TWITTER ANALYSIS

(IN RSTUDIO USING R PROGRAMMING LANGUAGE)

**Table of Contents**

[Introduction 4](#_gjdgxs)

[Limitations 4](#_30j0zll)

[Tools and Packages used 5](#_1fob9te)

[Twitter Analysis: 6](#_3znysh7)

[Creating a Twitter Application 6](#_2et92p0)

[Working on RStudio- Building the corpus 8](#_tyjcwt)

[Saving Tweets 11](#_1t3h5sf)

[Sentiment Function 12](#_4d34og8)

[Scoring tweets and adding column 13](#_2s8eyo1)

[Import the csv file 14](#_17dp8vu)

[Visualizing the tweets 15](#_3rdcrjn)

[Analysis & Conclusion 16](#_26in1rg)

[Text Analysis 17](#_lnxbz9)

[Final code for Twitter Analysis 19](#_35nkun2)

[Final code for Text Mining 20](#_1ksv4uv)

[References 21](#_44sinio)

# Introductions

**Twitter** is an amazing micro blogging tool and an extraordinary communication medium. In addition, twitter can also be an amazing open mine for text and social web analyses. Among the different softwares that can be used to analyze twitter, **R** offers a wide variety of options to do lots of interesting and fun things. In this project I have used **RStudio** as its pretty much easier working with scripts as compared to **R**.

According to Wikipedia, Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

**Sentiment analysis**, also referred to as Opinion Mining, implies extracting opinions, emotions and sentiments in text. As you can imagine, one of the most common applications of sentiment analysis is to track attitudes and feelings on the web, especially for tacking products, services, brands or even people. The main idea is to determine whether they are viewed positively or negatively by a given audience.

The purpose of **Text Mining** is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, you can analyze words, clusters of words used in documents, or you could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining project. In the most general terms, text mining turns “text into number” which can then be incorporated in other analyses.

Applications of text Mining are analyzing open-ended survey responses, automatic processing of messages, emails, etc., analyzing warranty or insurance claims, diagnostic interviews, etc., investigating competitors by crawling their web sites.

# Limitations

There are certain limitations while doing Twitter Analysis using R. Firstly, while getting Status of user timeline the method can only return a fixed maximum number of tweets which is limited by the Twitter API.

Secondly, while requesting tweets for a particular keyword, it sometime happens that the number of retrieved tweets are less than the number of requested tweets.

Thirdly, while requesting tweets for a particular keyword , the older tweets cannot be retrieved.

# Tools and Packages used

In this project “Twitter Analysis using R” I have used RStudio GUI and following packages:

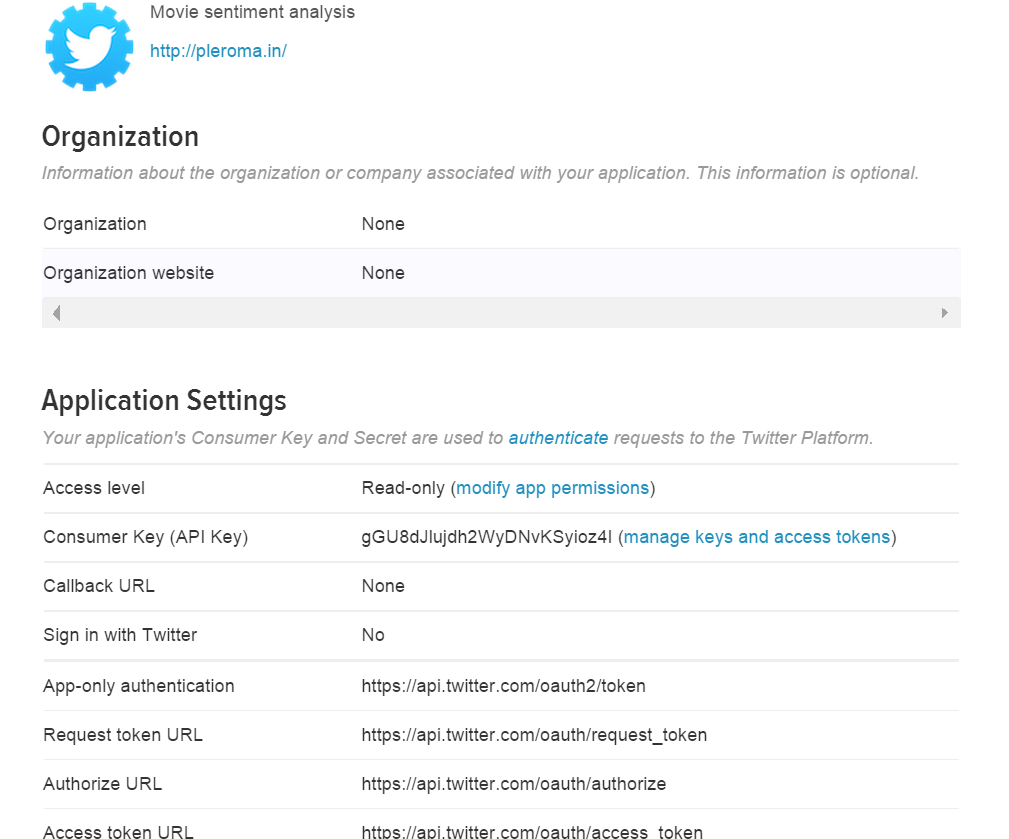
* **twitteR** : Provides an interface to the Twitter web API.
* **ROAuth** : This package provides an interface to the OAuth 1.0 specification, allowing users to authenticate via OAuth to the server of their choice.
* **plyr** : This package is a set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each pieces and then put all the pieces back together.
* **stringr** : stringr is a set of simple wrappers that make R's string functions more consistent, simpler and easier to use. It does this by ensuring that: function and argument names (and positions) are consistent, all functions deal with NA's and zero length character appropriately, and the output data structures from each function matches the input data structures of other functions.
* **ggplot2** : An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources.
* **RColorBrewer** : The packages provides palettes for drawing nice maps shaded according to a variable.
* **tm** : A framework for text mining applications within R.
* **wordcloud** : This package helps in creating pretty looking word clouds in Text Mining.

# Twitter Analysis:

## Creating a Twitter Application

First step to perform Twitter Analysis is to create a twitter application. This application will allow you to perform analysis by connecting your R console to the twitter using the Twitter API. The steps for creating your twitter applications are:

* Go to https://apps.twitter.com/ and login by using your twitter account.
* Then go to My Applications → Create a new application.
* Give your application a name, describe about your application in few words, provide your website’s URL or your blog address (in case you don’t have any website). Leave the Callback URL blank for now. Complete other formalities and create your twitter application. Once, all the steps are done, the created application will show as below. Please note the Consumer key and Consumer Secret numbers as they will be used in RStudio later.



## Working on RStudio- Building the corpus

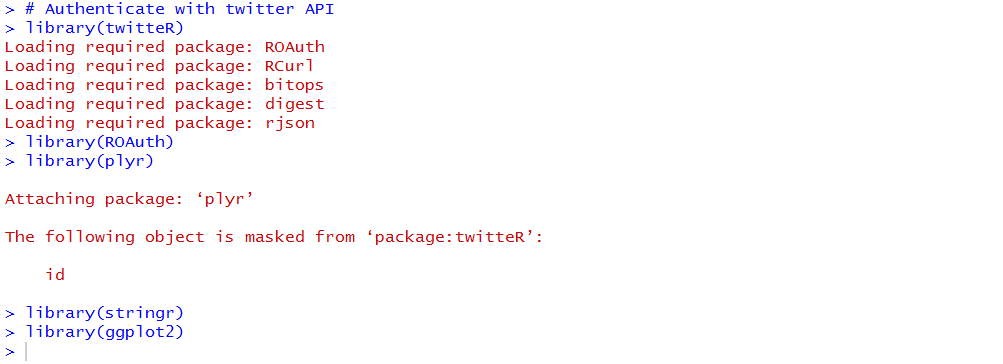
In this section, I will first use some packages in R. These are twitter, ROAuth, plyr, stringr and ggplot2. You can install these packages by the following commands:

C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png

Now run the following R script code snippet



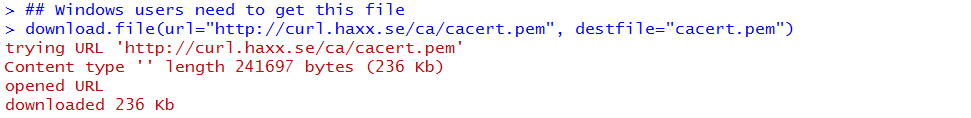
After running this script section, the console will look like this



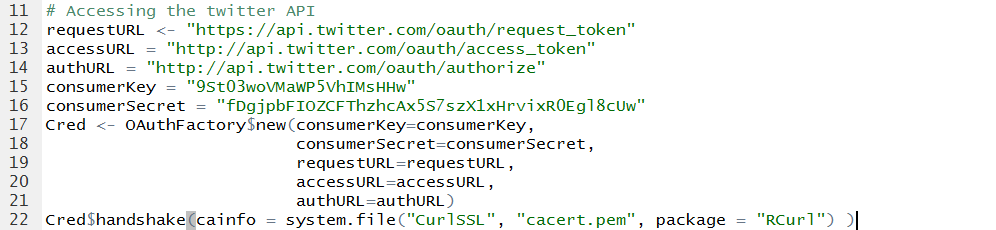
Now, windows users need to download a small file by following command

C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png

After running this code, RStudio will look for the following file at the given url, and will download it for you. Your console will look like this:



Now once this file is downloaded, we are now moving on to accessing the twitter API. This step include the script code to perform handshake using the Consumer Key and Consumer Secret number of your own application. You have to change these entries by the keys from your application. Following is the code you have to run to perform handshake.



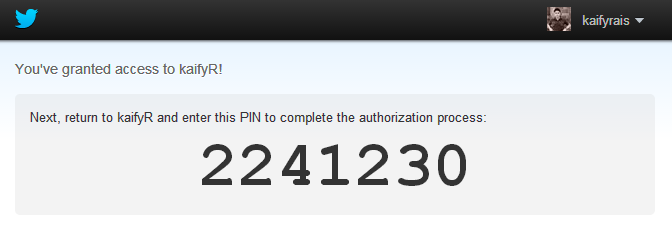
Here, line number 12,13 and 14 assign the request url, access url and authorization url of twitter application to the variables requestURL, accessURL and authURL respectively. consumerKey and consumerSecret are unique to a twitter applicartion. Running this gives following message on the R console :



The last three lines of the console are a message to the user. To enable the connection, please direct your web browser to:

<http://api.twitter.com/oauth/authorize?oauth_token=dHwEGXdxbjJ093sG0tVjYVT0NQrkjU3DuCxcC1YQyc>

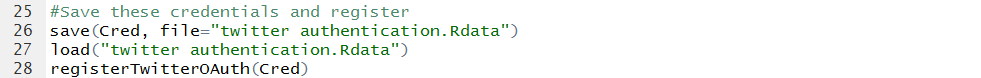
After opening the above link in your browser, authorize application by providing you username and password. And the app will be authorized. You will receive a code like this:



Write this code in the console. The console will give a message like this.

C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png

Now register the handshake by following command



The console will give a message with TRUE, which means that the handshake is complete. Now we can get the tweets from the twitter timeline.

C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png

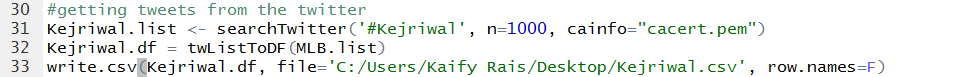
## Saving Tweets

Once the handshake is done and authorized by twitter, we can fetch most recent tweets related to any keyword. I have used #Kejriwal as Mr. Arvind Kejriwal is the most talked about person in Delhi now a day.

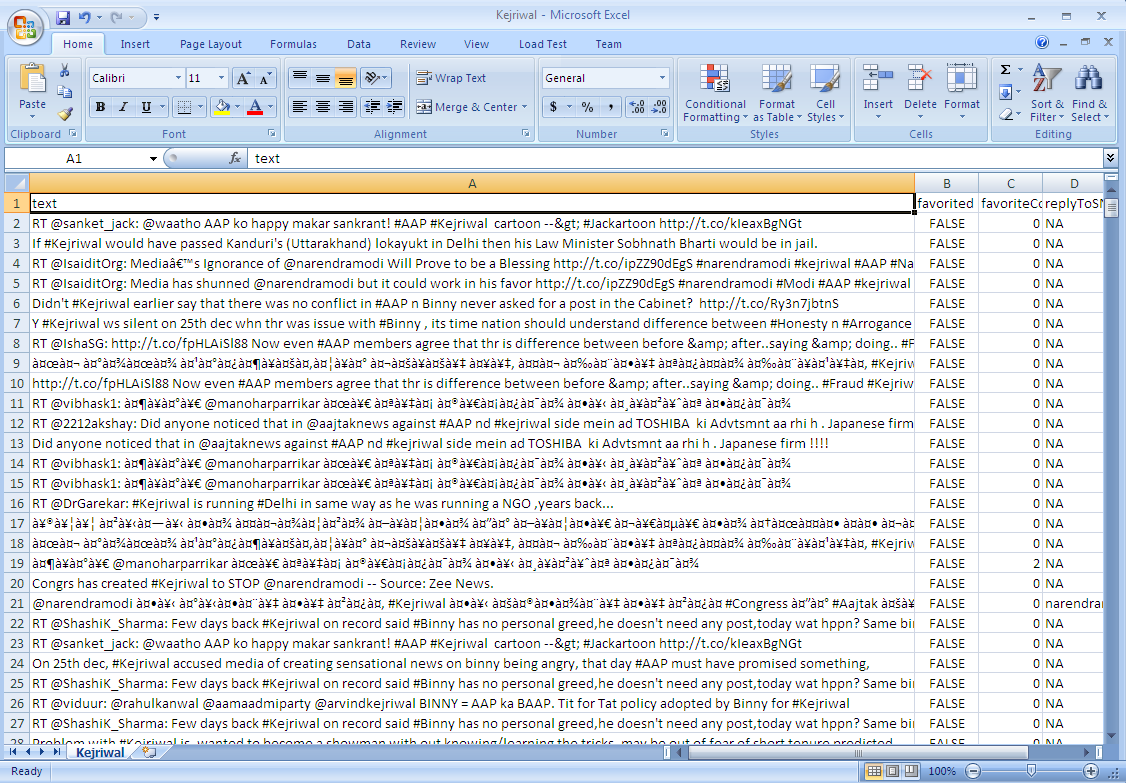
The code for getting tweets related to #Kejriwal is:

C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png

This command will get 1000 tweets related to Kejriwal. The function “searchTwitter” is used to download tweets from the timeline. Now we need to convert this list of 1000 tweets into the data frame, so that we can work on it. Then finally we convert the data frame into .csv file



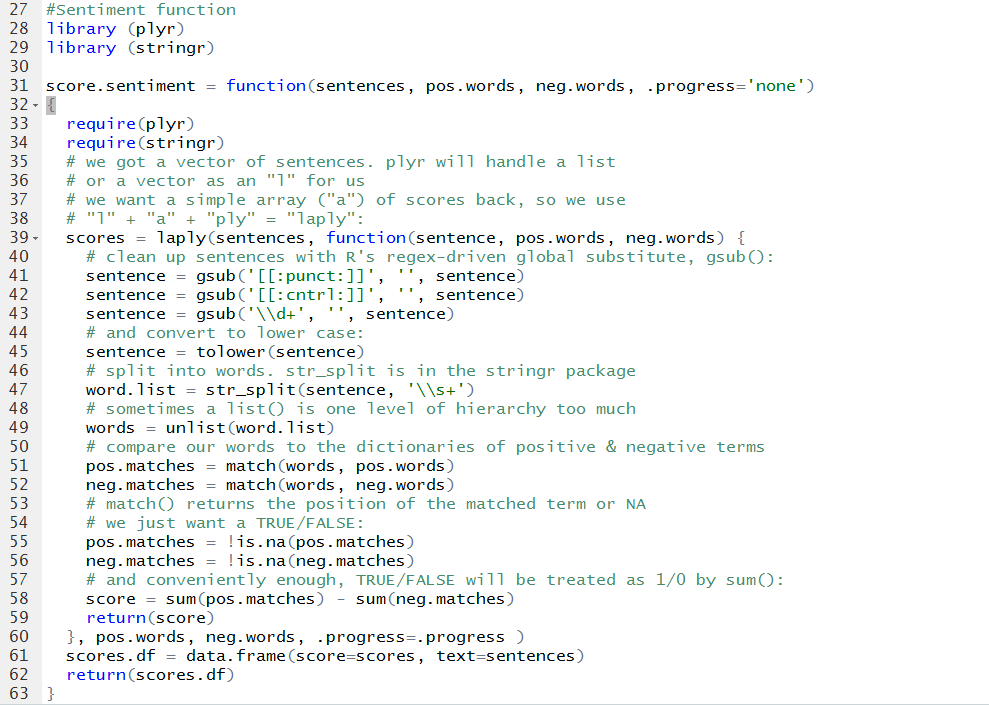
The Kejriwal.csv file contains all the information about the tweets. A snapshot of the csv file is given below:



## Sentiment Function

Once we have the tweets we just need to apply some functions to convert these tweets into some useful information. The main working principle of sentiment analysis is to find the words in the tweets that represent positive sentiments and find the words in the tweets that represent negative sentiments. For this we need a list of words that contains positive and negative sentiment words. I have downloaded the list from Google and it is easily available.

After downloading the list, save it in your working directory. The sentiment analysis uses two packages plyr and stringr to manipulate strings. The function is



The sentiment function calculate score for each individual tweet. It first calculate the positive score by comparing words with the negative words list and then calculate negative score by comparing words with negative words list. The final score is calculated as

score= positive score – negative score.

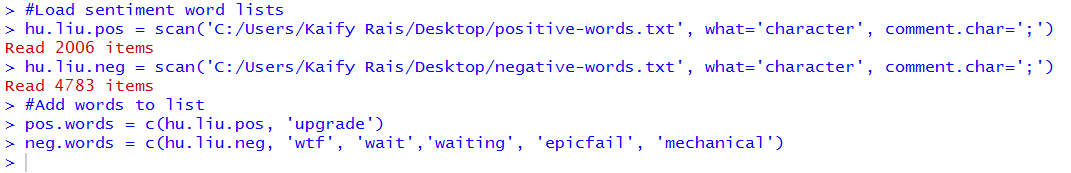
## Scoring tweets and adding column

In this step we score tweets from the above sentiment function.

C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png

C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png

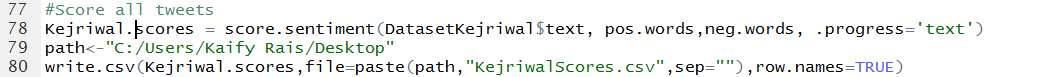
The console gives the following output



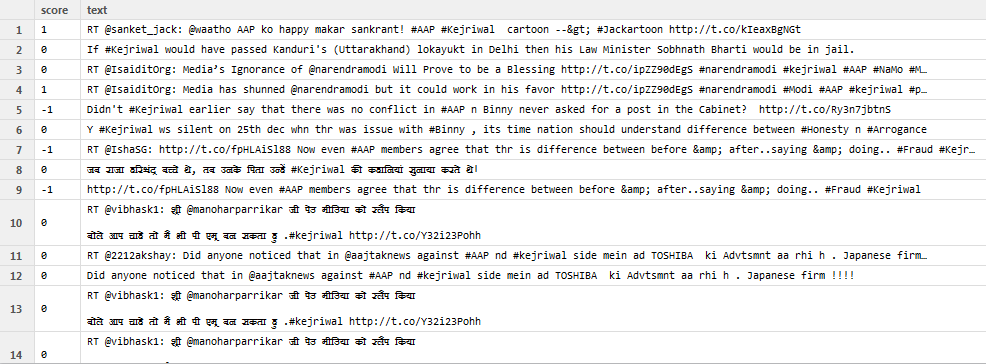
## Import the csv file

C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png

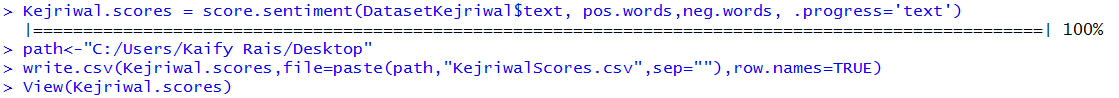
When we import this csv file, a dataset file is created in the working directory. Next step is to score the tweets, this can be done by creating a separate csv file which contains the score of each tweet. This can be done as follows:



The snapshot of the score file shows the score of each tweet as an integer in front of every tweet.

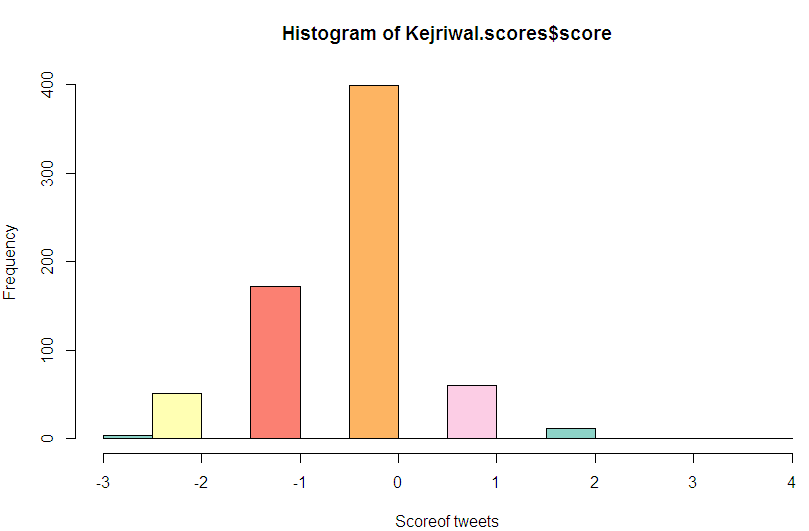


And this is the status shown on console



## Visualizing the tweets

Now that all the work is done. We can create visual histograms and other plots to visualize the sentiments of the user. This can be done by using hist function. I have used a package RColorBrewer to play with colors. The code for creating histogram is



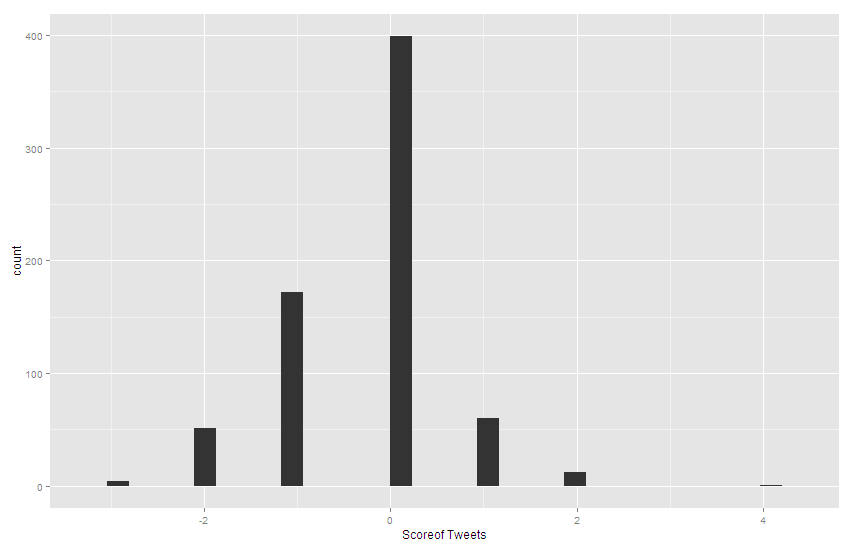
# Analysis & Conclusion

The above histogram shows the frequency of tweets with respect of scores allotted to each tweets. The x-axis shows the score of each tweet as a negative and positive integer or zero. A positive score represents positive or good sentiments associated with that particular tweet whereas a negative score represents negative or bad sentiments associated with that tweet. A score of zero indicates a neutral sentiment. The more positive the score, the more positive the sentiments of the person tweeting and vice-versa.

The above histogram is skewed towards negative score which shows that the sentiments of people regarding Mr. Kejriwal are negative. This can be justified as in recent days , some of his schemes has backfired such as holding Janta Darbar has been scrapped because of chaos it created.

This can also be stated using a Quick plot with following commands:

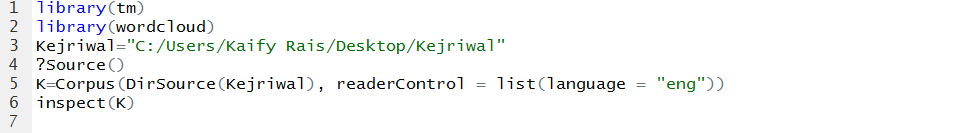
C:\Users\KAIFYR~1\AppData\Local\Temp\msohtmlclip1\01\clip_image001.png



Out of 1000 tweets that were fetched from the twitter, A majority of them (400) are neutral, whereas around 225 were having negative sentiments. Less than 100 tweets were having positive sentiments but the overall score is negative as can be seen from the plot.

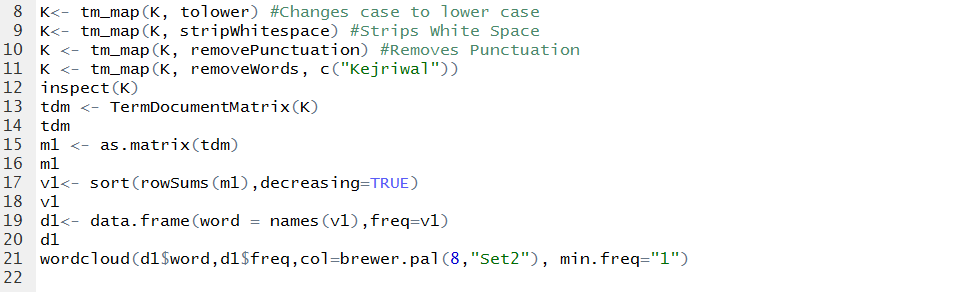
# Text Analysis

Now that we have created a csv file containing tweets of the #Kejriwal. We can do the text mining of the tweets. The procedure for this use wordcloud and tm package for text mining. The code for text mining is:

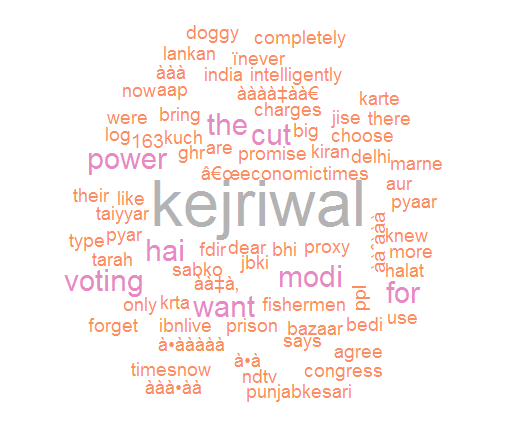


The output of console will be a collection of all the tweets saved in that file.

Now, we perform some text mining functions to refine the text and filter the text according to our need.



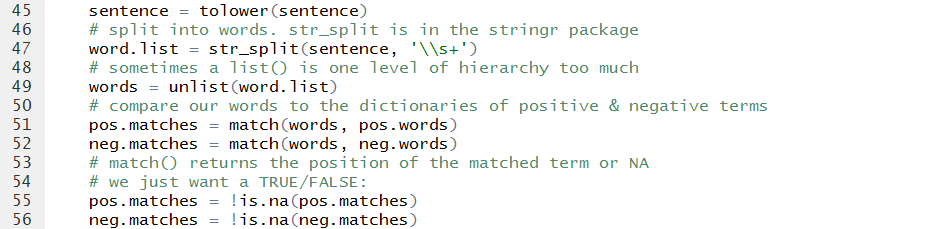
The final wordcloud obtained is as follows:

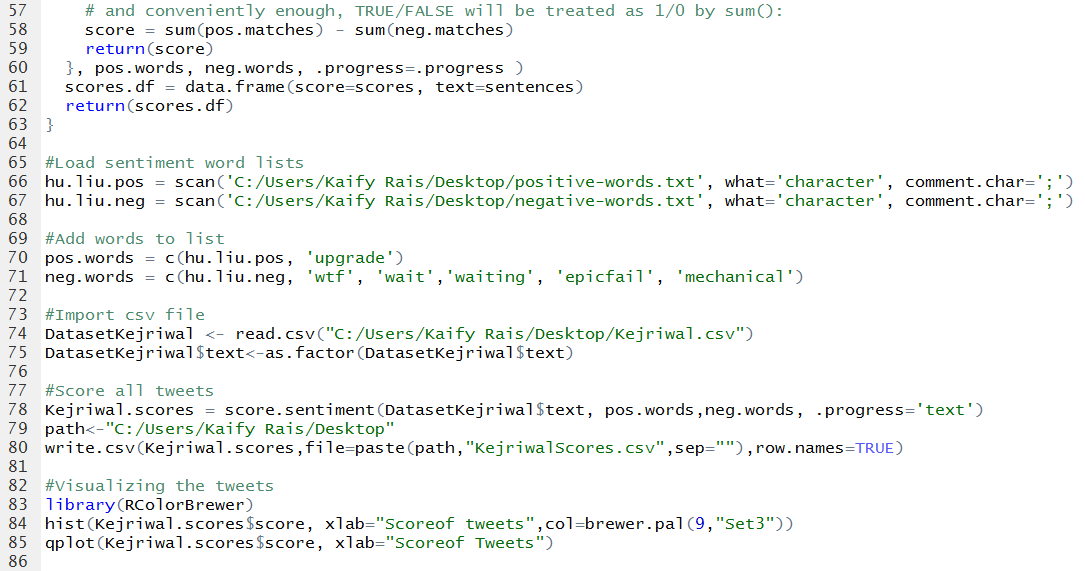


According to this wordcloud , we can see that Kejriwal is the most used term in the tweets followed by power, cut, modi which shows that while tweeting about Kejriwal the person also connect the term Kejriwal with the words like voting, power cuts and Modi.

# Final code for Twitter Analysis







# Final code for Text Mining



# References

<http://txcdk.unt.edu/iralab/sentiment_analysis>

<https://sites.google.com/site/miningtwitter/questions/sentiment/analysis>