**Data science project on sentiment analysis using twitter API**

**A Project submitted**

**in Partial Fulfillment of the requirements**

**for the Degree, of**

**Bachelors of Science (Honors)**

**in**

COMPUTER SCIENCE

**by**

**SATYAM PANDEY**

**ROLL NO. 15013570044**

**Under the Supervision of**

**Mrs. Geetika Vashisht**



**Department of Computer Science**

**College of Vocational Studies, DU**

**to the**

**Department of Computer Science  
University of Delhi**



**Index**

1. Problem statement

2. Introduction

3. Packages used

4. Why not the base package?

5. Comparison/Analysis of why a particular plot type has been used

6. Concept

7. Implementation

8. Conclusion

9. References

**Problem statement**

The problem is to make an application in R which does sentiment analysis using tweets from Twitter and we analyze the tweets of two rival teams and predict the score based on the mood within the teams by calculating their scores on certain predetermined parameters.

**Introduction**

1. Text based sentiment analysis: To check whether the word entered has been used in a positive or negative sense in a tweet.

2. Football Match tweet analysis: To know whether the tweets of the football matches are positive or negative and hence utilize this data to find out the global appeal and the popularity of the team and to know whether which team is winning the online game.

3. Football Match comparison: To compare the chances of winning of two teams. And we can also know which twitter account has better fan engagement.

**Packages used**

The packages used are:

1. **purr:** purrr provides several versions of map that allow you to specify the structure of your output.
2. **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.
3. **twitteR**: twitteR is an R package which provides access to the Twitter API. Most functionality of the API is supported, with a bias towards API calls that are more useful in data analysis as opposed to daily interaction.
4. **plyr**: plyr is an R package that makes it simple to split data apart, do stuff to it, and mash it back together. This is a common data-manipulation step. Importantly, plyr makes it easy to control the input and output data format with a consistent syntax.
5. **dplyr**: dplyr provides simple “verbs”, functions that correspond to the most common data manipulation tasks, to help you translate your thoughts into code. It uses efficient backends, so you spend less time waiting for the computer.

**Why not the base package?**

The base package is not used because purrr provides several versions of map that allow you to specify the structure of your output.. The plyr package makes it easy to split data apart. The twitteR package provides access to the Twitter API. The stringr package makes strings easier to use.

**Comparison/Analysis of why a particular plot type has been used**

**Histogram**

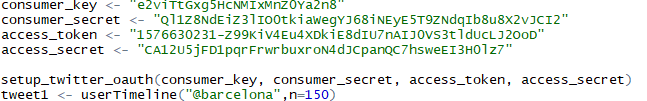
It is used to calculate the score to display of whether the text is entered in positive or negative sense in a tweet. We want to know the level of positivity or negativity in the tweet that contains the particular word entered in text based sentiment analysis hence we use histogram in it.

**Concept of the project**

* We are doing the sentiment analysis suppose you have an application which is very famous, and has a billion users, and you decide to add a new functionality to your application, How will you get the feedback of it ?
* The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer’s attitude towards a particular topic product, etc. is positive, negative or neutral.
* We are going to do an El Clasico comparison of real Madrid and Barcelona and see who is getting more positive tweets.

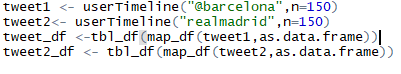
1. **Using TwitteR package to get data from Twitter API in R**

First get a twitter account if you don’t have one by registering on twitter.com website. Next go to (https://dev.twitter.com) with the same credentials. Here we will be setting up the API. This API is required for authentication to search tweets from a third party application. It uses an industry standard process called OAuth. OAuth creates the handshake between twitter and R using something called as “Consumer Key” and “Consumer Secret”.



1. **Obtaining data from twitter API**

userTimeline(user,n) function is used to get/retrieve tweets from various timelines within the Twitter universe. The twListToDF() is used to convert a list of tweets to a data frame.



1. **Importing positive and negative words from text file for sentiment analysis**.



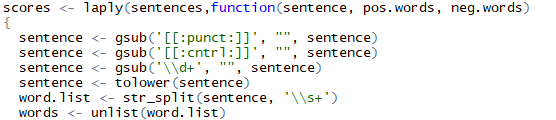
1. **Cleaning data obtained from twitter in the sentiment score calculation function described next.**

The gsub(pattern, replacement, x) function searches for the pattern in x and then replaces with replacement variable in x.

The tolower(sentence) converts each and every character in sentence to lowercase.

The str\_split(sentence,separator) is a function in stringr package to split a string into pieces.

The unlist(x) function when given a list structure x, unlist simplifies it to produce a vector which contains all the atomic components which occur in x.

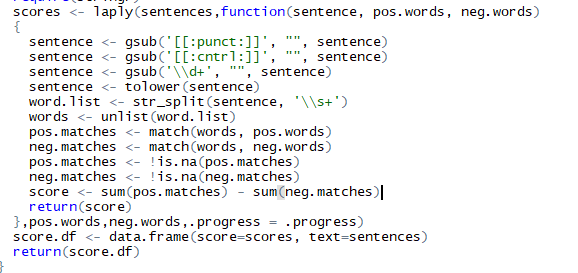


1. **Matching positive and negative words after fetching them from files against tweets retrieved and then giving the sentiment score.**

match() function returns a vector of the positions of (first) matches of its first argument in its second.

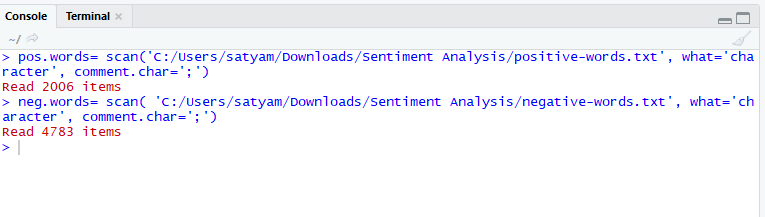
isna() function is used to check whether the value in the vector is not available.

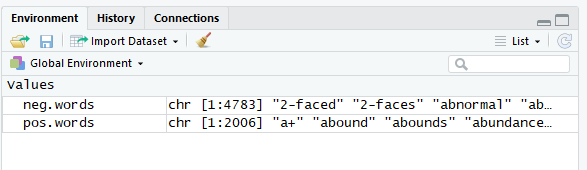
data.frame() function is used to make a dataframe.



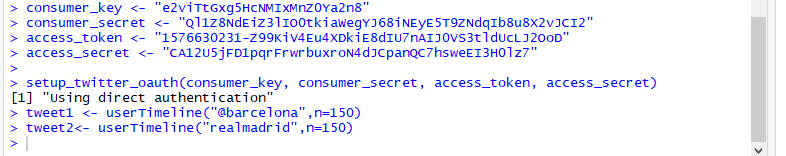
**Implementation of code**

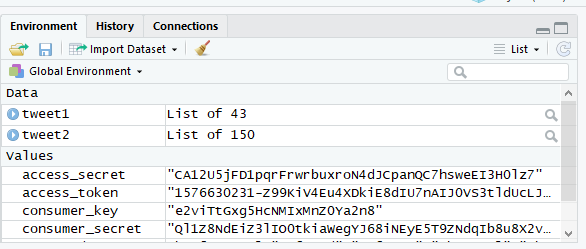
* Including positive and negative words from environment





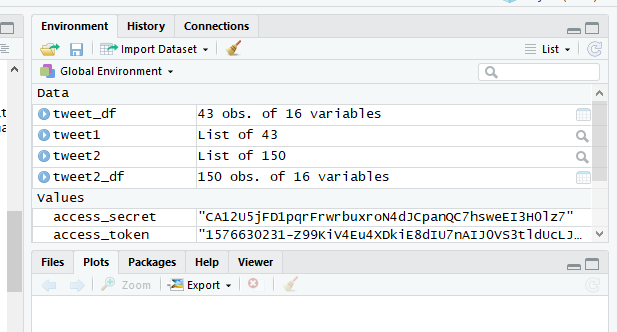
* Streaming tweets from twitter





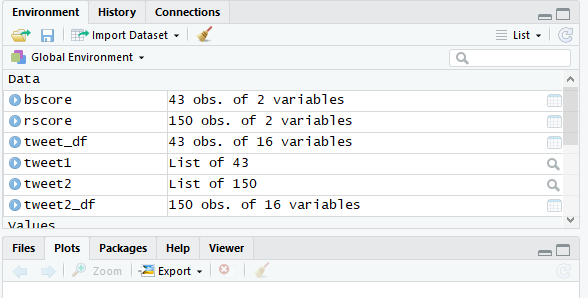
* Converting tweets into data frame

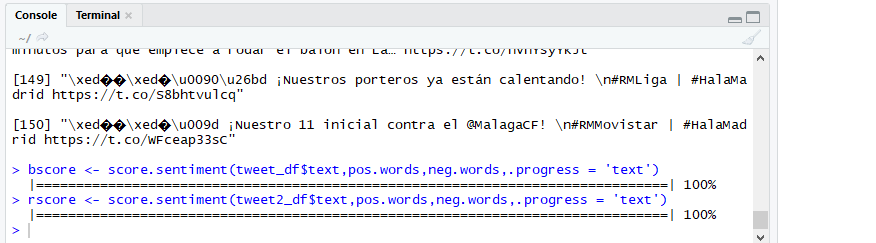




* Executing our function

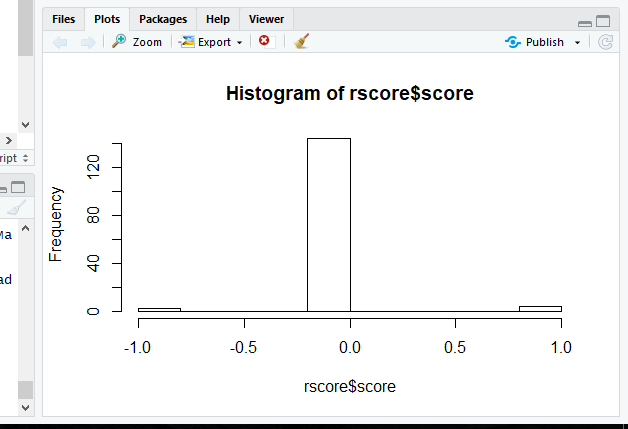




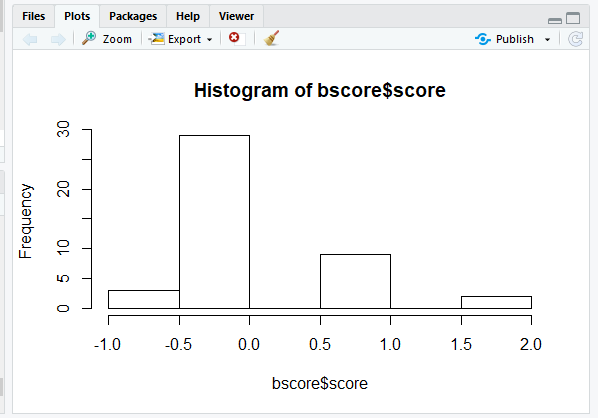


* Displaying text based sentiment analysis using histogram

1. histogram of r score



1. histogram of b score



**Analysis of histogram**

**Application of text based sentiment analysis**

**Conclusion**

* The text based sentiment analysis help us to analyze in general which word is used in a positive or negative sense in a tweet.
* It helps us to conclude which team is better among the two teams the user enters and helps us to know the winning chances of both the teams.
* The El Clasio team tweet analysis helps us to conclude about the behaviour/mood of the person who is following that particular team.

**References**

* https://www.rdocumentation.org/
* https://www.edureka.co/
* https://youtube.com/
* https://developer.twitter.com/