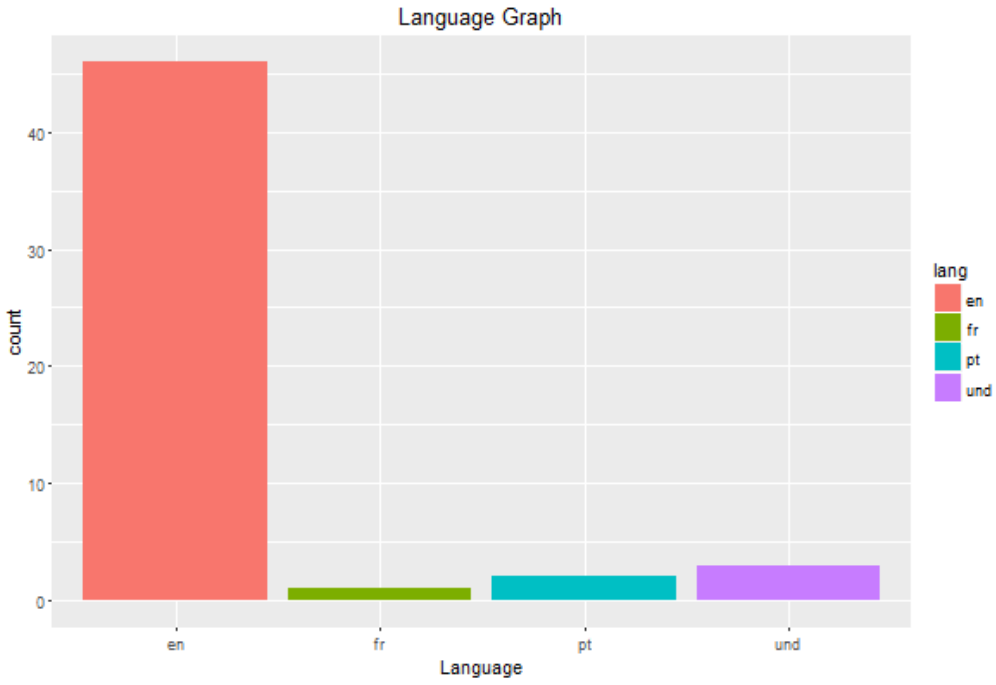
We use the filterStream() function in streamR to stream tweets which get written to a file in JSON format and use the parseTweets() function to retrieve the tweets from the JSON file into a dataframe in R.

Usage:
filterStream(file.name = "tweets.json", timeout=5,track = data, oauth = my_oauth)
parseTweets("tweets.json", verbose = FALSE)

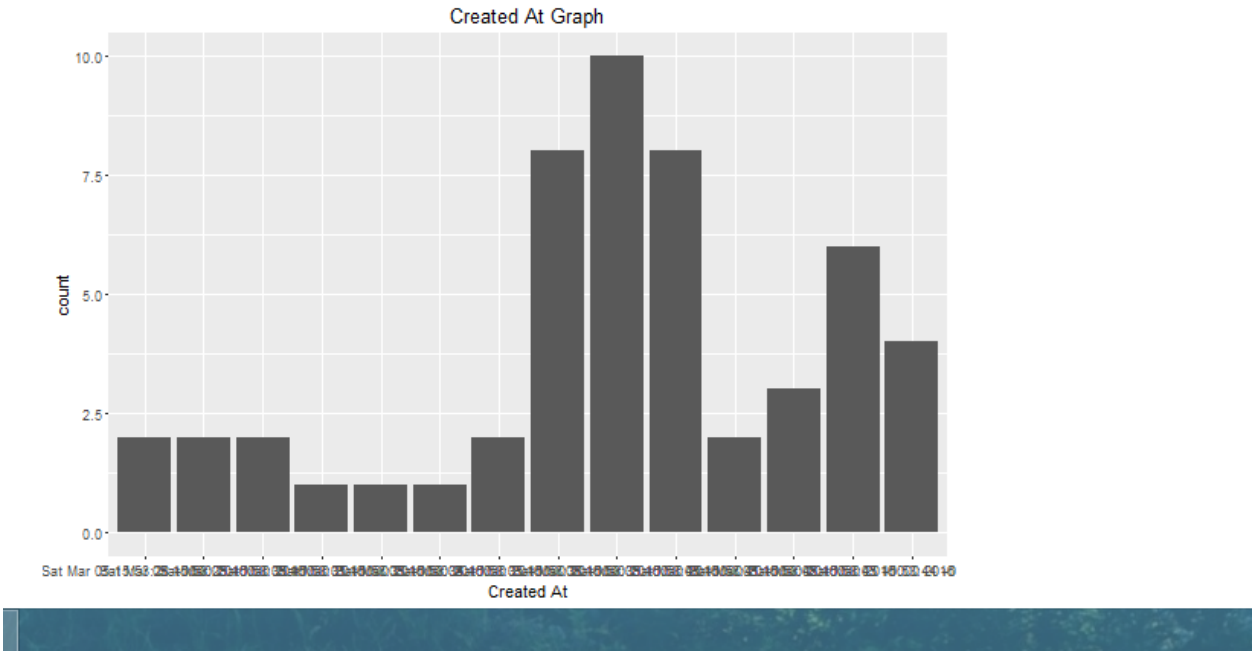
We get 43 fields for each tweet using the above API and we make use of some of them for generating the real time plots.

PLOTS (All get updated in real time)

I used a map to display the locations from the streaming tweets. The red points indicate the locations which get updated in real time.



This language graph displays the tweets collected for each language which gets updated constantly.



This graph keeps track of count of new tweets against their created time.

CONCLUSION

We have learnt to use many statistical, NLP and graph tools in R and use them to derive useful conclusions and display them using R Shiny.