SENTIMENT ANALYSIS FOR HATE SPEECH ACTIVITY DETECTION ONLINE IN KENYAN

By

WILFRED WAKANYA GATHINGIRAH

BOBIT/NRB/4181/16

A Project Proposal Submitted for the study leading to a Project Report in partial fulfillment of the requirements for the award of a Bachelor of Business and Information Technology of St. Paul's University

Supervisor,

PIUS NYAANGA MOMANYI

DATE, 13th April 2019

ABSTRACT

Hate speech on social media has unfortunately become a common occurrence in the Kenyan online community largely due to advances in mobile computing and the internet. Incidents of hate speech on social media have the potential of quickly disseminating amidst online users and escalating into acts of violence and hate crimes due to incitement, as was the case during the 2007-2008 Post Election Violence.

Current efforts by the National Cohesion and Integration Commission to monitor hate speech on social media involve the use of web crawlers to collect possible instances of hate speech based on specific keywords. Human monitors then have to analyze the collected data to determine instances that are actually hate speech. This human analysis is not only time consuming and overwhelming but also introduces subjective notions of what constitutes hate speech.

This research proposed the application of machine learning techniques system that analyzes the sentiments in users' messages on online pages to build a text binary classifier to detect hate speech.

LIST OF ANCRYNOMS

NCIC - National Cohesion and Integration Commission

IT - information technology

PEV - Post Election Violence

SA - sentiments analysis

DECLARATION

This research proposal is my original work and to the best of my knowledge, it has not been presented for academic award in any other university.

....

WILFRED WAKANYA (BOBIT/NRB/4181/16) **DATE**

CANDIDATE

This project proposal has been submitted for examinations with my approval as the university supervisor.

....

Name **Date**

Lecturer

Department of Business Administration and Information Technology

Faculty of Information Technology

Nairobi Campus

ACKNOWLEDGEMENT

My greatest appreciation goes to the almighty God who gave me the strength and hope to do this business plan. I would also like to thank my lecturer Mrs Charity Makau for her continued guidance regarding preparation of the business plan. Last but not least, the appreciation goes to the family members and friends who tirelessly worked day and night to see that this business plan is completed.

Contents

SENTIMENT ANALYSIS FOR HATE SPEECH ACTIVITY DETECTION ONLINE IN KENYAN

CHAPTER 1: INTRODUCTION

1.1 Background

Globally, there is no consensus on the meaning of the term hate speech. Researchers have tried to define hate speech as speech which either promotes acts of violence or creates an environment of prejudice that may eventually result in actual violent acts against a group of people. Speech in this sense includes any kind of expression including pictures and videos \cite {SambuliN.MoraraF. &Mahihu2013}.

The term sentiments analysis is used to mean a process where a piece of writing is analyzed and determine the opinion on the writing. Is the writing positive, negative or neutral? It is simply deriving the opinion or attitude of a speaker.

\cite {Cohen-Almagor2011} defines hate speech as hateful comments towards a person or group of people based on inherent attributes such as gender, ethnicity, and color among others. The definition of hate speech in Kenya, emphasizes on the use of hateful words with an intention to bring about ethnic hatred, where ethnic hatred is defined as hatred against a group of people based on their color, race, nationality or ethnic origins, \cite {National Council for Law Reporting.2008}.

There exists a strong relationship between hate speech and actual hate crime \cite {WaseemZ. & Hovy}. Widely propagated hate speech can easily result into incitement and consequent escalation into actual acts of violence against a group of people. This was clearly witnessed in the 2007-2008 Post Election Violence (PEV) in Kenya. The 2007-2008 PEV is partly blamed on widespread hate speech based on ethnic stereotypes and coded language (Hate speech was widely spread through a number of channels in the times preceding and during the PEV conflict. Such kind of speech resulted in the incitement of individuals to use violence and the galvanization of groups against one another \cite {Hirsch.2009}. This strong connection between hate speech and actual hate crime illustrates the importance of monitoring hate speech to avoid widespread incitement and potential incidents of hate crime.

1.2 Problem Statement

Monitoring hate content in traditional mainstream media such as radio and television is much easier than monitoring online hate speech content such as social media and micro blogging sites, \cite {Mugambi2017}. This is largely due to the fact that social media consists of a large amount of user generated content that would need to be monitored.

Current efforts by the NCIC to monitor hate speech on social media involve the use of web crawlers to collect text from social media platforms and human monitors to analyze the collected text. The NCIC's research department provides keywords of most frequently occurring terms in hate speech text, most of which are based on common stereotypes and coded language. Web crawlers search social media platforms collecting text matching the keywords. Once collected, human monitors have to go through all collected text to identify which ones are hate speech and which ones are not \cite{Mugambi2017}. This human processing of collected text is inadequate as the amount of content on social media is huge, significantly limiting how much a human monitor can review.

This work proposed the development of a model that applies machine learning techniques to automatically classify tweets as hate speech or not. This automatic classification will significantly improve the process of detecting hate speech on social media by reducing the amount of time and human effort required.

1.3 Objectives of the System

1.3.1 General Objectives

a) To investigate the existing techniques used in hate speech detection in social media,
b) To review the current machine learning techniques applied in hate speech detection,
c) To develop a model for hate speech detection,
d) To validate the model on online posts.

1.3.2 Specific Objectives

a) The system will have the ability to analyze sentiments and state the opinion of the data if it is a hate speech or not.
b) Help to reduce hate speech spread around the country.
c) Helps agencies to detect suspicious web pages and track them from their sources.
d) It'll be limited first to the use of authorities in Kenya who are involved in the security of the nation.
e) To provide an integrated user-friendly online activity management system
f) To manage risk and threats from online activity.

1.4 Scope of the System

This study limited its analysis to detecting hate speech on the social media platform such as Twitter, Facebook and only considered tweets expressed in English and Swahili. The use of sheng', vernacular languages, memes, audios and videos within tweets were not considered.

The system will utilize R studio as the main programming language. It will focus on the fusion of the codes that allow one to analyze the sentiments on the data from the webpage and gauge whether the words or speech is negative or the positive.

The software requirements for this system to work as it is supposed to will be Windows 10, WAMP Server, My SQL 5.6. The hardware requirements for system will be as follows Processor will at least have to be Dual Core, the Hard Disk will require a space of 50 GB and finally the Memory will need to be1GB RAM.

1.5 Justification

Hate messages disseminated online are increasingly common, largely attributed to issues of anonymity, itinerancy, permanency and cross-jurisdiction of online content \cite {UNESCO2015}. Notably, social media usage during the

PEV was not only to promote peace and justice but also as a channel for spreading of biased information, tribal prejudices and hate speech \cite {MakinenM.&Kuira.2008}

Text classification is an important technique for the handling and organization of text data with a wide range of applications in information retrieval. Currently, NCIC human monitors have to sift through numerous online content to identify hate speech in social media. This human analysis is overwhelming, time consuming and introduces personal interpretation of what is considered as hate speech. Text classification would enable categorization of the huge amounts of online data into hate speech or non-hate speech text, significantly reducing the amount of data that human monitors have to review, making the process of hate speech detection faster.

CHAPTER 2: LITERATURE REVIEW

2.0 Introduction

The nature of hate speech in Kenya and current processes to monitor hate speech on social media is reviewed. Significant and relevant publications and research are further reviewed to understand the application of machine learning techniques in text classification. A conceptual framework is then presented at the completion of the literature review.

2.1 Hate Speech in Kenya

The National Cohesion and Integration Commission (NCIC) was instituted as a consequent of the 2007-2008 PEV to oversee and monitor content in media such as radio, television, mobile phones and television in a bid to govern hate speech \cite {National Council for Law Reporting.2008}. According to the NCIC, a statement does not amount to hate speech unless it: causes hatred, makes a group or community look inferior, makes a community or group be viewed with contempt, degrades a group or community, or dehumanizes a group or community \cite {Commission2011}. To be quantified as hate speech, the statement should contain: threatening, abusive or insulting messages, sometimes using coded language. These messages must be directed towards a targeted group and intended to stir hatred based on the group's identity including: ethnicity, race, colour or any other national origin \cite {Commission2011}.

2.2 Hate Speech Detection

Hate speech in Kenyan online forums has unfortunately become a common occurrence with the growth of the internet, social media and mobile computing in the recent past. Social media has created a new space for the dissemination of hate speech. Since 200/, the NCIC, Kenyan civil society as well as police authorities have put measures to monitor hate speech on traditional mainstream media but hate speech on social media remains to hardly monitored \cite {SambuliN.MoraraF. &Mahihu2013}. However, more recently NCIC have put effort into monitoring hate speech on social media through the use of web crawlers.

While investigating and monitoring hate speech, investigators must take into consideration five key aspects: context, ripple effect, fear, possible retaliation and violence (National Cohesion and Integration Commission, 2011). A statement can be considered hate speech in one context but not in another. Additionally, the same statement might have different levels of impact depending on the context, for example ethnic statements may have a higher impact in political environments than social settings. The second aspect, ripple effect, and third effect, fear mean that the statement should cause some discomfort and fear amongst members of the group being targeted, respectively. The fourth aspect, possible retaliation means that the statement should provoke counterattacks and finally the statement promotes acts of violence or hate crimes (National Cohesion and Integration Commission, 2013).

2.3 Online Platform

Twitter, despite being a popular social media platform, is famously known for its cruelty in how people vent out their emotions from politicians sharing their political stances to people sharing about their normal everyday lives.

Analyzing Tweets makes it easy to understand what people think, be it good or bad concerning a particular topic of conversation or tweet. One may be curious about what people think about a personality media, trending topics or about the political atmosphere ratings. If people are tweeting about this topic, then Analyzing Tweets can help you categorize that conversation \cite {Mejova2009}.

Analyzing Tweets is especially treasured when there is a large amount of tweets around a subject. An example of this would be to analyze a hashtag like #KenyansOnTwitter. Using Analyze Tweets, I might be able to see which tweets are reacting positively or negatively to. Given the real-time nature of Twitter, Analyze Tweets lets you tap into what's going on in real-time \cite {Mukherjee2012}.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

Research can be defined as the process of systematically solving problems \cite {Bhatnagar M. &Singh2013}. This section describes the various methods and procedures that were adopted in carrying out the research.

This research proposal focuses on the case study approach of social media platforms such as twitter that use sentiments analysis \cite {Mejova2009}. This research method is suitable for this case study because it displays clearly the use of the systems that are required to build a sentiments analysis system and how they detect the key words within the websites.

3.2 System Development Methodology

3.2.1 Data Collection

Interviews were used to gain additional insight on the techniques currently used by NCIC to detect hate speech on social media, to determine the user requirements of a system to detect hate speech on twitter, and to provide further guidelines on the type of keywords to be used in the mining of twitter.

3.3 Requirements Analysis

This research aimed at developing a model for monitoring hate speech on twitter. Based on this objective, this section outlines the various requirements to be provided for by the proposed solution.

3.3.1 Functional requirements

i. The application should display to the user the tweets labelled as hate speech.
ii. The application should allow a user to enter keywords to be used as search parameters.
iii. The application should classify the tweet as hate speech or not hate speech.
iv. The application should retrieve tweets from Twitter using the Twitter Search API matching the keywords specified by the user.

3.3.2 Non-Functional Requirements

i. Usability **-** The intended users of the proposed solution are the Information and Communication Technology (ICT) staff at NCIC. It is intended that the interaction between these users and the model shall be simple to allow them collect data from twitter easily and view prediction results.

ii. Availability - is a non- functional of the system by ensuring that the system function all the time when and not needed to ensure that the data stored in the system is ready all the time.

iii. Scalability - If there is an increase in the amount of twitter posts matching user keywords searched, the proposed solution should be able to handle the extra load, collecting the tweets and predicting their labels without breaking down.

3.4 System Architecture

The system architecture shows the general layout of the twitter hate speech monitoring prototype and the components it is made up of. The hate speech detection process begins with the user entering a keyword to be used to retrieve matching tweets. The tweets collector module receives the keyword and collects tweets matching the keyword from the Twitter Search API and stores them in a database.

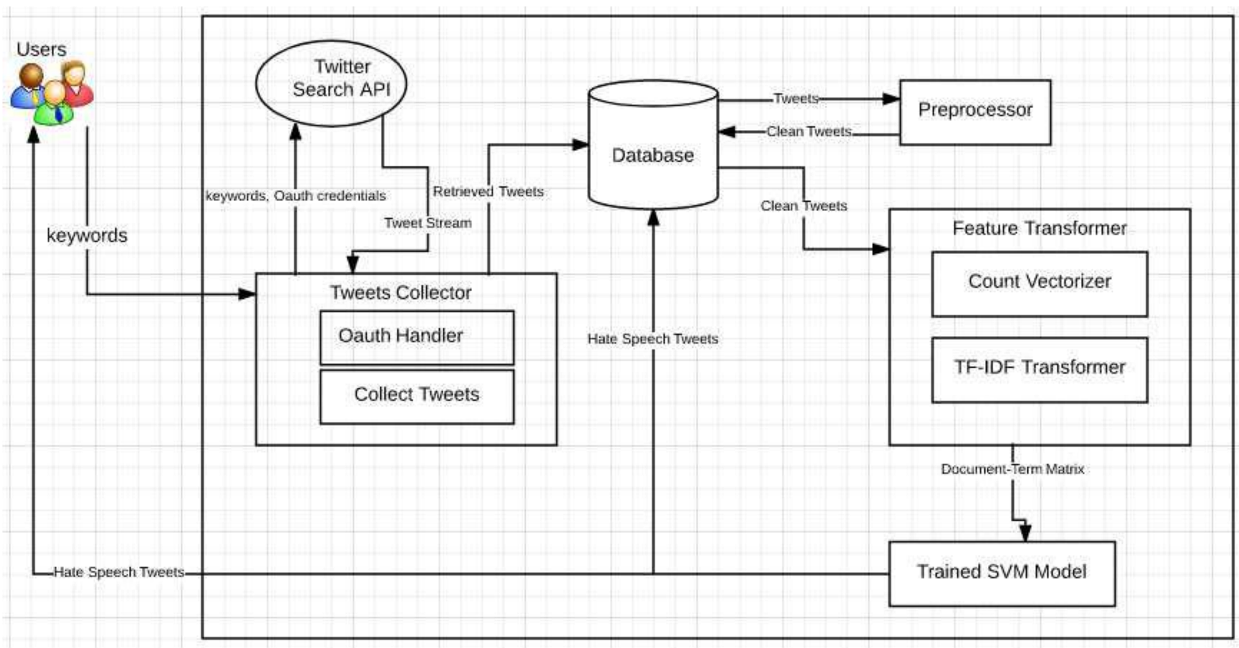


Figure 1. System Architecture

3.4.1 Use Case Diagram

Use case diagrams are used to illustrate interaction between actors and the system. Figure 1. illustrates these interactions between the various actors and the proposed hate speech detection prototype. The diagram also depicts the functionality that the proposed system should have.

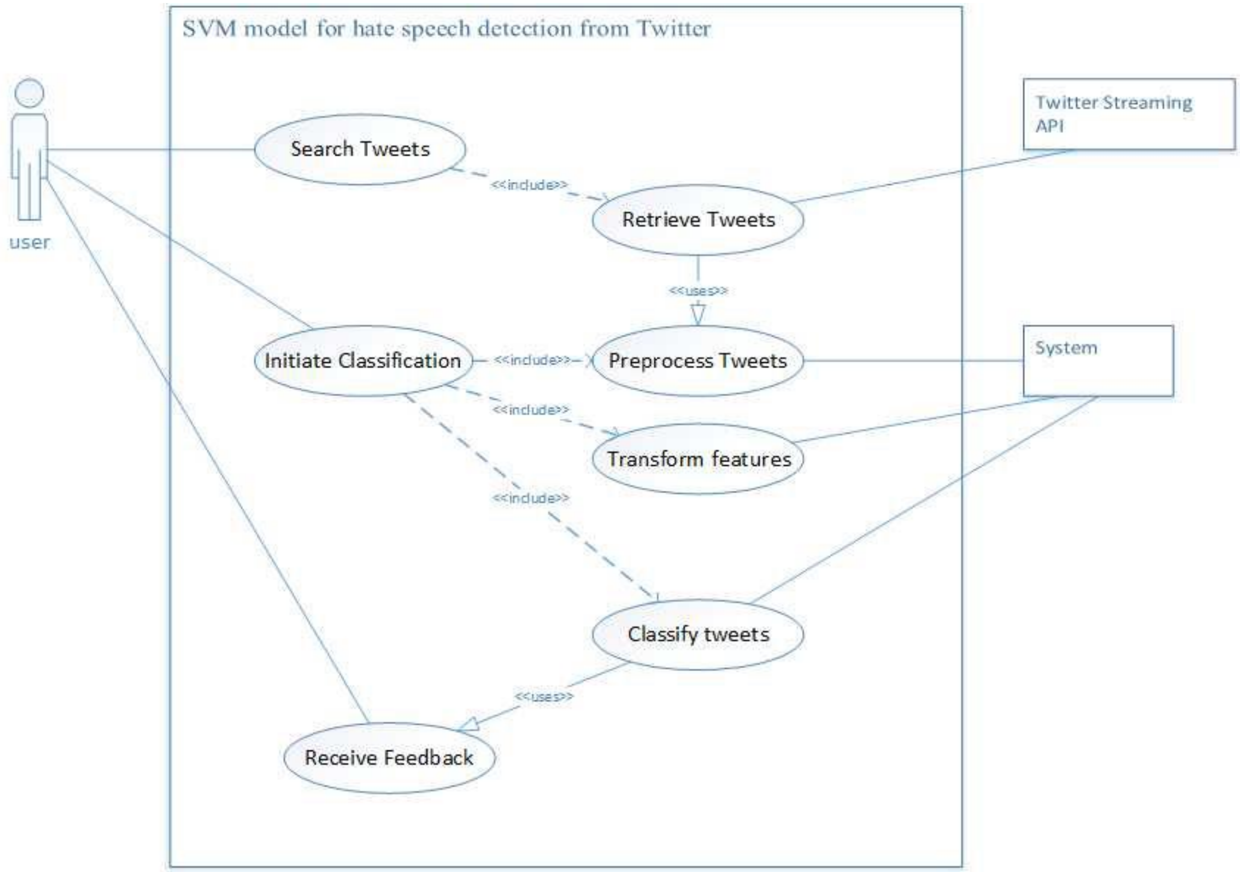


Figure 2: Use Case Diagram

3.4.2 Context Diagram

The context diagram as depicted in Figure 4.4 illustrates the boundary of the prototype, its environment and the entities that interact with it. It also shows the various inputs and outputs from the prototype to the entities. The main entities interacting with the proposed prototype are a user and the Twitter Search API.

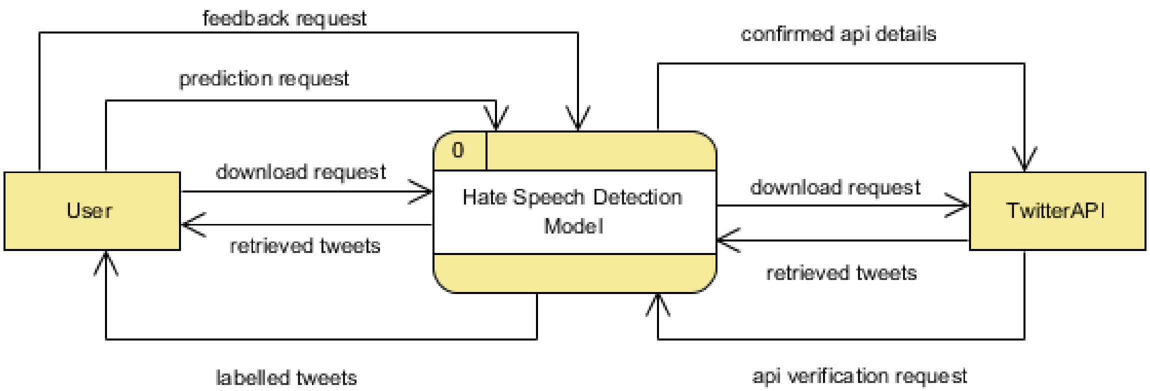


Figure 3: Context Diagram

CHAPTER 4: SYSTEM IMPLEMENTATION

Sentiments Analysis System To detect hate speech online was designed to allow authorities to scan online websites as well as emails for any insightful words or activity by mostly politicians. The system would be in the custody of the authorities who are involved in the maintain political and economic stability or the department involved in country security those that protect against outside threats. The architecture in the software consisted of the database and application program. The system was implemented using R studio which has a storage of the document that will be used for the analysis, then at this point then one can run the code that is presented and simply run the entire program.

4.1 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.

Slam : Data structures and algorithms for sparse arrays and matrices, based on index arrays and simple triplet representations, respectively.

SnowballC : An R interface to the C 'libstemmer' library that implements Porter's word stemming algorithm for collapsing words to a common root to aid comparison of vocabulary.

NLP : By human language, we're simply referring to any language used for everyday communication. It refers to the way we communicate to each other using speech and text.

Syuzhet : The package attempts to reveal the latent structure of narrative by means of sentiment analysis.

StreamR : This package provides a series of functions that allow R users to access Twitter's filter, sample, and user streams, and to parse the output into data frames.

Httr : The httr package makes it easy to talk to web APIs from R.

Dplyr : dplyr is a new package which provides a set of tools for efficiently manipulating datasets in R. dplyr is the next iteration of plyr, focussing on only data frames.

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.

4.2 Twitter Analysis:

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 application are:

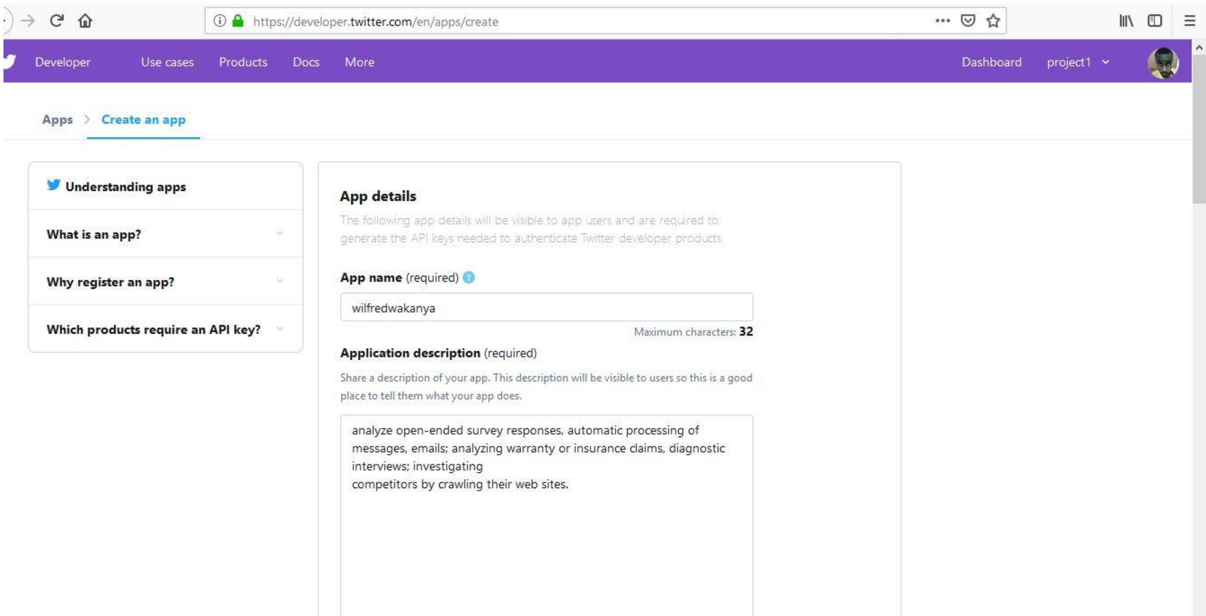


Figure 4: Twitter API

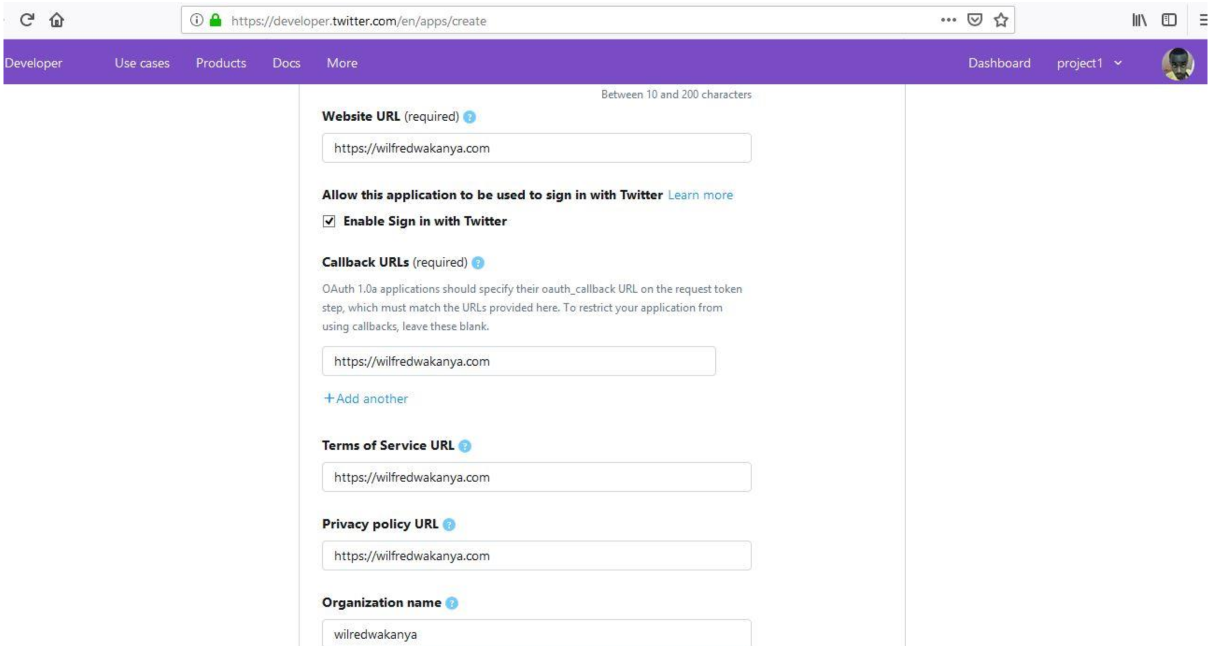


Figure 4.1: Twitter API

Give your application a name, describe your application in a few words, provide your website's URL or your address in case you do not 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.

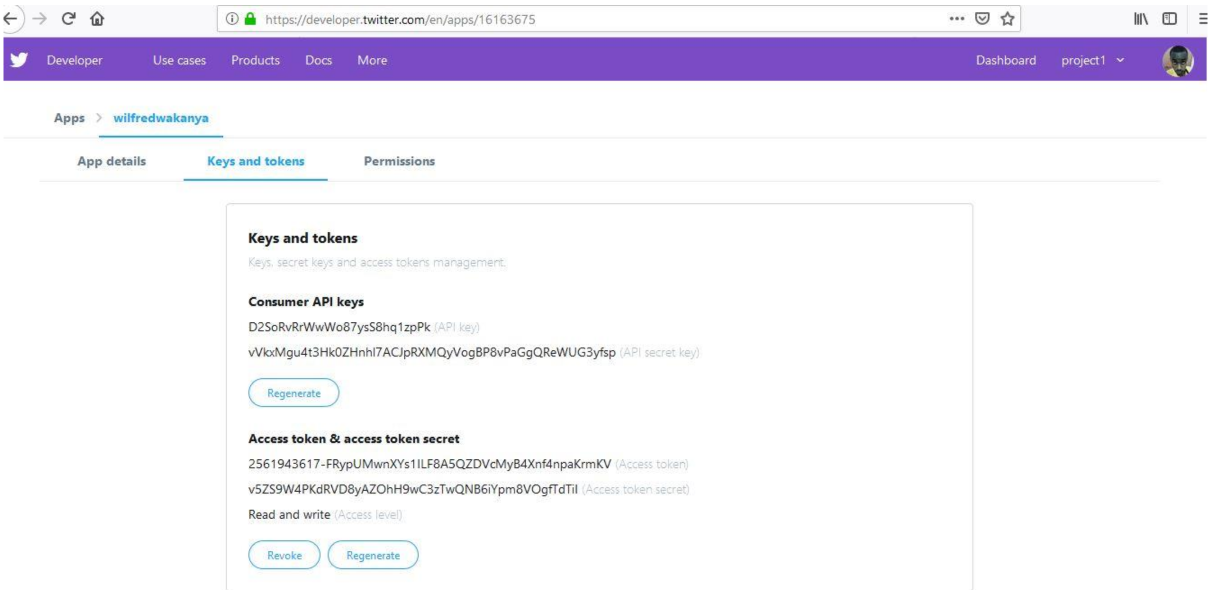
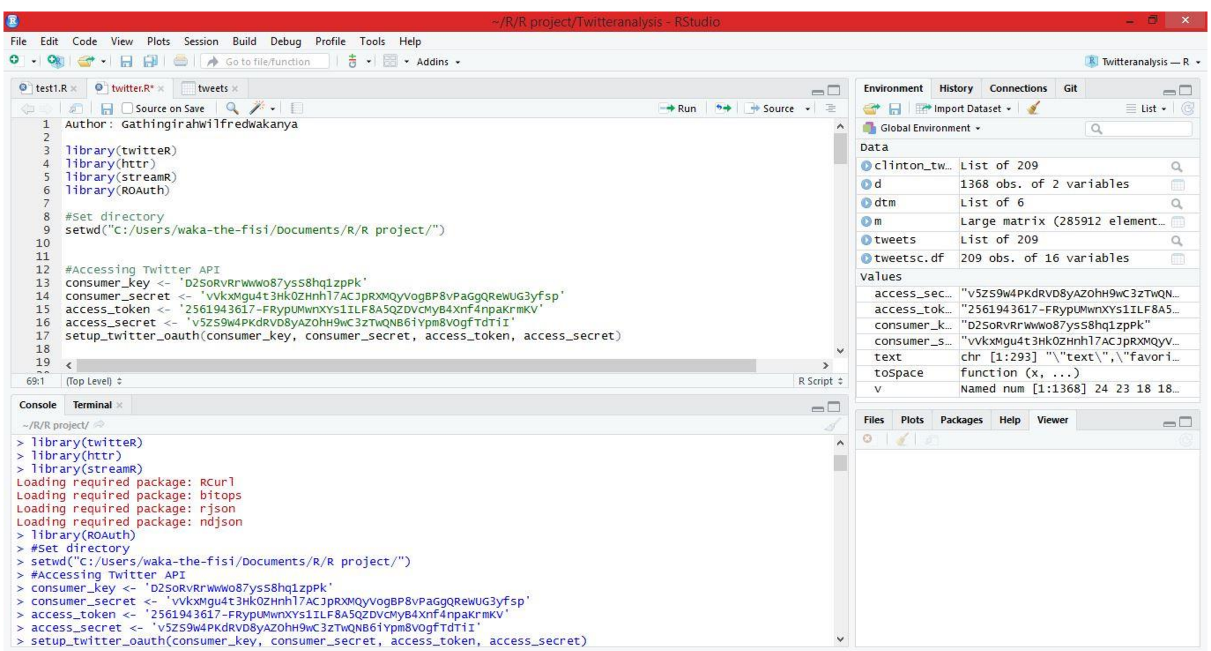
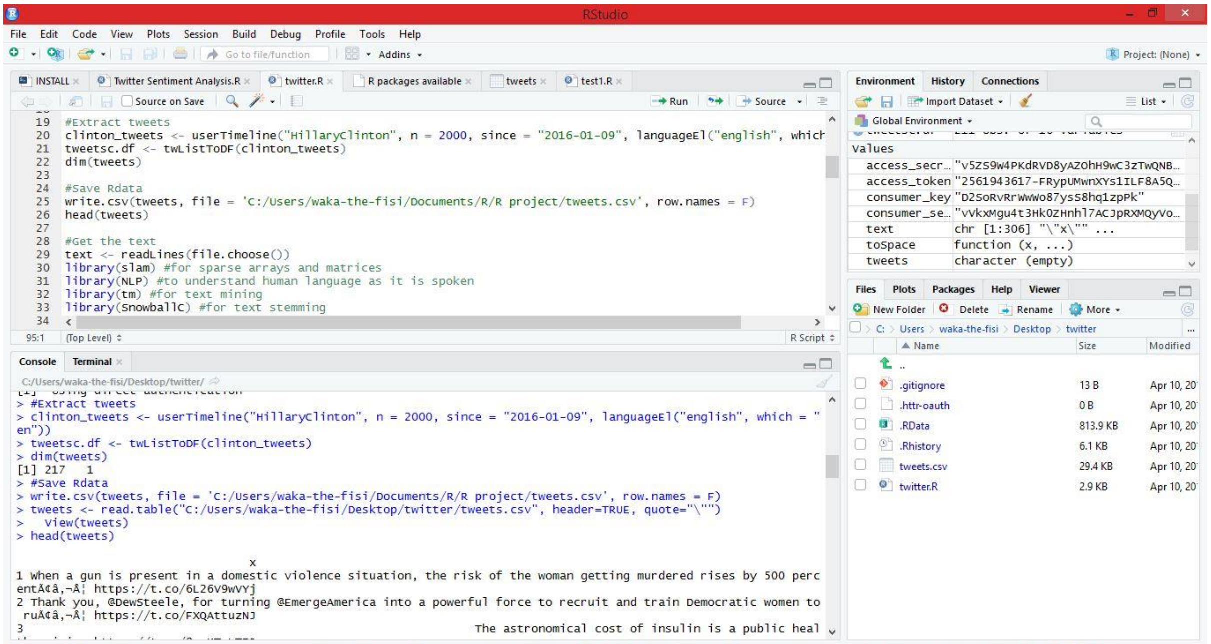


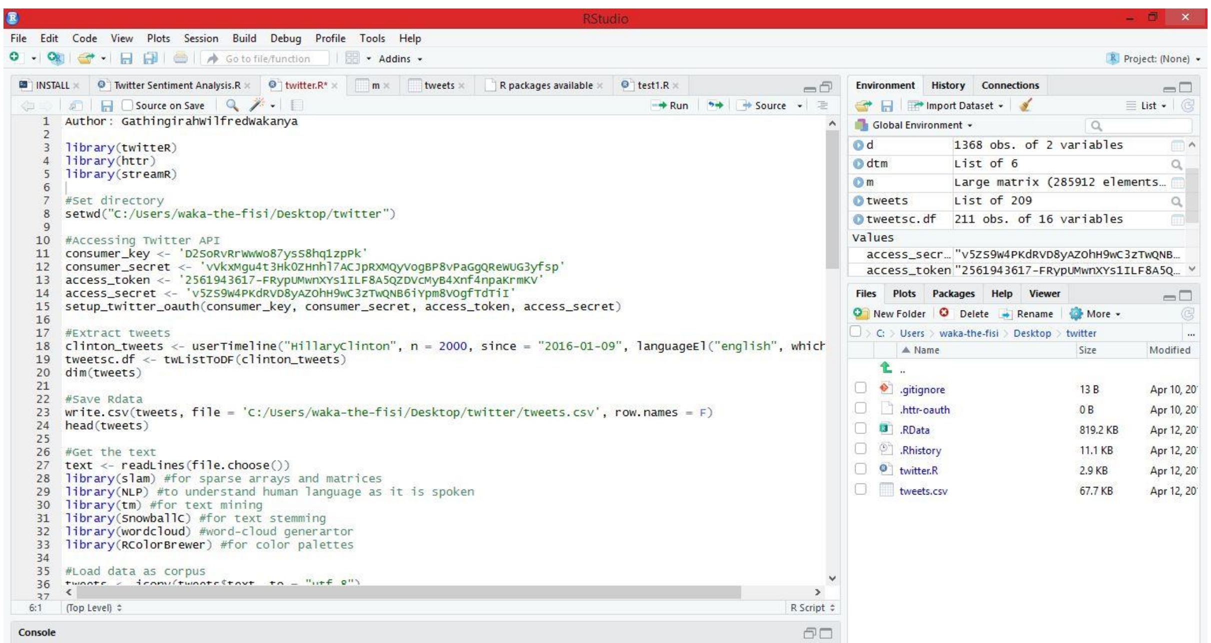
Figure 4.3: Twitter Keys and Tokens

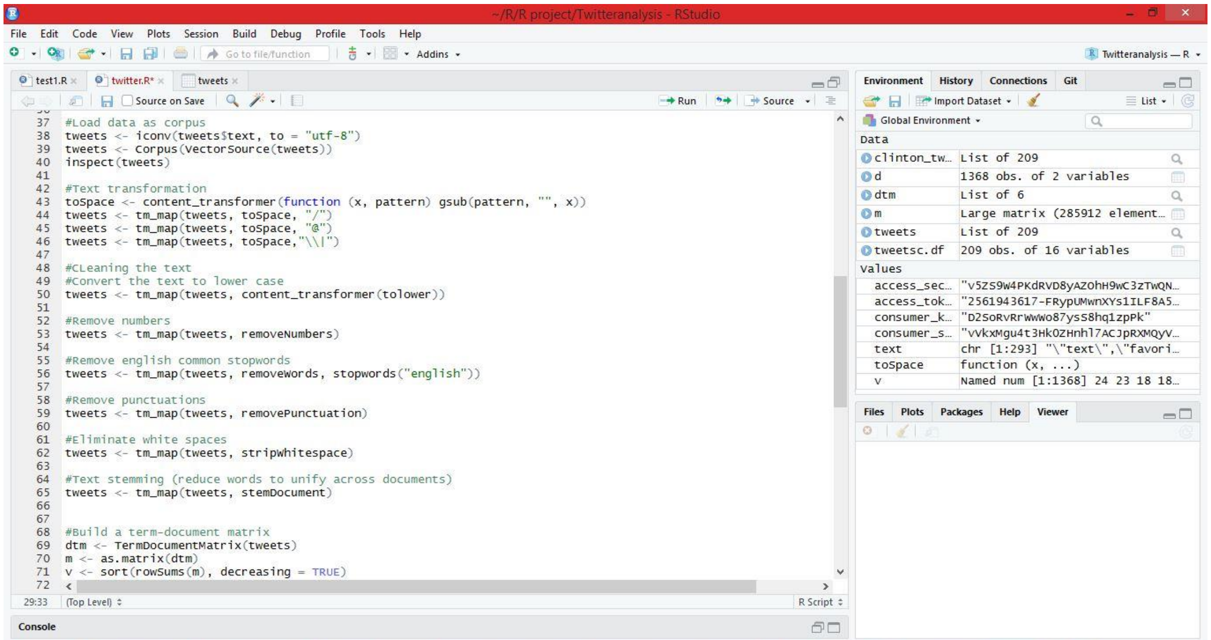
Once this step is done. Next, I will work on my RStudio.

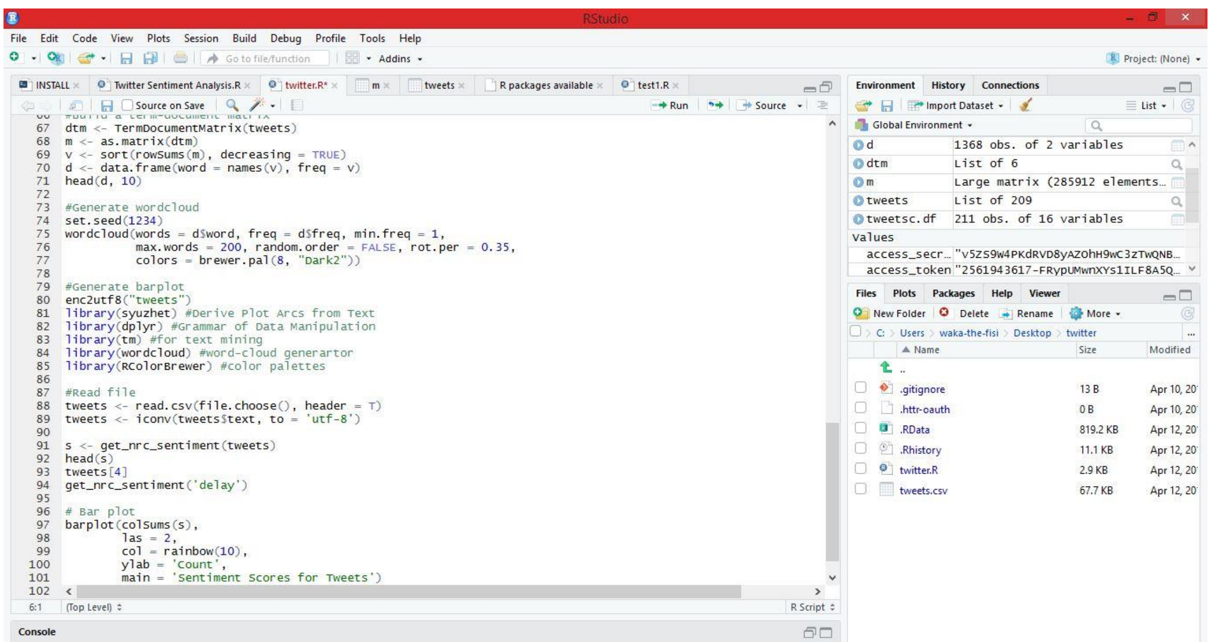
Figure 4.4: SCREENSHOTS OF R STUDIO WITH CODES FOR SENTIMENTS ANALYSIS

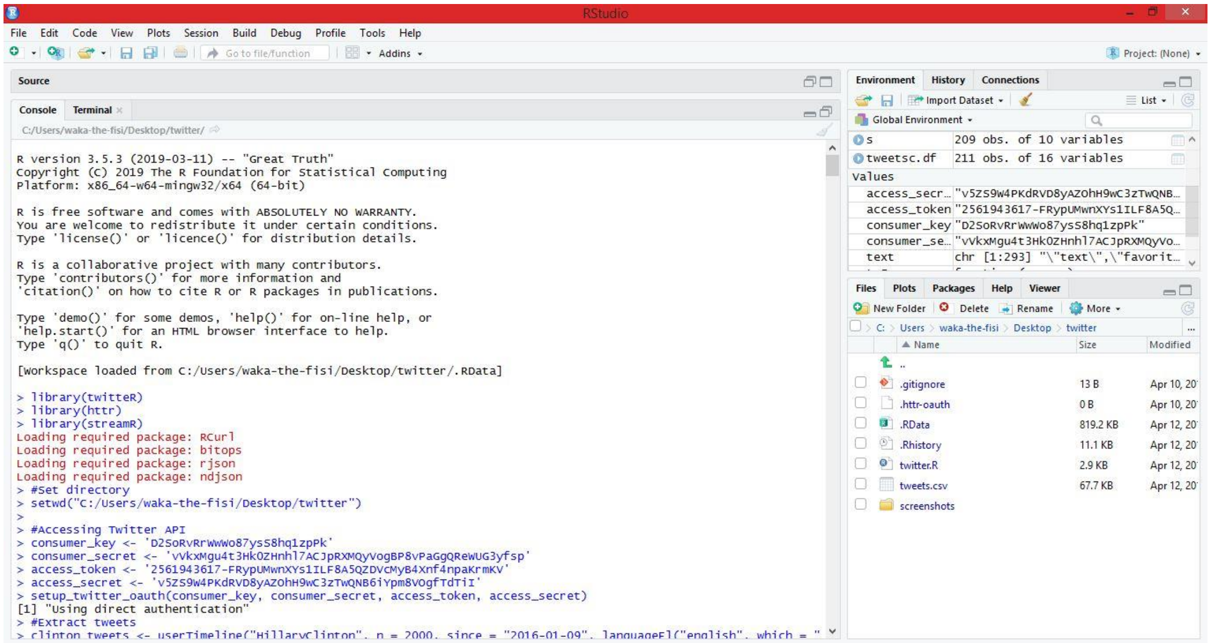


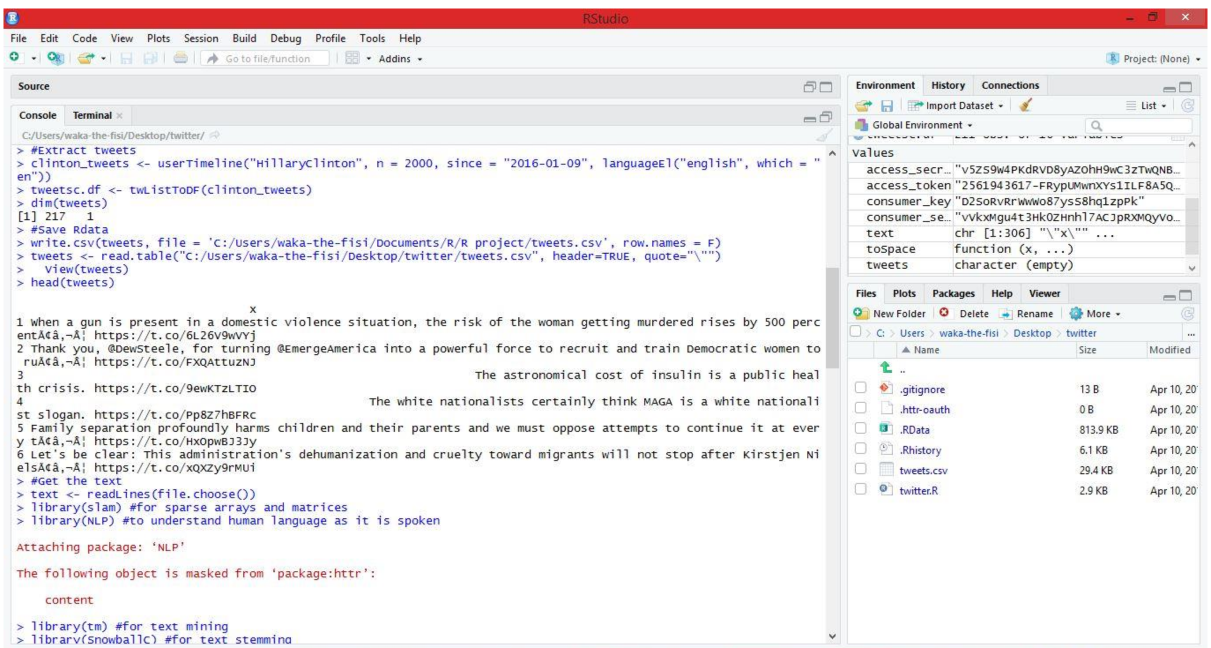

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
#Extract tweets
clinton_tweets <- userTimeline("HillaryClinton", n = 2000, since = "2016-01-09", languageEl("english", which
tweetsc.df <- twListToDF(clinton_tweets)
dim(tweets)
#Save Rdata
write.csv(tweets, file = 'C:/Users/waka-the-fisi/Documents/R/R project/tweets.csv', row.names = F)
head(tweets)
#Get the text
text <- readLines(file.choose())
library(slam) #for sparse arrays and matrices
library(NLP) #to understand human language as it is spoken
library(tm) #for text mining
library(SnowballC) #for text stemming
Console Terminal
C:/Users/waka-the-fisi/Desktop/twitter/
> #Extract tweets
> clinton_tweets <- userTimeline("HillaryClinton", n = 2000, since = "2016-01-09", languageEl("english", which = "en"))
> tweetsc.df <- twListToDF(clinton_tweets)
> dim(tweets)
[1] 217 1
> #Save Rdata
> write.csv(tweets, file = 'C:/Users/waka-the-fisi/Documents/R/R project/tweets.csv', row.names = F)
> tweets <- read.table("C:/Users/waka-the-fisi/Desktop/twitter/tweets.csv", header=TRUE, quote="\"")
> View(tweets)
> head(tweets)
1 When a gun is present in a domestic violence situation, the risk of the woman getting murdered rises by 500 perc
2 Thank you, @DewSteele, for turning @EmergeAmerica into a powerful force to recruit and train Democratic women to
The astronomical cost of insulin is a public heal
Environment History Connections
Global Environment
Values
access_secr… "v5ZS9W4PKdRVD8yAZOhH9wC3zTWQNB…
access_token "2561943617-FRypUMwnXYS1ILF8A5Q…
consumer_key "D2SoRvRrWwWo87ysS8hq1zpPk"
consumer_se… "vvkxMgu4t3Hk0ZHnhl7ACJpRXMQyVo…
text chr [1:306] "\"x\"" …
toSpace function (x, …)
tweets character (empty)
Files Plots Packages Help Viewer
.gitignore 13 B
.httr-oauth 0 B
.RData 813.9 KB
.Rhistory 6.1 KB
tweets.csv 29.4 KB
twitter.R 2.9 KB

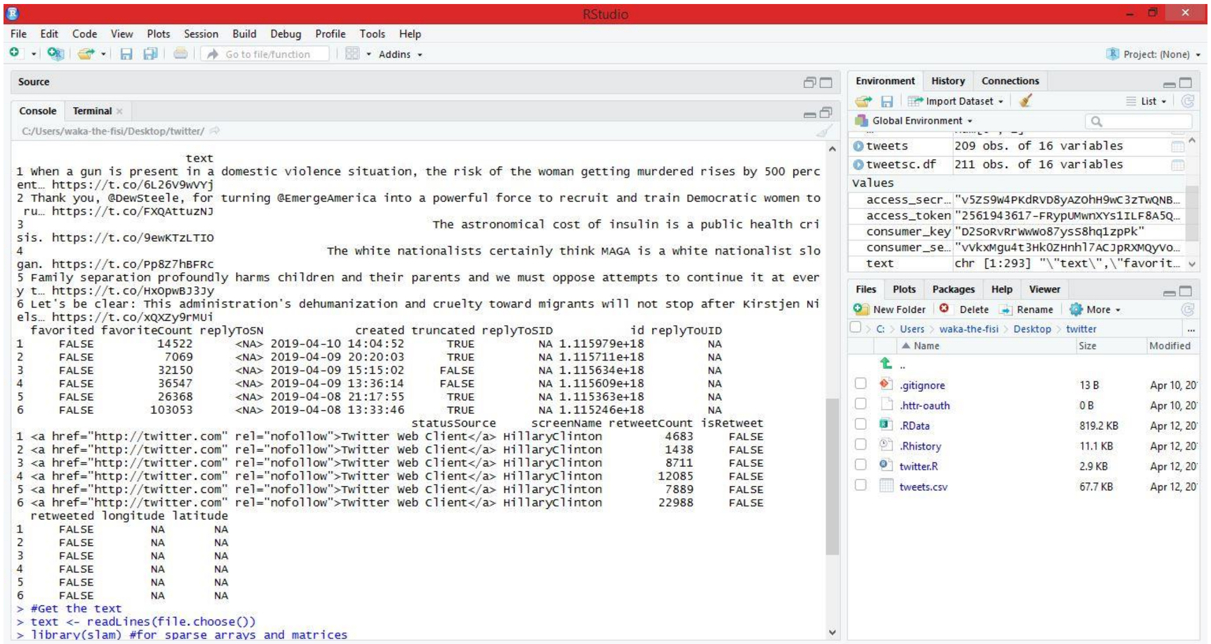

RStudio
Author: Gathingirahwilfredwakanya
library(twitteR)
library(httr)
library(streamR)
#Set directory
setwd("C:/Users/waka-the-fisi/Desktop/twitter")
#Accessing Twitter API
consumer_key <- 'D2SoRvRrWwWo87ysS8hq1zpPk'
consumer_secret <- 'vvkxMgu4t3Hk0ZHnhl7ACJpRXMQyVogBP8vPaGgQReWUG3yfsp'
access_token <- '2561943617-FRypUMwnXYS1ILF8A5QZDVcMyB4Xnf4npaKrmKV'
access_secret <- 'v5ZS9W4PKdRVD8yAZOhH9wC3zTWQNB6iYpm8VOgfTdTiI'
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
#Extract tweets
clinton_tweets <- userTimeline("HillaryClinton", n = 2000, since = "2016-01-09", languageEl("english", which
tweetsc.df <- twListToDF(clinton_tweets)
dim(tweets)
#Save Rdata
write.csv(tweets, file = 'C:/Users/waka-the-fisi/Desktop/twitter/tweets.csv', row.names = F)
head(tweets)
#Get the text
text <- readLines(file.choose())
library(slam) #for sparse arrays and matrices
library(NLP) #to understand human language as it is spoken
library(tm) #for text mining
library(SnowballC) #for text stemming
library(wordcloud) #word-cloud generartor
library(RColorBrewer) #for color palettes
#Load data as corpus
Console
Global Environment
d 1368 obs. of 2 variables
dtm List of 6
m Large matrix (285912 elements…
tweets List of 209
tweetsc.df 211 obs. of 16 variables
Values
access_secr… "v5ZS9W4PKdRVD8yAZOhH9wC3zTWQNB…
access_token "2561943617-FRypUMwnXYS1ILF8A5Q…
.gitignore 13 B
.httr-oauth 0 B
.RData 819.2 KB
.Rhistory 11.1 KB
twitter.R 2.9 KB
tweets.csv 67.7 KB

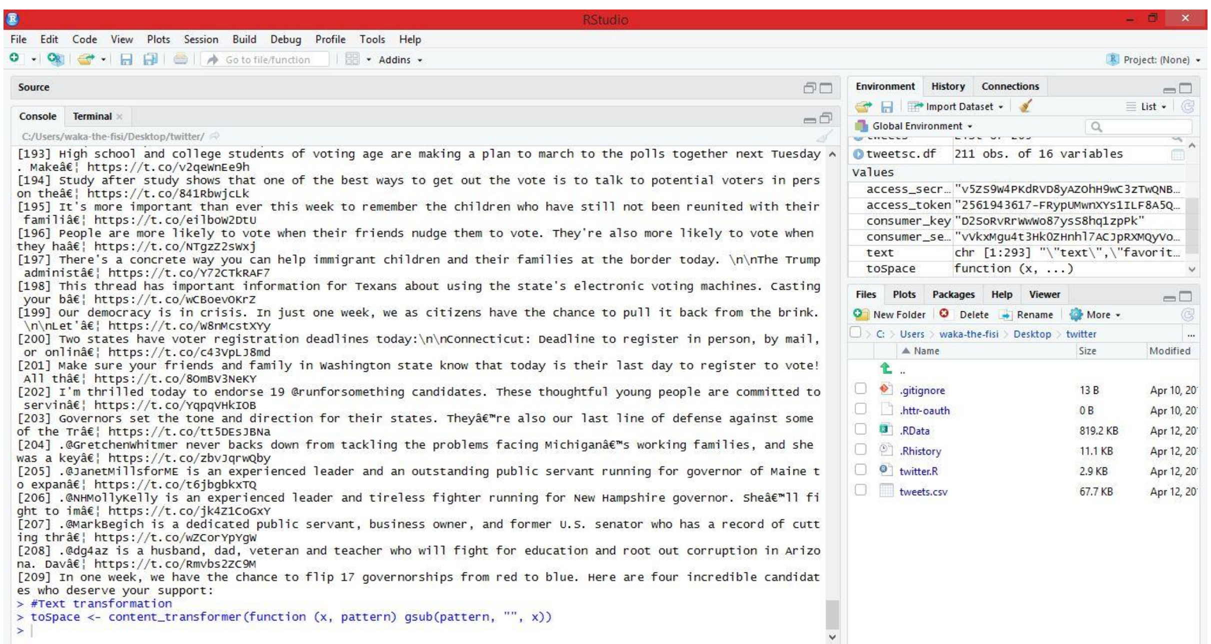

~/R/R project/Twitteranalysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
test1.R twitter.R* tweets
Source on Save Run Source
#Load data as corpus
tweets <- iconv(tweets$text, to = "utf-8")
tweets <- Corpus(VectorSource(tweets))
inspect(tweets)
#Text transformation
toSpace <- content_transformer(function (x, pattern) gsub(pattern, "", x))
tweets <- tm_map(tweets, toSpace, "/")
tweets <- tm_map(tweets, toSpace, "@")
tweets <- tm_map(tweets, toSpace,"\\|")
#CLeaning the text
#Convert the text to lower case
tweets <- tm_map(tweets, content_transformer(tolower))
#Remove numbers
tweets <- tm_map(tweets, removeNumbers)
#Remove english common stopwords
tweets <- tm_map(tweets, removeWords, stopwords("english"))
#Remove punctuations
tweets <- tm_map(tweets, removePunctuation)
#Eliminate white spaces
tweets <- tm_map(tweets, stripWhitespace)
#Text stemming (reduce words to unify across documents)
tweets <- tm_map(tweets, stemDocument)
#Build a term-document matrix
dtm <- TermDocumentMatrix(tweets)
m <- as.matrix(dtm)
v <- sort(rowSums(m), decreasing = TRUE)
Environment History Connections Git
Import Dataset
Global Environment
Data
clinton_tw… List of 209
d 1368 obs. of 2 variables
dtm List of 6
m Large matrix (285912 element…
tweets List of 209
tweetsc.df 209 obs. of 16 variables
Values
toSpace function (x, ...)
v Named num [1:1368] 24 23 18 18…
Files Plots Packages Help Viewer
Console

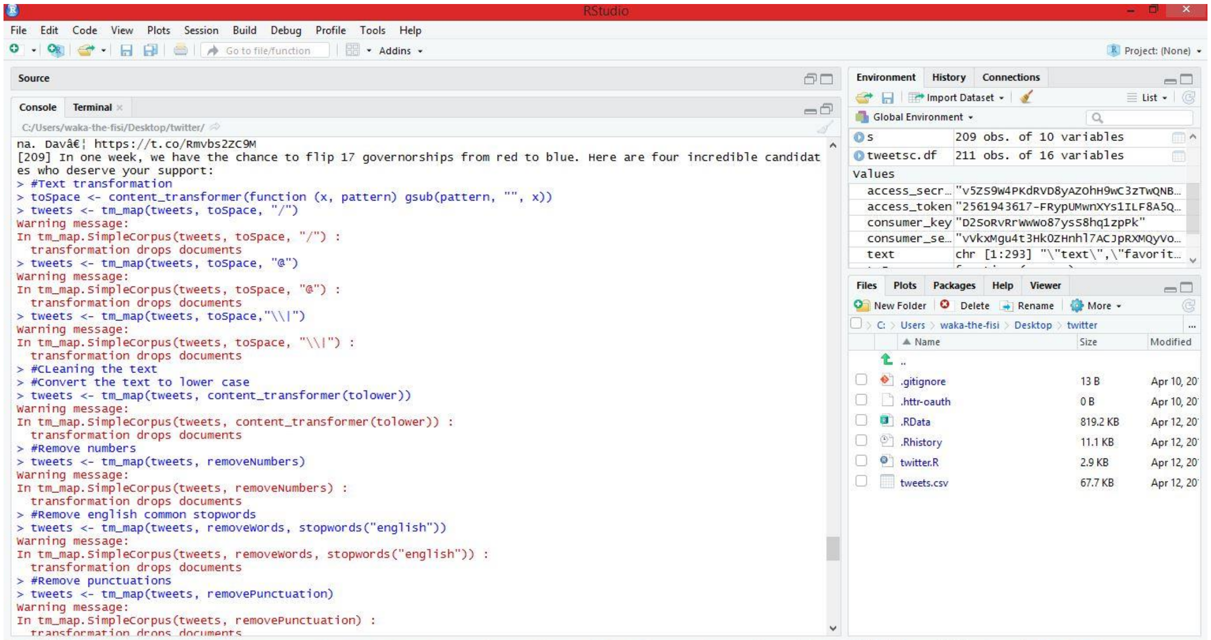

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
INSTALL Twitter Sentiment Analysis.R twitter.R* m tweets R packages available test1.R
dtm <- TermDocumentMatrix(tweets)
m <- as.matrix(dtm)
v <- sort(rowSums(m), decreasing = TRUE)
d <- data.frame(word = names(v), freq = v)
head(d, 10)
#Generate wordcloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
max.words = 200, random.order = FALSE, rot.per = 0.35,
colors = brewer.pal(8, "Dark2"))
#Generate barplot
enc2utf8("tweets")
library(syuzhet) #Derive Plot Arcs from Text
library(dplyr) #Grammar of Data Manipulation
library(tm) #for text mining
library(wordcloud) #word-cloud generartor
library(RColorBrewer) #color palettes
#Read file
tweets <- read.csv(file.choose(), header = T)
tweets <- iconv(tweets$text, to = 'utf-8')
s <- get_nrc_sentiment(tweets)
head(s)
tweets[4]
get_nrc_sentiment('delay')
# Bar plot
barplot(colSums(s),
las = 2,
col = rainbow(10),
ylab = 'Count',
main = 'Sentiment Scores for Tweets')
Environment History Connections
Global Environment
d 1368 obs. of 2 variables
dtm List of 6
m Large matrix (285912 elements…
tweets List of 209
tweetsc.df 211 obs. of 16 variables
Values
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users waka-the-fisi Desktop twitter
Name Size Modified
.gitignore 13 B
.httr-oauth 0 B
.RData 819.2 KB
.Rhistory 11.1 KB
twitter.R 2.9 KB
tweets.csv 67.7 KB
Console

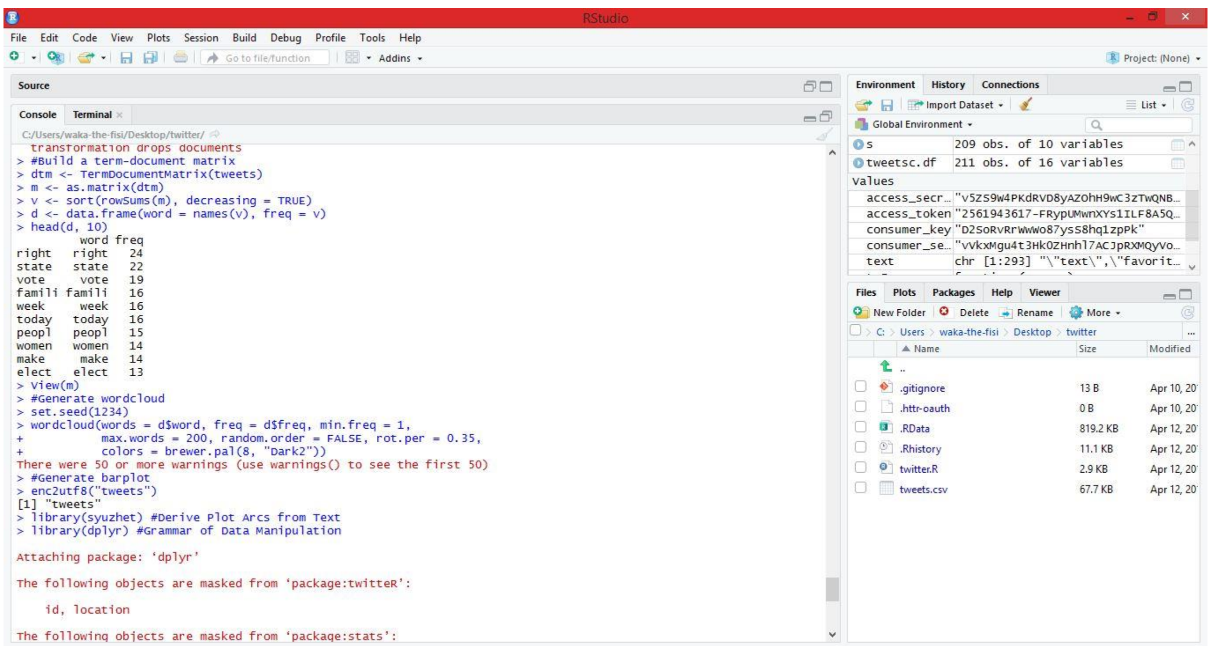

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console Terminal
C:/Users/waka-the-fisi/Desktop/twitter/
R version 3.5.3 (2019-03-11) -- "Great Truth"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from C:/Users/waka-the-fisi/Desktop/twitter/.RData]
> library(twitteR)
> library(httr)
> library(streamR)
Loading required package: RCurl
Loading required package: bitops
Loading required package: rjson
Loading required package: ndjson
> #Set directory
> setwd("C:/Users/waka-the-fisi/Desktop/twitter")
>
> #Accessing Twitter API
> consumer_key <- 'D2SoRvRrWwWo87ysS8hq1zpPk'
> consumer_secret <- 'vVkxMgu4t3Hk0zHnhl7ACJpRXMQyVogBP8vPaGgQReWUG3yfsp'
> access_token <- '2561943617-FRypUMwnXYs1ILF8A5QZDVcMyB4Xnf4npaKrmKV'
> access_secret <- 'v5ZS9W4PKdRVD8yAZOhH9wC3zTWQNB6iYpm8VOgfTdTiI'
> setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
[1] "Using direct authentication"
> #Extract tweets
> clinton_tweets <- userTimeline("HillaryClinton", n = 2000, since = "2016-01-09", languageEl("english", which = "
Environment History Connections
Import Dataset
List
Global Environment
s 209 obs. of 10 variables
tweetsc.df 211 obs. of 16 variables
Values
access_secr… "v5ZS9W4PKdRVD8yAZOhH9wC3zTWQNB…
access_token "2561943617-FRypUMwnXYs1ILF8A5Q…
consumer_key "D2SoRvRrWwWo87ysS8hq1zpPk"
consumer_se… "vVkxMgu4t3Hk0zHnhl7ACJpRXMQyVo…
text chr [1:293] "\"text\",\"favorit…
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users waka-the-fisi Desktop twitter
Name Size Modified
.gitignore 13 B Apr 10, 20
.httr-oauth 0 B Apr 10, 20
.RData 819.2 KB Apr 12, 20
.Rhistory 11.1 KB Apr 12, 20
twitter.R 2.9 KB Apr 12, 20
tweets.csv 67.7 KB Apr 12, 20
screenshots

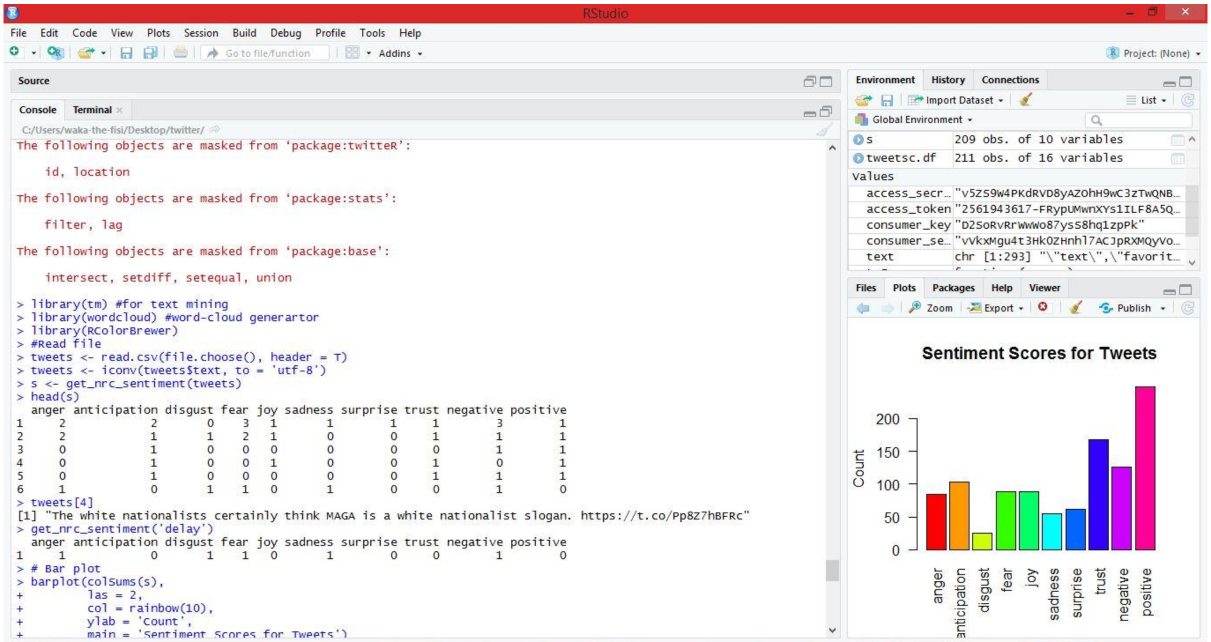

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console Terminal
C:/Users/waka-the-fisi/Desktop/twitter/
> #Extract tweets
> clinton_tweets <- userTimeline("HillaryClinton", n = 2000, since = "2016-01-09", languageEl("english", which = "en"))
> tweetsc.df <- twListToDF(clinton_tweets)
> dim(tweets)
[1] 217 1
> #Save Rdata
> write.csv(tweets, file = 'C:/Users/waka-the-fisi/Documents/R/R project/tweets.csv', row.names = F)
> tweets <- read.table("C:/Users/waka-the-fisi/Desktop/twitter/tweets.csv", header=TRUE, quote="\"")
> View(tweets)
> head(tweets)
x
1 When a gun is present in a domestic violence situation, the risk of the woman getting murdered rises by 500 percentÃ¢â,¬Â¦ https://t.co/6L26V9wVYj
2 Thank you, @DewSteele, for turning @EmergeAmerica into a powerful force to recruit and train Democratic women to ruÃ¢â,¬Â¦ https://t.co/FXQAttuzNJ
3 The astronomical cost of insulin is a public health crisis. https://t.co/9ewKTzLTIO
4 The white nationalists certainly think MAGA is a white nationalist slogan. https://t.co/Pp8z7hBFRc
5 Family separation profoundly harms children and their parents and we must oppose attempts to continue it at every tÃ¢â,¬Â¦ https://t.co/HxOpwBJ3Jy
6 Let's be clear: This administration's dehumanization and cruelty toward migrants will not stop after Kirstjen NielsÃ¢â,¬Â¦ https://t.co/xQXZy9rMUi
> #Get the text
> text <- readLines(file.choose())
> library(slam) #for sparse arrays and matrices
> library(NLP) #to understand human language as it is spoken
Attaching package: 'NLP'
The following object is masked from 'package:httr':
content
> library(tm) #for text mining
> library(SnowballC) #for text stemming
Environment History Connections
Import Dataset
List
Global Environment
Values
access_secr… "v5ZS9W4PKdRVD8yAZOhH9wC3zTWQNB…
access_token "2561943617-FRypUMwnXYs1ILF8A5Q…
consumer_key "D2SoRvRrWwWo87ysS8hq1zpPk"
consumer_se… "vVkxMgu4t3Hk0zHnhl7ACJpRXMQyVo…
text chr [1:306] "\"x\"" ...
toSpace function (x, ...)
tweets character (empty)
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users waka-the-fisi Desktop twitter
Name Size Modified
.gitignore 13 B Apr 10, 20
.httr-oauth 0 B Apr 10, 20
.RData 813.9 KB Apr 10, 20
.Rhistory 6.1 KB Apr 10, 20
tweets.csv 29.4 KB Apr 10, 20
twitter.R 2.9 KB Apr 10, 20

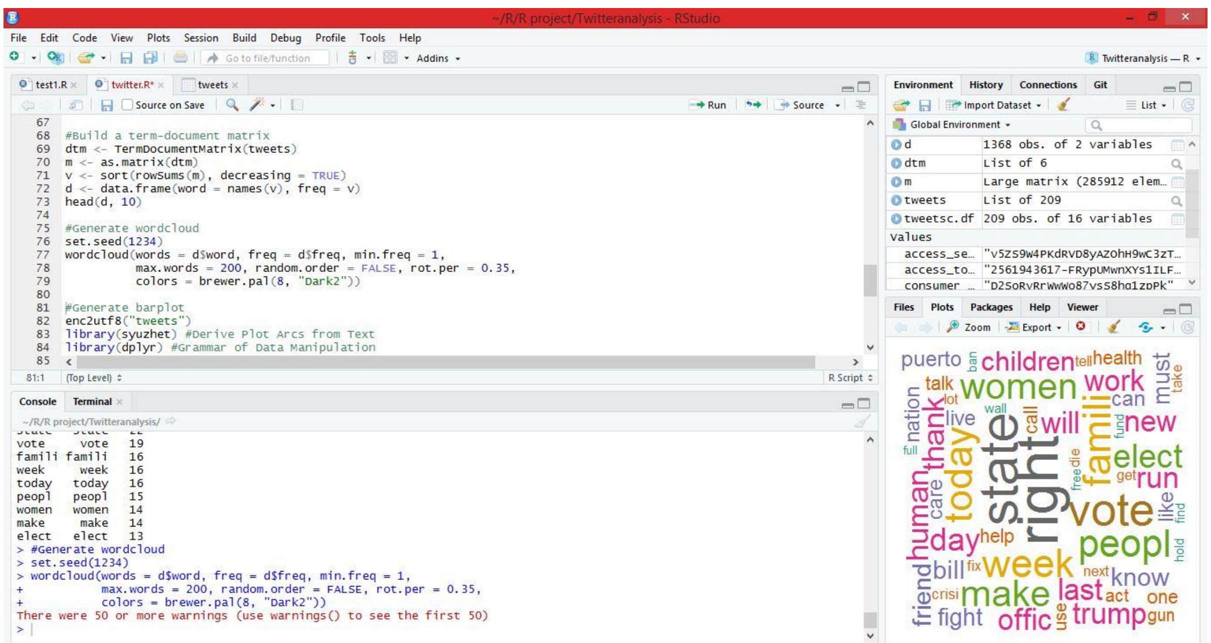

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal
C:/Users/waka-the-fisi/Desktop/twitter/
text
1 When a gun is present in a domestic violence situation, the risk of the woman getting murdered rises by 500 percent… https://t.co/6L26V9wVYj
2 Thank you, @DewSteele, for turning @EmergeAmerica into a powerful force to recruit and train Democratic women to ru… https://t.co/FXQAttuzNJ
3 The astronomical cost of insulin is a public health crisis. https://t.co/9ewKTzLTIO
4 The white nationalists certainly think MAGA is a white nationalist slogan. https://t.co/Pp8Z7hBFRc
5 Family separation profoundly harms children and their parents and we must oppose attempts to continue it at every t… https://t.co/HxOpwBJ3Jy
6 Let's be clear: This administration's dehumanization and cruelty toward migrants will not stop after Kirstjen Niels… https://t.co/XQXZy9rMUi
favorited favoriteCount replyToSN created truncated replyToSID id replyToUID
1 FALSE 14522 <NA> 2019-04-10 14:04:52 TRUE NA 1.115979e+18 NA
2 FALSE 7069 <NA> 2019-04-09 20:20:03 TRUE NA 1.115711e+18 NA
3 FALSE 32150 <NA> 2019-04-09 15:15:02 FALSE NA 1.115634e+18 NA
4 FALSE 36547 <NA> 2019-04-09 13:36:14 FALSE NA 1.115609e+18 NA
5 FALSE 26368 <NA> 2019-04-08 21:17:55 TRUE NA 1.115363e+18 NA
6 FALSE 103053 <NA> 2019-04-08 13:33:46 TRUE NA 1.115246e+18 NA
statusSource screenName retweetCount isRetweet
1 Twitter Web Client HillaryClinton 4683 FALSE
2 Twitter Web Client HillaryClinton 1438 FALSE
3 Twitter Web Client HillaryClinton 8711 FALSE
4 Twitter Web Client HillaryClinton 12085 FALSE
5 Twitter Web Client HillaryClinton 7889 FALSE
6 Twitter Web Client HillaryClinton 22988 FALSE
retweeted longitude latitude
1 FALSE NA NA
2 FALSE NA NA
3 FALSE NA NA
4 FALSE NA NA
5 FALSE NA NA
6 FALSE NA NA
> #Get the text
> text <- readLines(file.choose())
> library(slam) #for sparse arrays and matrices
Environment History Connections
Import Dataset
Global Environment
tweets 209 obs. of 16 variables
tweetsc.df 211 obs. of 16 variables
Values
access_secr… "v5ZS9W4PKdRVD8yAZOhH9wC3zTWQNB…
access_token "2561943617-FRypUMwnXYs1ILF8A5Q…
consumer_key "D2SoRvRrWwWo87ysS8hq1zpPk"
consumer_se… "vVkxMgu4t3Hk0ZHnhl7ACJpRXMQyVo…
text chr [1:293] "\"text\",\"favorit…
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users waka-the-fisi Desktop twitter
Name Size Modified
.gitignore 13 B Apr 10, 20
.httr-oauth 0 B Apr 10, 20
.RData 819.2 KB Apr 12, 20
.Rhistory 11.1 KB Apr 12, 20
twitter.R 2.9 KB Apr 12, 20
tweets.csv 67.7 KB Apr 12, 20

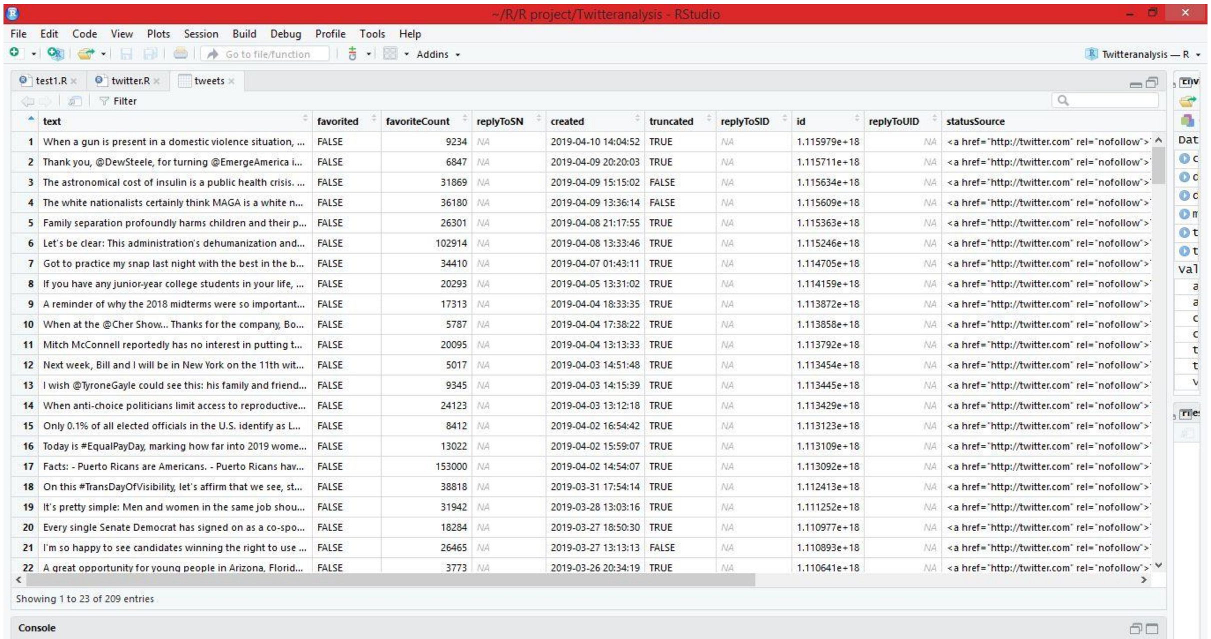

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal
C:/Users/waka-the-fisi/Desktop/twitter/
[193] High school and college students of voting age are making a plan to march to the polls together next Tuesday. Makeâ€¦ https://t.co/v2qewnEe9h
[194] Study after study shows that one of the best ways to get out the vote is to talk to potential voters in person on theâ€¦ https://t.co/841RbwjcLk
[195] It's more important than ever this week to remember the children who have still not been reunited with their familiâ€¦ https://t.co/eilbow2DtU
[196] People are more likely to vote when their friends nudge them to vote. They're also more likely to vote when they haâ€¦ https://t.co/NTgzZ2swxj
[197] There's a concrete way you can help immigrant children and their families at the border today. \n\nThe Trump administâ€¦ https://t.co/Y72CTkRAF7
[198] This thread has important information for Texans about using the state's electronic voting machines. Casting your bâ€¦ https://t.co/wCBoevOKrZ
[199] Our democracy is in crisis. In just one week, we as citizens have the chance to pull it back from the brink. \n\nLet'â€¦ https://t.co/W8nMcstXYy
[200] Two states have voter registration deadlines today:\n\nConnecticut: Deadline to register in person, by mail, or onlinâ€¦ https://t.co/c43VpLJ8md
[201] Make sure your friends and family in Washington state know that today is their last day to register to vote! All thâ€¦ https://t.co/8OmBV3NeKY
[202] I'm thrilled today to endorse 19 @runforsomething candidates. These thoughtful young people are committed to servinâ€¦ https://t.co/YqpqVHkIOB
[203] Governors set the tone and direction for their states. Theyâ€™re also our last line of defense against some of the Trâ€¦ https://t.co/tt5DEsJBNa
[204] .@GretchenWhitmer never backs down from tackling the problems facing Michiganâ€™s working families, and she was a keyâ€¦ https://t.co/zbvJqrwQby
[205] .@JanetMillsforME is an experienced leader and an outstanding public servant running for governor of Maine to expanâ€¦ https://t.co/t6jbgbkxTQ
[206] .@NHMollyKelly is an experienced leader and tireless fighter running for New Hampshire governor. Sheâ€™ll fight to imâ€¦ https://t.co/jk4Z1CoGXY
[207] .@MarkBegich is a dedicated public servant, business owner, and former U.S. senator who has a record of cutting thrâ€¦ https://t.co/wZCorYpYgW
[208] .@dg4az is a husband, dad, veteran and teacher who will fight for education and root out corruption in Arizona. Davâ€¦ https://t.co/Rmvbs2ZC9M
[209] In one week, we have the chance to flip 17 governorships from red to blue. Here are four incredible candidates who deserve your support:
> #Text transformation
> toSpace <- content_transformer(function (x, pattern) gsub(pattern, "", x))
>
Environment History Connections
Import Dataset
Global Environment
tweetsc.df 211 obs. of 16 variables
Values
access_secr… "v5ZS9W4PKdRVD8yAZOhH9wC3zTWQNB…
access_token "2561943617-FRypUMwnXYs1ILF8A5Q…
consumer_key "D2SoRvRrWwWo87ysS8hq1zpPk"
consumer_se… "vVkxMgu4t3Hk0ZHnhl7ACJpRXMQyVo…
text chr [1:293] "\"text\",\"favorit…
toSpace function (x, ...)
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users waka-the-fisi Desktop twitter
Name Size Modified
.gitignore 13 B Apr 10, 20
.httr-oauth 0 B Apr 10, 20
.RData 819.2 KB Apr 12, 20
.Rhistory 11.1 KB Apr 12, 20
twitter.R 2.9 KB Apr 12, 20
tweets.csv 67.7 KB Apr 12, 20


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
C:/Users/waka-the-fisi/Desktop/twitter/
na. Davâ€¦ https://t.co/Rmvbs2ZC9M
[209] In one week, we have the chance to flip 17 governorships from red to blue. Here are four incredible candidat
es who deserve your support:
> #Text transformation
> toSpace <- content_transformer(function (x, pattern) gsub(pattern, "", x))
> tweets <- tm_map(tweets, toSpace, "/")
Warning message:
In tm_map.SimpleCorpus(tweets, toSpace, "/") :
transformation drops documents
> tweets <- tm_map(tweets, toSpace, "@")
Warning message:
In tm_map.SimpleCorpus(tweets, toSpace, "@") :
transformation drops documents
> tweets <- tm_map(tweets, toSpace,"\\|")
Warning message:
In tm_map.SimpleCorpus(tweets, toSpace, "\\|") :
transformation drops documents
> #CLeaning the text
> #Convert the text to lower case
> tweets <- tm_map(tweets, content_transformer(tolower))
Warning message:
In tm_map.SimpleCorpus(tweets, content_transformer(tolower)) :
transformation drops documents
> #Remove numbers
> tweets <- tm_map(tweets, removeNumbers)
Warning message:
In tm_map.SimpleCorpus(tweets, removeNumbers) :
transformation drops documents
> #Remove english common stopwords
> tweets <- tm_map(tweets, removeWords, stopwords("english"))
Warning message:
In tm_map.SimpleCorpus(tweets, removeWords, stopwords("english")) :
transformation drops documents
> #Remove punctuations
> tweets <- tm_map(tweets, removePunctuation)
Warning message:
In tm_map.SimpleCorpus(tweets, removePunctuation) :
transformation drops documents
Environment History Connections
Import Dataset
List
Global Environment
s 209 obs. of 10 variables
tweetsc.df 211 obs. of 16 variables
Values
access_secr… "v5ZS9w4PKdRVD8yAZOhH9wC3zTWQNB…
access_token "2561943617-FRypUMwnXYs1ILF8A5Q…
consumer_key "D2SoRvRrWwWo87ysS8hq1zpPk"
consumer_se… "vVkxMgu4t3Hk0ZHnhl7ACJpRXMQyVo…
text chr [1:293] "\"text\",\"favorit…
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users waka-the-fisi Desktop twitter
Name Size Modified
.gitignore 13 B Apr 10, 20
.httr-oauth 0 B Apr 10, 20
.RData 819.2 KB Apr 12, 20
.Rhistory 11.1 KB Apr 12, 20
twitter.R 2.9 KB Apr 12, 20
tweets.csv 67.7 KB Apr 12, 20


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
C:/Users/waka-the-fisi/Desktop/twitter/
transformation drops documents
> #Build a term-document matrix
> dtm <- TermDocumentMatrix(tweets)
> m <- as.matrix(dtm)
> v <- sort(rowSums(m), decreasing = TRUE)
> d <- data.frame(word = names(v), freq = v)
> head(d, 10)
word freq
right right 24
state state 22
vote vote 19
famili famili 16
week week 16
today today 16
peopl peopl 15
women women 14
make make 14
elect elect 13
> View(m)
> #Generate wordcloud
> set.seed(1234)
> wordcloud(words = d$word, freq = d$freq, min.freq = 1,
+ max.words = 200, random.order = FALSE, rot.per = 0.35,
+ colors = brewer.pal(8, "Dark2"))
There were 50 or more warnings (use warnings() to see the first 50)
> #Generate barplot
> enc2utf8("tweets")
[1] "tweets"
> library(syuzhet) #Derive Plot Arcs from Text
> library(dplyr) #Grammar of Data Manipulation
Attaching package: ‘dplyr’
The following objects are masked from ‘package:twitteR’:
id, location
The following objects are masked from ‘package:stats’:
Environment History Connections
Import Dataset
List
Global Environment
s 209 obs. of 10 variables
tweetsc.df 211 obs. of 16 variables
Values
access_secr… "v5ZS9w4PKdRVD8yAZOhH9wC3zTWQNB…
access_token "2561943617-FRypUMwnXYs1ILF8A5Q…
consumer_key "D2SoRvRrWwWo87ysS8hq1zpPk"
consumer_se… "vVkxMgu4t3Hk0ZHnhl7ACJpRXMQyVo…
text chr [1:293] "\"text\",\"favorit…
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users waka-the-fisi Desktop twitter
Name Size Modified
.gitignore 13 B Apr 10, 20
.httr-oauth 0 B Apr 10, 20
.RData 819.2 KB Apr 12, 20
.Rhistory 11.1 KB Apr 12, 20
twitter.R 2.9 KB Apr 12, 20
tweets.csv 67.7 KB Apr 12, 20


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal
The following objects are masked from 'package:twitteR':
id, location
The following objects are masked from 'package:stats':
filter, lag
The following objects are masked from 'package:base':
intersect, setdiff, setequal, union
> library(tm) #for text mining
> library(wordcloud) #word-cloud generartor
> library(RColorBrewer)
> #Read file
> tweets <- read.csv(file.choose(), header = T)
> tweets <- iconv(tweets$text, to = 'utf-8')
> s <- get_nrc_sentiment(tweets)
> head(s)
anger anticipation disgust fear joy sadness surprise trust negative positive
1 2 2 0 3 1 1 1 1 3 1
2 2 1 1 2 1 0 0 1 1 1
3 0 1 0 0 0 0 0 0 1 1
4 0 1 0 0 1 0 0 1 0 1
5 0 1 0 0 0 0 0 1 1 1
6 1 0 1 1 0 1 0 0 1 0
> tweets[4]
[1] "The white nationalists certainly think MAGA is a white nationalist slogan. https://t.co/Pp8Z7hBFRc"
> get_nrc_sentiment('delay')
anger anticipation disgust fear joy sadness surprise trust negative positive
1 1 0 1 1 0 1 0 0 1 0
> # Bar plot
> barplot(colSums(s),
+ las = 2,
+ col = rainbow(10),
+ ylab = 'Count',
+ main = 'Sentiment Scores for Tweets')
Environment History Connections
Global Environment
s 209 obs. of 10 variables
tweetsc.df 211 obs. of 16 variables
Values
text chr [1:293] "\"text\",\"favorit…
Files Plots Packages Help Viewer
Sentiment Scores for Tweets
Count
0 50 100 150 200
anger anticipation disgust fear joy sadness surprise trust negative positive


~/R/R project/Twitteranalysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
test1.R twitter.R* tweets
Source on Save Run Source
67
68 #Build a term-document matrix
69 dtm <- TermDocumentMatrix(tweets)
70 m <- as.matrix(dtm)
71 v <- sort(rowSums(m), decreasing = TRUE)
72 d <- data.frame(word = names(v), freq = v)
73 head(d, 10)
74
75 #Generate wordcloud
76 set.seed(1234)
77 wordcloud(words = d$word, freq = d$freq, min.freq = 1,
78 max.words = 200, random.order = FALSE, rot.per = 0.35,
79 colors = brewer.pal(8, "Dark2"))
80
81 #Generate barplot
82 enc2utf8("tweets")
83 library(syuzhet) #Derive Plot Arcs from Text
84 library(dplyr) #Grammar of Data Manipulation
85
Console Terminal
~/R/R project/Twitteranalysis/
vote vote 19
famili famili 16
week week 16
today today 16
peopl peopl 15
women women 14
make make 14
elect elect 13
> #Generate wordcloud
> set.seed(1234)
> wordcloud(words = d$word, freq = d$freq, min.freq = 1,
+ max.words = 200, random.order = FALSE, rot.per = 0.35,
+ colors = brewer.pal(8, "Dark2"))
There were 50 or more warnings (use warnings() to see the first 50)
Environment History Connections Git
Global Environment
d 1368 obs. of 2 variables
dtm List of 6
m Large matrix (285912 elem…
tweets List of 209
tweetsc.df 209 obs. of 16 variables
Values
Files Plots Packages Help Viewer
puerto children health women work state right today vote peopl week make last trump fight offic new elect run will famili human thank day help bill know one gun


~/R/R project/Twitteranalysis - RStudio

	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource
1	When a gun is present in a domestic violence situation, ...	FALSE	9234	NA	2019-04-10 14:04:52	TRUE	NA	1.115979e+18	NA	<a href="http://twitter.com" rel="nofollow">
2	Thank you, @DewSteele, for turning @EmergeAmerica i...	FALSE	6847	NA	2019-04-09 20:20:03	TRUE	NA	1.115711e+18	NA	<a href="http://twitter.com" rel="nofollow">
3	The astronomical cost of insulin is a public health crisis. ...	FALSE	31869	NA	2019-04-09 15:15:02	FALSE	NA	1.115634e+18	NA	<a href="http://twitter.com" rel="nofollow">
4	The white nationalists certainly think MAGA is a white n...	FALSE	36180	NA	2019-04-09 13:36:14	FALSE	NA	1.115609e+18	NA	<a href="http://twitter.com" rel="nofollow">
5	Family separation profoundly harms children and their p...	FALSE	26301	NA	2019-04-08 21:17:55	TRUE	NA	1.115363e+18	NA	<a href="http://twitter.com" rel="nofollow">
6	Let's be clear: This administration's dehumanization and...	FALSE	102914	NA	2019-04-08 13:33:46	TRUE	NA	1.115246e+18	NA	<a href="http://twitter.com" rel="nofollow">
7	Got to practice my snap last night with the best in the b...	FALSE	34410	NA	2019-04-07 01:43:11	TRUE	NA	1.114705e+18	NA	<a href="http://twitter.com" rel="nofollow">
8	If you have any junior-year college students in your life, ...	FALSE	20293	NA	2019-04-05 13:31:02	TRUE	NA	1.114159e+18	NA	<a href="http://twitter.com" rel="nofollow">
9	A reminder of why the 2018 midterms were so important...	FALSE	17313	NA	2019-04-04 18:33:35	TRUE	NA	1.113872e+18	NA	<a href="http://twitter.com" rel="nofollow">
10	When at the @Cher Show... Thanks for the company, Bo...	FALSE	5787	NA	2019-04-04 17:38:22	TRUE	NA	1.113858e+18	NA	<a href="http://twitter.com" rel="nofollow">
11	Mitch McConnell reportedly has no interest in putting t...	FALSE	20095	NA	2019-04-04 13:13:33	TRUE	NA	1.113792e+18	NA	<a href="http://twitter.com" rel="nofollow">
12	Next week, Bill and I will be in New York on the 11th wit...	FALSE	5017	NA	2019-04-03 14:51:48	TRUE	NA	1.113454e+18	NA	<a href="http://twitter.com" rel="nofollow">
13	I wish @TyroneGayle could see this: his family and friend...	FALSE	9345	NA	2019-04-03 14:15:39	TRUE	NA	1.113445e+18	NA	<a href="http://twitter.com" rel="nofollow">
14	When anti-choice politicians limit access to reproductive...	FALSE	24123	NA	2019-04-03 13:12:18	TRUE	NA	1.113429e+18	NA	<a href="http://twitter.com" rel="nofollow">
15	Only 0.1% of all elected officials in the U.S. identify as L...	FALSE	8412	NA	2019-04-02 16:54:42	TRUE	NA	1.113123e+18	NA	<a href="http://twitter.com" rel="nofollow">
16	Today is #EqualPayDay, marking how far into 2019 wome...	FALSE	13022	NA	2019-04-02 15:59:07	TRUE	NA	1.113109e+18	NA	<a href="http://twitter.com" rel="nofollow">
17	Facts: - Puerto Ricans are Americans. - Puerto Ricans hav...	FALSE	153000	NA	2019-04-02 14:54:07	TRUE	NA	1.113092e+18	NA	<a href="http://twitter.com" rel="nofollow">
18	On this #TransDayOfVisibility, let's affirm that we see, st...	FALSE	38818	NA	2019-03-31 17:54:14	TRUE	NA	1.112413e+18	NA	<a href="http://twitter.com" rel="nofollow">
19	It's pretty simple: Men and women in the same job shou...	FALSE	31942	NA	2019-03-28 13:03:16	TRUE	NA	1.111252e+18	NA	<a href="http://twitter.com" rel="nofollow">
20	Every single Senate Democrat has signed on as a co-spo...	FALSE	18284	NA	2019-03-27 18:50:30	TRUE	NA	1.110977e+18	NA	<a href="http://twitter.com" rel="nofollow">
21	I'm so happy to see candidates winning the right to use ...	FALSE	26465	NA	2019-03-27 13:13:13	FALSE	NA	1.110893e+18	NA	<a href="http://twitter.com" rel="nofollow">
22	A great opportunity for young people in Arizona, Florid...	FALSE	3773	NA	2019-03-26 20:34:19	TRUE	NA	1.110641e+18	NA	<a href="http://twitter.com" rel="nofollow">

Showing 1 to 23 of 209 entries

CHAPER 5: CONCLUSION

5.1 Introduction

This chapter is supposed to show the conclusion of the entire system as a whole showing the how the process of creating the system has been as well as making clear some of the challenge the that I faced whilst striving to create the system.It will also show the milestones that were covered and the new experience acquired when dealing with it.

5.2 Results

The end result of the software is pretty good given that R Studio comes with an already fully functional interface it was very helpful to see that the studio came with easy to use pre - installed packages that on simply needs to call from library they are stored in. After this inserting the code to perform the sentiment analysis is pretty straight forward which helped the entire system even more by creating a CSV file of the data one wants to analyse the rest is up to R studio which runs the code and gives an output of the results.

5.3 Problem Faced

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.

5.4 Database Creation

I had an easy time creating the database as I was deriving direct live tweets from twitter and then saving the tweets in a.CSV using excel. This made it easier in creating, organizing and retrieving my database.

5.5 Experience

Through this process of creating a personal system I learnt the hardship of creating a system all on your own though it is very possible and doable I began to value the ease that comes with teamwork which enables one to distribute roles and tasks evenly across the group making the work easier and more efficient, I learnt to code using R studio which was totally new to me.

5.6 Source code

Author: GathingirahWilfredWakanya

```
library(twitteR)
library(httr)
library(streamR)

#Set directory
setwd("C:/Users/waka-the-fisi/Desktop/twitter")

#Accessing Twitter API
consumer_key <- 'D2SoRvRrWwWo87ysS8hq1zpPk'
consumer_secret <- 'vVkxMgu4t3Hk0ZHnhl7ACJpRXMQyVogBP8vPaGgQReWUG3yfsp'
access_token <- '2561943617-FRypUMwnXYs1ILF8A5QZDVcMyB4Xnf4npaKrmKV'
access_secret <- 'v5ZS9W4PKdRVD8yAZOhH9wC3zTwQNB6iYpm8VOgfTdTiI'
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

#Extract tweets
```

```
clinton_tweets <- userTimeline("HillaryClinton", n = 2000, since = "2016-01-09",
languageEl("english", which = "en"))
tweetsc.df <- twListToDF(clinton_tweets)
dim(tweets)

#Save Rdata
write.csv(tweets, file = 'C:/Users/waka-the-fisi/Desktop/twitter/tweets.csv', row.names = F)
head(tweets)

#Get the text
text <- readLines(file.choose())
library(slam) #for sparse arrays and matrices
library(NLP) #to understand human language as it is spoken
library(tm) #for text mining
library(SnowballC) #for text stemming
library(wordcloud) #word-cloud generartor
library(RColorBrewer) #for color palettes

#Load data as corpus
tweets <- iconv(tweets$text, to = "utf-8")
tweets <- Corpus(VectorSource(tweets))
inspect(tweets)

#Text transformation
toSpace <- content_transformer(function (x, pattern) gsub(pattern, "", x))
tweets <- tm_map(tweets, toSpace, "/")
tweets <- tm_map(tweets, toSpace, "@")
tweets <- tm_map(tweets, toSpace,"\\|")

#CLeaning the text
#Convert the text to lower case
```

```
tweets <- tm_map(tweets, content_transformer(tolower))

#Remove numbers
tweets <- tm_map(tweets, removeNumbers)

#Remove english common stopwords
tweets <- tm_map(tweets, removeWords, stopwords("english"))

#Remove punctuations
tweets <- tm_map(tweets, removePunctuation)

#Eliminate white spaces
tweets <- tm_map(tweets, stripWhitespace)

#Text stemming (reduce words to unify across documents)
tweets <- tm_map(tweets, stemDocument)



#Build a term-document matrix
dtm <- TermDocumentMatrix(tweets)
m <- as.matrix(dtm)
v <- sort(rowSums(m), decreasing = TRUE)
d <- data.frame(word = names(v), freq = v)
head(d, 10)

#Generate wordcloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
        max.words = 200, random.order = FALSE, rot.per = 0.35,
        colors = brewer.pal(8, "Dark2"))
```

```
#Generate barplot
enc2utf8("tweets")
library(syuzhet) #Derive Plot Arcs from Text
library(dplyr) #Grammar of Data Manipulation
library(tm) #for text mining
library(wordcloud) #word-cloud generartor
library(RColorBrewer) #color palettes

#Read file
tweets <- read.csv(file.choose(), header = T)
tweets <- iconv(tweets$text, to = 'utf-8')

s <- get_nrc_sentiment(tweets)
head(s)
tweets[4]
get_nrc_sentiment('delay')

# Bar plot
barplot(colSums(s),
        las = 2,
        col = rainbow(10),
        ylab = 'Count',
        main = 'Sentiment Scores for Tweets')
```

REFERENCES

Bhatnagar, M., & Singh, K. (2013). Research Methodology as SDLC Process in Image Processing. *International Journal of Computer Applications*, *Vol 77 No 2*.

Cohen-Almagor, R. (2011). Fighting Hate and Bigotry on the Internet. *Policy & Internet*, *Article 6*.

Commission, N. C. and I. (2011). National Cohesion and Integration Commission. *Police Training Manual- On the Enforcement of the Law on Hate Speech*, (Nairobi: National Cohesion and Integration Commission).

Commission, N. C. and I. (2013). National Cohesion and Integration Commission. *The Use of Coded Language and Stereotypes among Kenyan Ethnic Communities*, (Nairobi: NCIC.).

Hirsch, S. (2009). Putting Hate Speech in Context: Observations on Speech, Power, and Violence in Kenya. George Mason University. (n.d.).

Makinen, M., & Kuira, M. (2008). S. M. and P.-E. C. in K. S., & Commons. (n.d.). Makinen, M., & Kuira, M. (2008). Social Media and Post-Election Crisis in Kenya. Scholarly Commons.

Mejova, Y. (2009). Sentiment Analysis: An Overview. Comprehensive Exam Paper. *Computer Science Department*, (May), 1–34.

Mugambi, S. K. (2017). Sentiment analysis for hate speech detection on social media : TF-IDF weighted N-Grams based approach Sentiment analysis for hate speech detection on social media : TF-IDF weighted N-Grams based approach.

Mukherjee, S., & Bhattacharyya, P. (2012). Sentiment Analysis in Twitter with Lightweight Discourse Analysis. *Proceedings of COLING 2012*, (December 2012), 1847–1864.

National Council for Law Reporting. (2008). *National Cohesion and Integration Act*, (Nairobi).

Sambuli, N., Morara, F., & Mahihu, C. (2013). M. O. D. S. in K., & Umati., N. (n.d.).

UNESCO. (2015). United Nations Educational, Scientific and Cultural Organization. *Countering Online Hate Speech*, (Paris: UNESCO).

Waseem, Z., & Hovy, D. (2016). H. S. or H. P. P. F. for H., For, S. D. on T. N.-H. (pp. 88-93). S. D. A., & Linguistics., C. (n.d.). No Title.