Problem 5: Stream processing

Problem Statement

In this question we need to stream data from Twitter and Analyze it at the same time. Also we need to analyze election data collected over 1 week and represent in a professional UI.

Data Set

We collected data using the Data Mining script from first problem. Data collected is on topic “election”. Data collected has created date ranging from 20th Feb to 27th Feb.

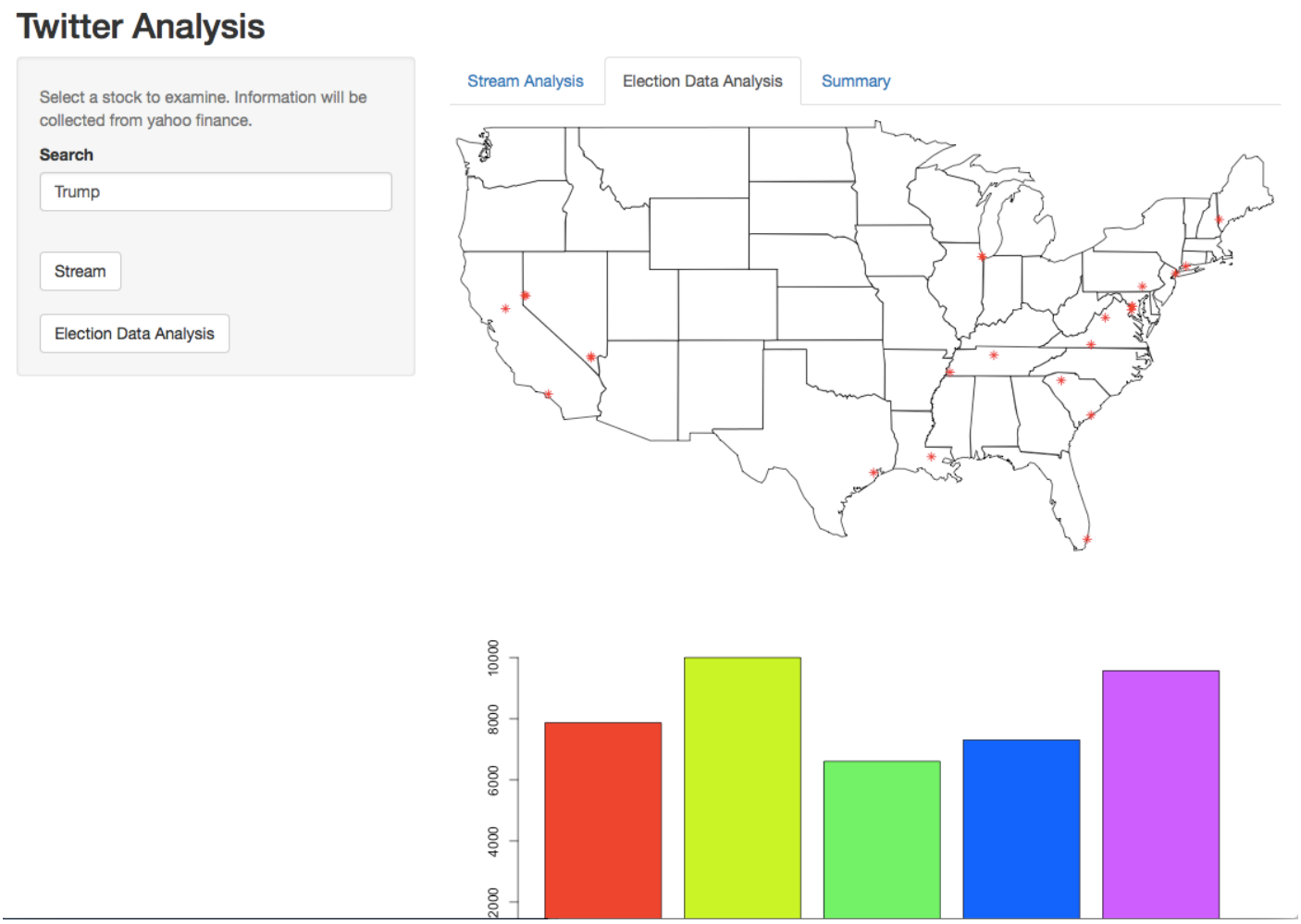
R Shiny

R Shiny is a web application framework for R. Shiny applications are automatically “live” in the same way that spreadsheets are live. Outputs change instantly as users modify inputs, without requiring a reload of the browser. It flaunts fast bidirectional communication between the web browser and R using the websockets package. We used R shiny to design the UI for this problem.

User Interface

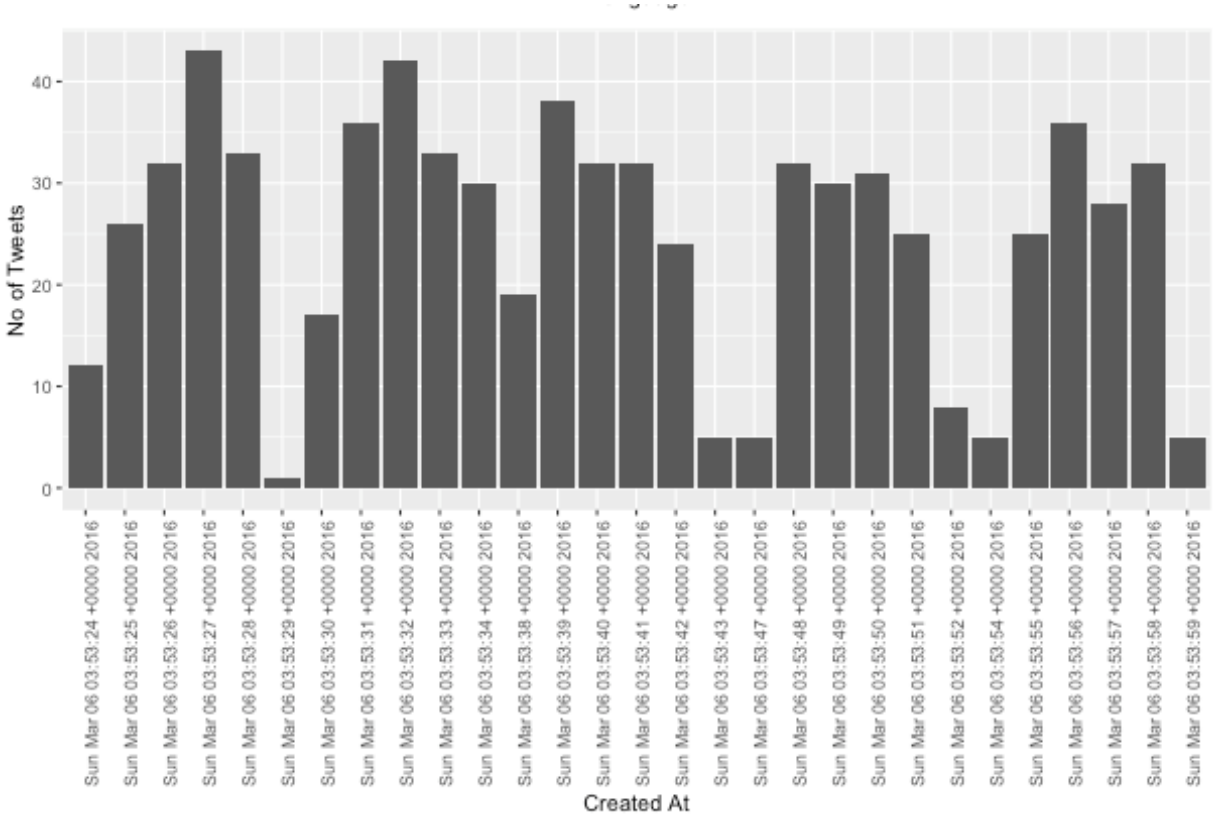
To display our analysis efficiently we used following UI components from R Shiny.

1. actionButton
2. textInput
3. tabsetPanel
4. tabPanel
5. plotOutput
6. verbatimTextOutput
7. tableOutput

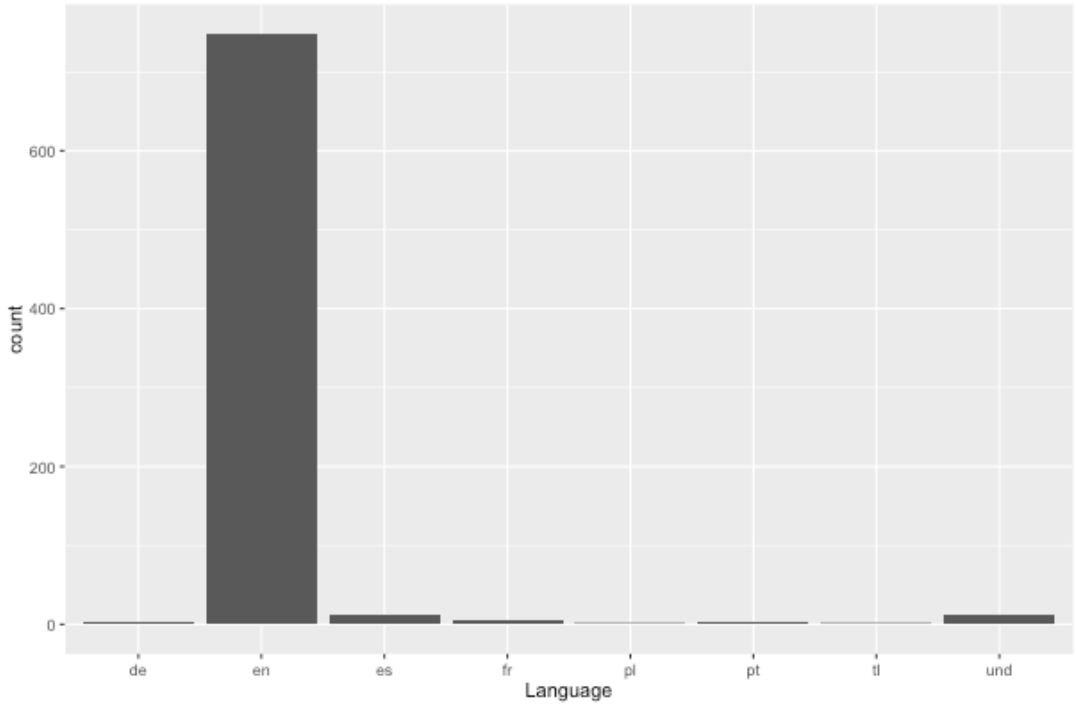


Stream Analysis

In this part, the UI facilitates user to input a query term. Using streamR package we open a stream of tweets and display the analysis of streamed tweets in runtime. The Graph and map are updated every 5 seconds.



This graph displays the number of tweets grouped by their creation time. This graph updates every 5 seconds to display almost live analysis of tweets of a given topic.

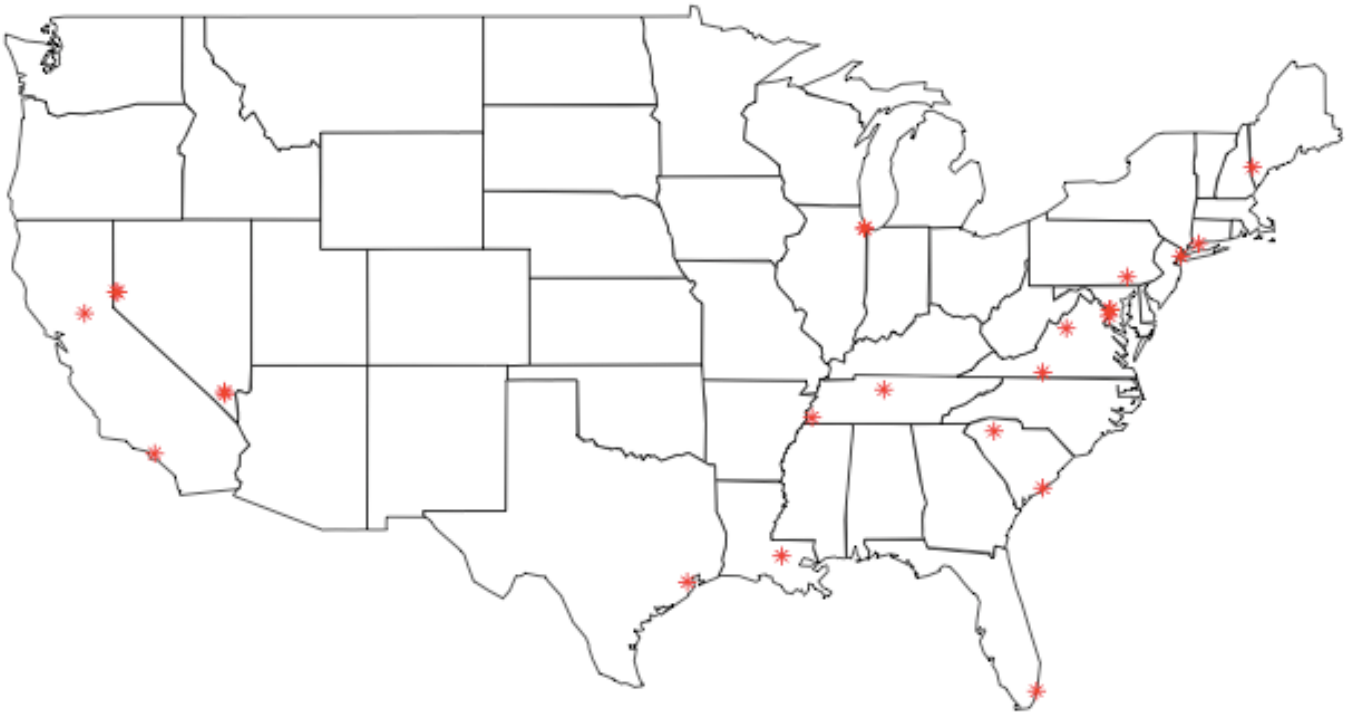


The above graph displays the language distribution of tweets. We are getting highest tweets for English as the searched topic “Trump” is popular in USA

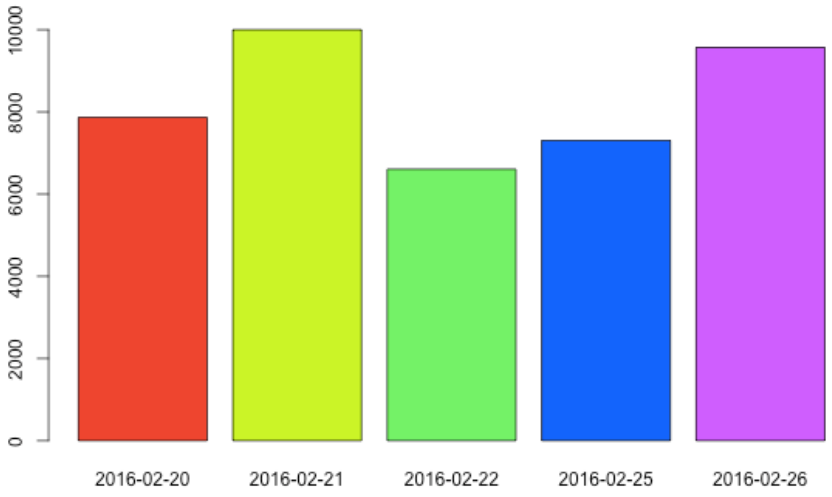
Election Data Analysis

From the election data in JSON file, we imported the data and cleaned it. Cleaning involved changing date in Date format, and latitude and longitude in numeric format to make it possible to display.

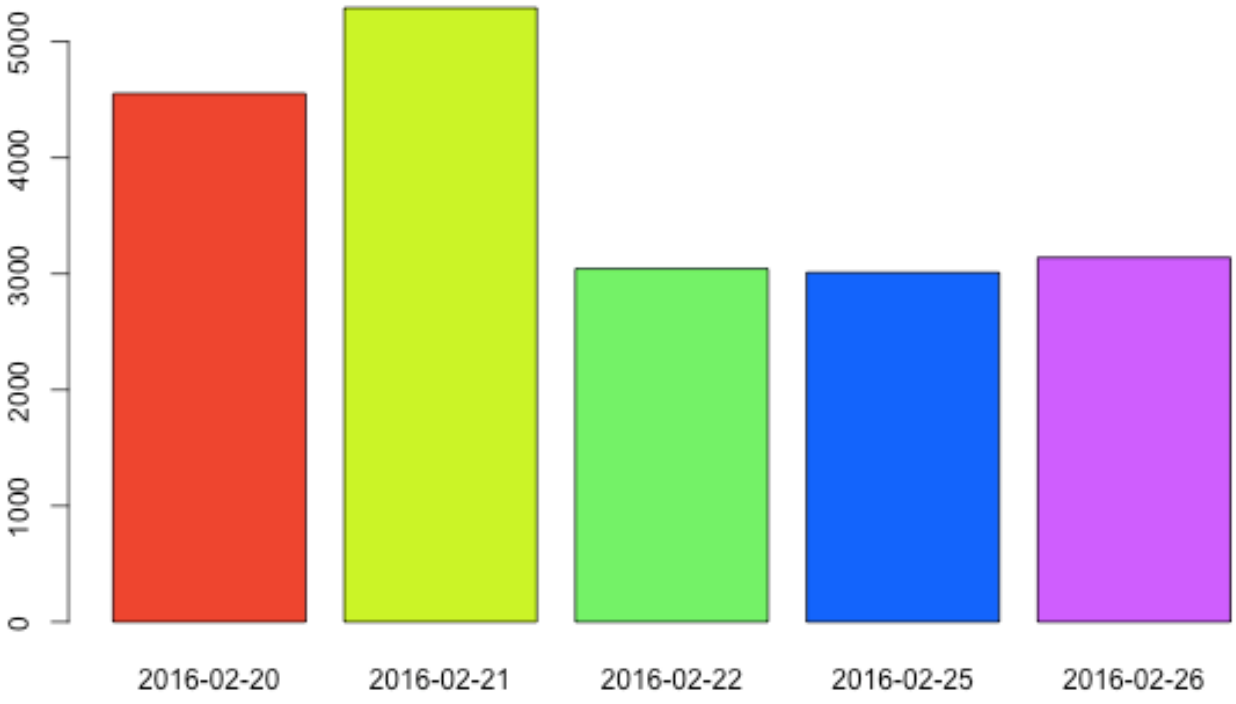
To correctly display trends of election data we represent the data collected in following formats



In the given map, Red * displays the location, where the tweet originated. We can see very few points because very few people provide their location in the tweet. More markers on east coast depicts that influence of election is more in regions in the east coast of United States.



This graph displays the distribution of tweets on over the dates.



This graph shows the count of retweeted tweets for given dates.

```
    text           favorited       favoriteCount        created          truncated
Length:41344       Mode :logical   Min.   :   0.0000   Length:41344       Mode :logical
Class :character   FALSE:41344     1st Qu.:   0.0000   Class :character   FALSE:41344
Mode  :character   NA's :0         Median :   0.0000   Mode  :character   NA's :0
                                   Mean   :   0.5866
                                   3rd Qu.:   0.0000
                                   Max.   :1370.0000

     id            statusSource        screenName         retweetCount    isRetweet
Length:41344       Length:41344       Length:41344       Min.   :   0.0   Mode :logical
Class :character   Class :character   Class :character   1st Qu.:   0.0   FALSE:22303
Mode  :character   Mode  :character   Mode  :character   Median :   1.0   TRUE :19041
                                                         Mean   : 153.2   NA's :0
                                                         3rd Qu.:  21.0
                                                         Max.   :1569.0

retweeted        replyToSN          replyToUID         replyToSID         longitude
Mode :logical    Length:41344       Length:41344       Length:41344       Length:41344
FALSE:41344      Class :character   Class :character   Class :character   Class :character
NA's :0          Mode  :character   Mode  :character   Mode  :character   Mode  :character



  latitude           created1            latitude1        longitude1
Length:41344       Min.   :2016-02-20   Min.   :-37.81   Min.   :-121.23
Class :character   1st Qu.:2016-02-21   1st Qu.: 17.97   1st Qu.: -81.57
Mode  :character   Median :2016-02-22   Median : 35.61   Median : -76.78
                   Mean   :2016-02-22   Mean   : 29.78   Mean   : -43.60
                   3rd Qu.:2016-02-25   3rd Qu.: 40.58   3rd Qu.:  -6.64
                   Max.   :2016-02-26   Max.   : 54.11   Max.   : 144.96
                                        NA's   :41294    NA's   :41294
```

This gives a brief summary of the tweets collected on election over 7 days.

Conclusion

We can conclude that using R Shiny we can represent our analysis is a very beautiful and interactive way. Functions like reactive and observe makes the application come to life.