Monitoring the Keyphrase of tweets using R

Reema Gupta
Assistant Professor
Department of Computer Science
Govt PG College for Women,Sector-14,Panchkula

Abstract—There is a great challenge how to manage big Data. As data storage and processing in a way that it becomes available to be consumed is really a tough task. Because data cannot be load into our traditional database system for analysis and storage, it leads to high cost and much time. Big data is described by 5 Vs (Volume, Variety, Velocity, Variability, Veracity). There are variety of tools available for big data analysis. R is one of the most popular statistical analysis tool. It works on Windows, Linux. R consists of numerous ready to use algorithms, formulae for analysis purpose.

Keywords—Analysis, Veracity, Package, Tweets, R

INTRODUCTION

Big data analytics include collection of data from various sources, mainly from social media, IOT, cloud computing, internet, databases and to process it in a way that it can be consumed by analysts [1]. Data analysis is important for a better decision making and also a challenging task as it includes data in unstructured, semi structured and structured form. Big data analysis faces lot of problems as data contain more ambiguity, errors, incomplete due to which storage, fault tolerance, and quality matters [2].

Big data analytics and statistics help in long term decisions. Big data is defined from 5 perspective also known as 5V’s.

1. Velocity
2. Volume
3. Variety
4. Veracity
5. Value

Velocity deals with the speed of accumulation of data generation. Volume is related to size. A variety deal with the nature of data i.e. Unstructured, Semi Structured and Structured. Veracity refers to the uncertainty in data and quality of data. Value is useful data as data is very big in quantity and useless until turn into a useful form [3]. In traditional analysis, statistical calculations consume more time. So, we need to use tools and techniques. There are many tools and techniques used for big data analysis. One of the General purpose programming language used for statistical analysis is R. R is a open source software which also provides the visual analytics and having strong graphical and statistical tool embed. R facilitates data management processes such as Transformations, subsetting and cleaning[4]. R consists of numerous ready to use algorithms, formulae for analysis purpose. It allows user to create data Output in the form of plots, graphs, and diagrams. R is also used by Google, facebook, Microsoft etc. for analysis.

This paper presents analysis conducted using R tool on Twitter-tweets.
R AND RSTUDIO

R is one of the most popular statistical analysis tools. It works on Windows, Linux. R consists of numerous ready to use algorithms, formulae for analysis purpose. It’s a programming language used for data manipulations, statistical analysis, and data visualization. R is an open source and free. R includes Conditional statements, recursive functions, and input/output commands, built in graphical tools for visualization[5][7]. Thousands of packages are available. To use a package, you must first install it and then load it.

Packages are installed using
install. Packages (name of package)
Loading of package is done using
library (name of package)

The R workspace consists of all the data objects you’ve created or loaded during your R session. When you quit R by either typing q() or exiting out of the application window, R will prompt you to save your workspace.

PROPOSED WORK

The methodology used to extract the tweets and further analyzed using R is described step by step as shown in figure 1. Pre-processing is a first step plays a very important role in text mining techniques and applications.

Figure 1. Process to analyze the tweets
RESULTS

In this section the words appeared in tweets, are extracted for further analysis[6]. The word counts based on their occurrences have been shown in figures.

Following packages are used to extract the results.

```r
library(twitteR)
library(rtweet)
library(syuzhet)
library(ggplot2)
library(dplyr)
library(tidyr)
library(tidytext)
library(widyr)
library(devtools)
```

1. Head(word_counts)

![Image](image1.png)

Figure 2. Word counts appeared single time

Head displays top 6 records with number of count of appeared words.

2. When a pair of words come together most of the time

![Image](image2.png)

Figure 3. Word counts appeared in pairs

When a pair of words come together most of the time in the tweets (data set taken)
3. When three words repeat maximum number of times

<table>
<thead>
<tr>
<th>paired_words</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>å å å å</td>
<td>3165</td>
</tr>
<tr>
<td>å å å å</td>
<td>2600</td>
</tr>
<tr>
<td>å å å å</td>
<td>2323</td>
</tr>
<tr>
<td>å å å å</td>
<td>1889</td>
</tr>
<tr>
<td>å å å å</td>
<td>1436</td>
</tr>
<tr>
<td>å å å å</td>
<td>1274</td>
</tr>
<tr>
<td>user user user</td>
<td>989</td>
</tr>
<tr>
<td>å å user</td>
<td>716</td>
</tr>
<tr>
<td>å å å å</td>
<td>659</td>
</tr>
<tr>
<td>å å å å å å</td>
<td>477</td>
</tr>
</tbody>
</table>

# ... with 190,064 more rows

Figure 4. Word counts appeared three times

When three words repeat more times it is counted with count displayed.

4. When two pairs(4 words) occur together

<table>
<thead>
<tr>
<th>paired_words</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>å å å å</td>
<td>1813</td>
</tr>
<tr>
<td>å å å å å å</td>
<td>1739</td>
</tr>
<tr>
<td>å å å å å å</td>
<td>1473</td>
</tr>
<tr>
<td>å å å å å å</td>
<td>1129</td>
</tr>
<tr>
<td>å å å å å å</td>
<td>1093</td>
</tr>
<tr>
<td>å å å å å å</td>
<td>1063</td>
</tr>
<tr>
<td>user user user</td>
<td>464</td>
</tr>
<tr>
<td>å å å å å å å å</td>
<td>461</td>
</tr>
<tr>
<td>å å å å å å å å</td>
<td>241</td>
</tr>
<tr>
<td>I am thankful for</td>
<td>221</td>
</tr>
</tbody>
</table>

# ... with 212,148 more rows

Figure 5. Word pairs occurred together

When 4 words appeared maximum number of times displayed with number of times of their repetition.

5. Count of unique words appeared in tweets

This graph indicates the unique words on Y-axis appeared in tweets and their count on X-axis.

Figure 6. Unique words appeared in tweets
TWEETS-WORD NETWORK

Word network is analyzed from the data set (tweets) taken for analysis using R. This analysis is done on the factor of their repetition in tweets.

- **N > 24**

This word network shows the number of words in the network of tweets having count more than 24.

- **When N > 50**

This (Word network) shows the no of words in the network of tweets having count more than 50.

- **When N > 250**

This (Word network) shows the no of words in the network of tweets having count more than 250.
When N>500
This (Word network) shows the no of words in the network of tweets having count more than 500.

When N>1024
This (Word network) shows the no of words in the network of tweets having count more than 1024.
Figure 10. Words count > 1024

Figure 11. Venn Diagram

N > 24 > N > 50 > N > 250 > N > 500 > N > 750 > N > 1024

The number of words with count more than 24 are highest as compared with the one with 50 count and same it is compared and in the last number of words with count 1024.
References: