THE TWIT TWEET MINER

Report on implementation, design and conclusions of the project

Hari Priya N -
II BSc Computer Science
Meenakshi College for women

Introduction

Context:

"The Twit Tweet Miner" is a project done as a part of my internship for the course of Bachelor of Science in Computer Science at Meenakshi College for Women.

Motivation:

I have always been interested and fascinated by Machine Learning and Data Science. This independent project became a blessing for me to explore and learn more about these fields. The self-paced project on Xcelerator provided me a great platform to experiment, play and understand the underlying concepts of Machine Learning and Data Science. Thanks to my teachers for suggesting me this project which has been a wonderful learning experience

The Project

In Brief:

Data Science and Machine Learning are two words that fascinate and take up utmost importance for computer graduates and scientists in today's era. I chose to work with these fields as I always felt that proper retrieval and analysis of data can do wonders and help us in many ways.

The "Twit Tweet Miner" is a project that analyses data retrieved from Twitter the popular social networking service. Data is examined in detail by finding frequently used words from tweets using R and machine learning techniques like K-means clustering and the results are graphically portrayed in the form of plots and word clouds.

Requirement:

The project involves the need to perform Text Mining on a trending hashtag (I have worked with #covid). Tweets mined have to be transformed to text by changing letters to lowercase, removing punctuation/numbers and removing stop words. The frequently used words in the tweets are visualised by using a word cloud.

The tweets are clustered using k-means algorithm into 5 main clusters. This technique is useful to further analyse the data collected for various other purposes like finding potential classification, sentiment analysis, etc.

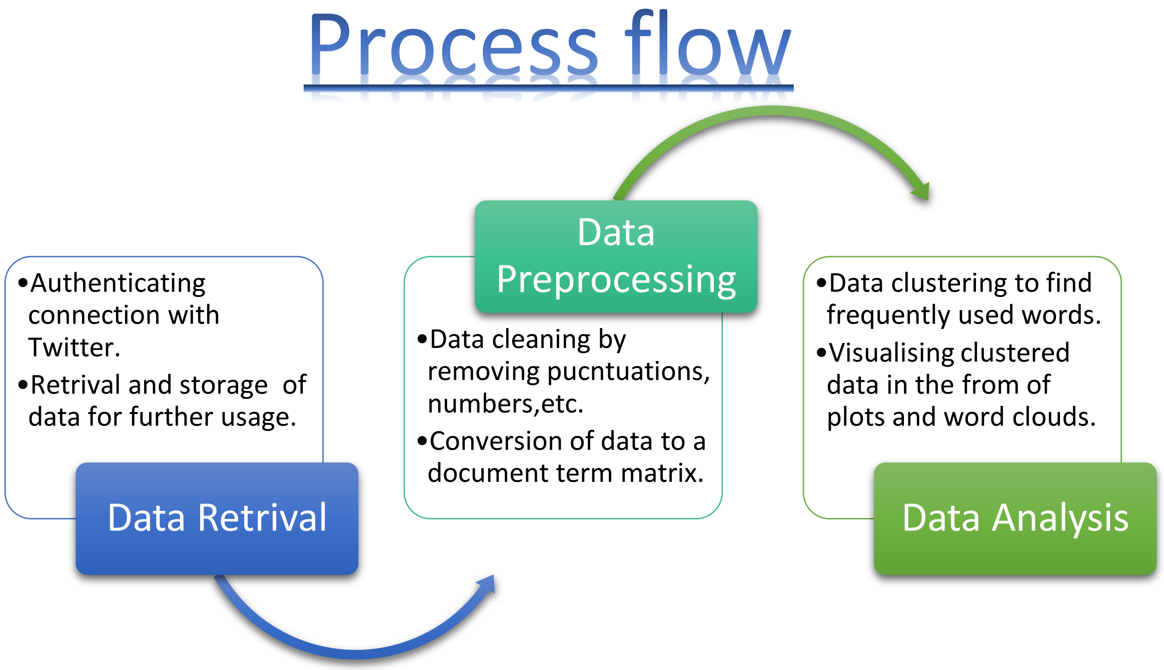
Design:

The Twit Tweet Miner is designed to retrieve tweets from Twitter based upon a common hashtag using language R. In this case I chose #covid as I felt this would be highly trending these days. Library "twittR" is used to create and authenticate connection to twitter. The functions in the package are also used to retrieve and store the tweets for further usage.

The next crucial step is to clean, pre-process and make required conclusions on the data that was retrieved. Packages "tm" and "SnowballC" are used to clean the data by removing punctuations, numbers, special characters and stop-words which otherwise could hinder in our analysis. The data obtained as a result of these steps is pre-processed.

The data is later converted into a document term matrix that enables mathematical calculations on the data for capturing interesting conclusions.

Now that basic things are set, data can be clustered and frequently used words can be found. Packages “factoextra”, “ggplot2”,”wordcloud” and “cluster” are loaded and functions like fviz_cluster(), clusplot(), wordcloud() from these libraries can used to visualise the data as clusters, plots and word clouds.


Process flow

- Authenticating connection with Twitter.
- Retrival and storage of data for further usage.

Data Retrival

Data Preprocessing

- Data cleaning by removing pucntuations, numbers,etc.
- Conversion of data to a document term matrix.

- Data clustering to find frequently used words.
- Visualising clustered data in the from of plots and word clouds.

Data Analysis

Implementation:

Initial setup:

Installation of R language and R Studio

R language and R studio have to be downloaded as the project is entirely done with language R. After the download, the “devtools” package which consists of the basic functions is to be installed

Creating a twitter developer account

The next step is to create a Twitter developer account. This would enable the user to use the Twitter API in order to retrieve data. This requires the user to sign up to Twitter. Later a form is to be submitted to Twitter requesting developer access in order to create an application. After the developer access is granted and an app is created, a set of API tokens are generated which is to be noted safely for further use.

R Studio set up to enable connection to Twitter:

Downloading “twitteR”

R uses the “twitteR” library to communicate with the Twitter API. Thus “twitteR” has to be downloaded.

Authentication

Twitter uses Open Authentication (OAuth) to grant access to the information. To setup the authentication, setup_twitter_oauth() function is to be used which takes consumer_key, consumer_secret, access_token and access_secret as the parameters. These parameters are generally provided by Twitter after we gain developer access.

```
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

Data Retrieval:

Querying Twitter

To access the Twitter data, we need to query our requirement to Twitter. In this project I have retrieved data based on #covid. In this project, function searchTwitter() retrieves the tweets based on #covid.

```
tweetCovid = searchTwitter(“#covid”, n=5000,  lang= “en”)
```

For further analysis these tweets are converted to a data frame and are stored locally.

Processing and Analysing the Data:

Cleaning the Data

Tweets retrieved have a lot of unwanted text and characters which would have to be removed in order to obtain meaningful results. This is done by removing punctuations, numbers, special, characters, extra whitespaces and non-ASCII characters. The “tm” package provides us with function tm_map() which helps in removing unwanted characters. But to use tm_map() we need to convert our data object into a Corpus. A Corpus is a collection of text document on which text mining and natural language processing routines can be applied.

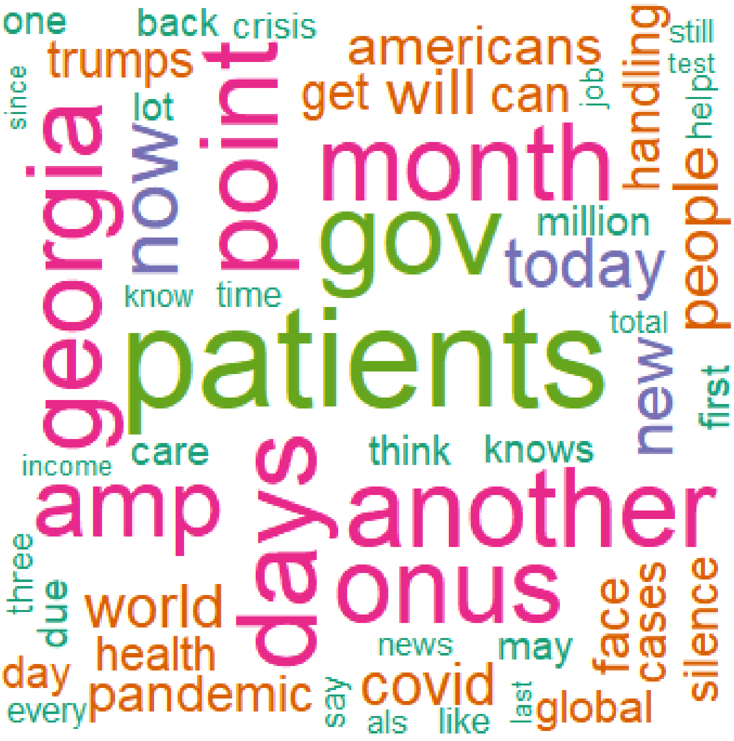
The Document Term Matrix

After cleaning the data, the data is converted to a Document Term Matrix. A Document Term Matrix consists of rows and columns in which rows correspond to the documents and columns corresponds to the terms. The Document Term Matrix enables us to perform numerical calculations on the data.

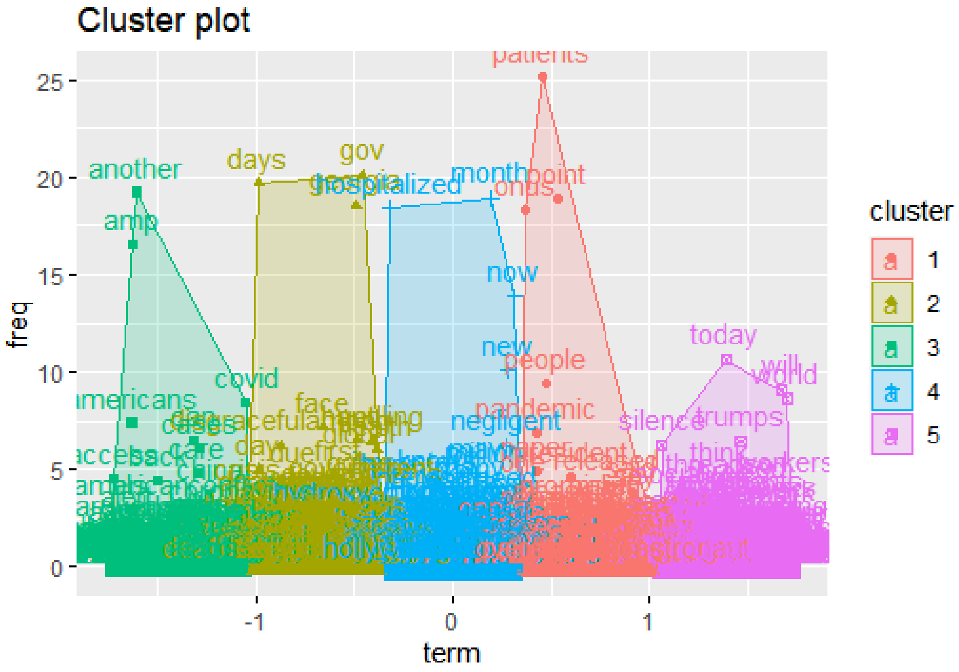
Visualising data using plots

The frequent terms with frequency of at least 50 are found and plotted as a word cloud. K-means clustering is performed on the data and is clustered into 5 groups. These clusters are visualised using function clusplot().

Conclusion



Word cloud of frequent words used in tweets with #covid.



Cluster plot using fviz_cluster() of frequent words used in tweets with #covid.

Thus the data was explored, cleaned and pre-processed. Cluster analysis was also successfully performed on the data. The plots show us the results obtained.

Challenges and scope for improvement

Challenges

The project takes in large amounts of data for frequent term analysis, results might not always yield distinct results. This is because tweets retrieved might not actually have key words that would contribute for the frequent term analysis, and hence is a challenge.

The other major constraint is compute and memory resources on the desktop. By retrieving 5000 tweets and converting them into a document term matrix (after cleaning) we end up at a humongous matrix with 34 million terms. This slows down the processor and occupies quite a large chunk of memory.

Scope for Improvement

This project scope is confined to analysing frequent terms obtained by a single trending hashtag. Different related hashtags (Eg: #covid, #corona, #covid-19) can be used and results of word clouds can be compared.

References

R (https://cran.r-project.org/)

R Studio (https://www.rstudio.com/products/rstudio/download/)

Twitter Developer (https://apps.twitter.com/)

TwitteR (http://geoffjentry.hexdump.org/twitteR.pdf)

R manual (https://cran.r-project.org/doc/manuals/r-devel/R-lang.pdf)

Tm (https://cran.r-project.org/web/packages/tm/tm.pdf)

Wordcloud2 (https://cran.r-project.org/web/packages/wordcloud2/vignettes/wordcloud.html)