Exploring the R Envinronment for Text Mining of Twitter Data

Mr. A. R. Ukalkar
Member CSI, Assistant Professor Department of Comp. Tech.
KITS, Ramtek Nagpur Maharashtra India
Email: arualkar@gmail.com

Abstract—The objective of the paper is to explore the R environment for text mining of twitter data in order to understand the underlying pattern in the social interaction of the user. We are going to explore the twitter Package provided by R data mining environment. We will try to download the data from twitter. The term document matrix, frequency of terms graph, and word cloud will be used to perform analysis of the data. The twitter provides Twitter API which can be used by the application to get access to twitter data. In our paper we are briefly exploring how to use twitter API to download data from the twitter account. The download of the data from twitter account requires the user to authenticate using OAuth object. This requirement is added since early 2013. We will also explore the tm package which provides a range of services to perform text mining of the data. The tm package accepts the data in various formats such as pdf, plain text, CSV format etc. It provides functions for the pre-processing of text data such as stemming, removing stop words, allowing filters etc. Using tm package we can perform classification, clustering and outlier analysis of the data. We can indentify associations among the terms using the term document matrix. The R provides efficient environment to perform data intensive computations. There are thousands of the packages provides in R environment which can be used by researchers for doing analysis and identifying the underlying patterns in the data. Using twitteR package we can download the tweets of the public timeline of users which can be analysed to understand the social behaviour of the user. The analysis of the twitter data will show the association of the user with the subject he/she is interested in.

Keywords*-Social network data mining, twitter data mining, text mining*

*****

I. What is Twitter?

On twitter website the user can post small messages up to 140 characters each. These messages can be read publicly. Twitter also allows its users to connect with other users. Once the user is added as friend to the account, tweets posted by the user are displayed in the friends twitter account. The twitter is very effective tool for communication among the team members working in small team which is geographically separated. Once the user registers on the twitter, he/she can personalize the user profile. The tweets of the user are displayed in reverse chronological order on public timeline page.

The twitter also provides the API using which other application can add the twitter functionality to their web site. [1]

II. Introduction to tm package Text Mining in R

The collection of documents in called Corpus. There is volatile and permanent corpus. The volatile corpus is lost when the R object is destroyed. The permanent corpus is stored externally. The R object is merely the pointer to the corpus stored externally. The changes to the corpus are reflected in all the R objects pointing to it.

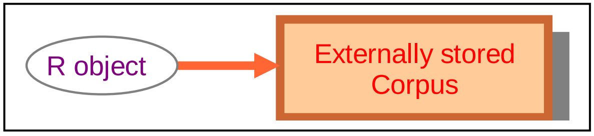

Fig 1. The R object is a pointer to corpus which stored separately outside R environment.

The volatile corpus can be created using constructor Corpus(x, readerControl) where parameter x represent input location such as directory, vector representing document, data frame like structure.

The second parameter is a list with two components reader and language. The reader creates text document specified by the input location or source x. Tm package come with different readers for example readPlain(), readGmane(), readRCV1(), readPDF(), readDoc() to name few. Second component language selected the text's language. We can save the corpus using writeCorpus() function. The writeCorpus() function saves corpus in multiple files. Each file corresponds to a document in the corpus. The full content of text document is displayed using inspect() function. tm_map function applies the transformation to all the documents present in corpus. The transformation may be converting one type of document to other type of document, removing the white spaces from the document, converting the document to lower case, removing stop words, stemming, applying filters et cetera.

We can create the term document matrix using tm package. Various operations can be performed on term document matrix such as finding frequent terms, finding associations among the terms, removing sparse terms. [2]

III. Introduction to Twitter Client for R

People are using twitter to connect with each other. Twitter has extended itself to the research groups and small teams that are separated from one another geographically. There has been an effort to tap the vast amount of data available on twitter for doing variety of research related to data mining and sentiments mining. In recent times people have done research on the data obtained from various twitter accounts to predict the trends and to identify individuals who contribute to setting trends etc. The companies who want to use online marketing strategies are interested in such research activities. Owing to greater interests of the researchers and users who want to access twitter data, the twitter transaction are needed to be authenticated using OAuth. To use the twitter data one has to first register on the twitter application website. After registering the application on twitter the user will be given *API key, API secret, Access Token, Access Secret.* Using these credentials the user can authenticate and get the access to the twitter data.

setup_twitter_oauth(“API_KEY”,”API_SECRET”). There is *httr* package that provides the functionality to the twitteR package regarding authentication.

The twitteR package provides various functions to access data on twitteR. The searchTwitter function provides the searching facility. It takes two parameters the first parameter is the keyword to be searched in the tweets and other is the number of tweets to be downloaded. There is a limitation for accessing the number of the tweets from the public timeline. We can collect the data related to the user. The getUser() function gives various details related to the twitter user such as their followers , who the users follow, their retweets etc. To convert the tweets into data frames there is a function called toDataFrame(). The stream of tweets is called the timeline. There are two timelines supported by twitteR, the user and home timeline. The user timeline displays the recently tweeted messages from the different users and home timeline displays the users own recent tweets. To identify the recent trends on the twitter the twitteR package has getTrends() function. It returns the recent trend information related to the location. The location is identified using WOEID. [3]

IV. RESULT AND ANALYSIS

A. *Term Document Matrices*

The most common way of organizing the text documents is to represent them using Term Document Matrix. The Term Document Matrix prints the frequency of terms along with the document in which these terms appear.

We have downloaded the 300 tweets containing word “big data” and its variations from the tweeter. After the tweets are downloaded, we have created the corpus of the tweets. The tweet is treated as single document. Thus we have created the corpus of 300 documents. In the subsequent steps we have removed the punctuation marks, numbers and URLs from the corpus. We have removed the noise words from the corpus. The corpus is used to build the term document matrix.

Then the resulting term document matrix is:

```
A term-document matrix (756 terms, 300 documents)
Non-/sparse entries: 2618/224182
Sparsity           : 99%
Maximal term length: 25
Weighting          : term frequency (tf)
```

B. *Finding Frequent Terms*

We can find the frequent occurring terms whose frequency is greater than 10.

```
[1] "almost"  "amp"   "analytics"   "big"  "bigdata"
[6] "chief"   "cloud"   "companies"   "data"  "double"
[11]"enterprises" "existed""future" "going" "hadoop"
[16] "ibmbigdata"  "internet"  "iot"   "janinebucks"  "job"
[21] "just" "kpipartners" "like" "make" "marketing"
[26]"mckinsey""midmarketibm""new""number" "power"
[31] "requires"  "sales" "storm"  "techinasia"   "things"
[36] "use" "will" "world"  "year"  "zettabytes"
```

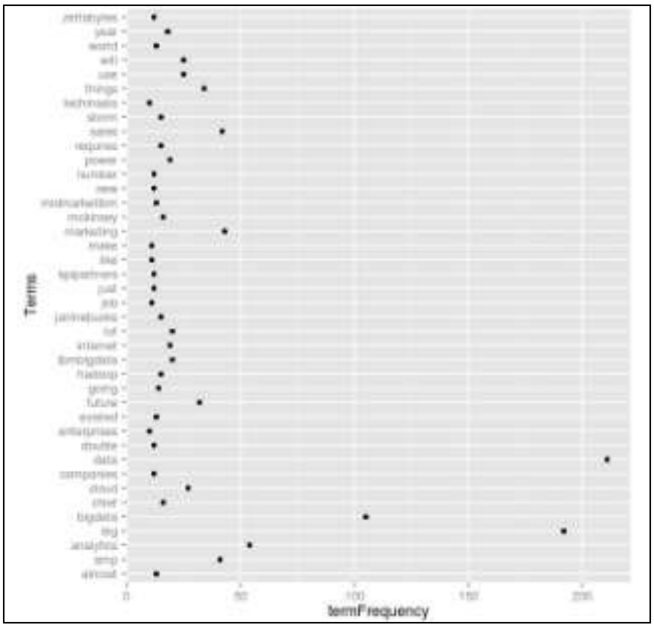

Fig 2.The above grapsh shows the terms and their frequency.

The term documents relates the terms with the documents. The row represents the terms and column represents the document. The column has entry which indicates how many times the particular term has occurred in the document. The corpus consists of large number of documents and normally the term document matrix maps many terms with the documents. Hence, the term document matrix is normally sparse in nature. The term document matrix can be compressed as most the entries in it are zeros. The *findfrequentterms()* finds the frequently occurring terms and their associations with each other.

C. *WordCloud*

The word cloud can be constructed to show the importance of words. The word cloud function takes two parameters, first parameter is the list of words and second parameter is frequency.

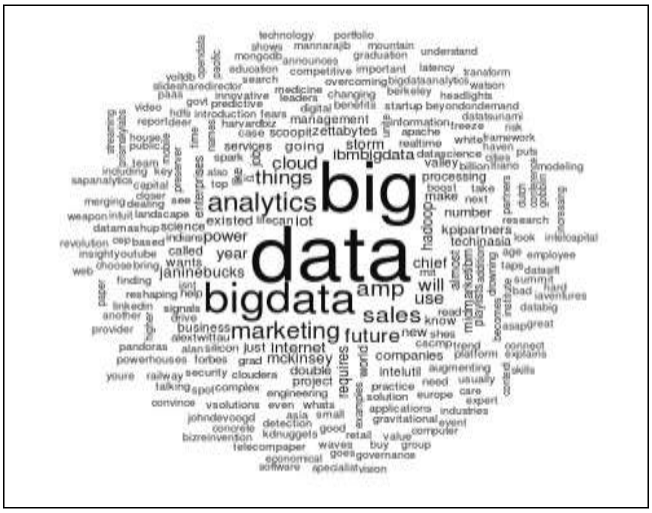

Fig 3. The above figure shows the big data word cloud

The word cloud can be build using term document matrix. The word cloud gives the pictorial view of the frequency of the words in the term document matrix. The most frequent terms are shown in big font and in bold.

D. Mining Twitter Data:

The twitter data can be useful to understand the interest of the user. The interest graph for the user depicts the correlation of the user with the things in which user is interested. There is huge amount of data that is uploaded live on the twitter by millions of users. This data can be mined for knowing the happening trends on the twitter. The Twitter API allows us to specify locations in search for the happening trends. Presently, only Twitter data is used for analysis but the Facebook and LinkedIn data can also be downloaded using respective API. The Facebook and LinkedIn data can provide the insight into user social and professional interests.

References

[1] Ingo Feinerer,"7 things you should know about twitter" Educase Learning Initiative. www.educase.edu/eli

[2] Ingo Feinerer,"Introduction to tm package text mining in R" January 13,2014

[3] Jeff Gentry, "Twitter Client for R" March 18, 2014

[4] Ingo Feinerer Wien,"A text mining framework in R and its applications", Agust 2014