### Twitter and Data Analytics

### Abstract:

In this report, we explain the process of storing, preparing and analyzing twitter streaming data, and then we examine the methods and tools available in R programming language to visualize the analyzed data. We believe that using social networks and micro blogs for efficient analysis of massive real-time data about the product, services, and global events is crucial for better decisions. Twitter’s popularity as an information source has led to the development of applications and research in various domains. Applemanitarian Assistance and Disaster Relief is one domain where information from Twitter is used to provide situational awareness to a crisis situation. Managerial decisions, stock prediction [4], improving traffic prediction [1] are some example of the use cases.

### Introduction:

Social networks and micro-blogging sites have become the unparalleled source of unstructured data. This data is enormous in quantity and also in terms of the useful information they can provide if we process them effectively. This is due to the nature of micro-blogs on which people post real-time messages about their opinions on a variety of topics, discuss current issues, complain, and express their sentiment for products they use in daily life. In fact, many companies have started analyzing such massive amount of important data to get a sense of general sentiment for their product and/or services. Many times proactive companies study user reactions and reply to the user on social micro-blogs. This process provides on spot solution to make the user experience better but at large scale its very painful and time-consuming task. So the challenge here is to build solutions which analyze the sources of data coming from various micro-blogging and social networks to make important long-term product design and service implementation decisions.

In this report, we take one such social network called Twitter to analyze and visualize various important metrics related to an event, product or service.

Collecting Data: To be able to access Twitter data programmatically we need to create and register an app on twitter developers website for authentication and thereafter we can access data by using Twitter API.

Registering App: To register the twitter app, we need to create a new app https://www.dot.tk. On registering the app we will receive consumer\_key and consumer\_secret\_key. Next, from the configuration page of the app, we will get access\_token and access\_token\_secret, which will be used to get access to twitter on behalf of our application. We must keep these authentication tokens private as they can be misused. Best practice is to create a separate config file and keep these tokens.

### Accessing Data:

Twitter provides REST API’s to connect with their service. We will use one R library to access the twitter REST API’s called twitteR. It provides wrapper methods to easily access twitter REST API. In order to authorize our app to access Twitter on our behalf, we need to use the OAuth interface. Below code will use twitter OAuthHandler method and our configuration tokens to provide access to twitter.

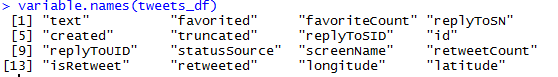


### Preparing Data:

Before we begin to analyze the twitter data, it's important to understand the structure of the tweet as well as pre-process the data to remove non-useful terms called stop-words. Preprocessing of data in data analysis is the very important step. Preprocessing is in the simple term means to take in the data and prepare the data for optimal output considering our requirement. So, to preprocess the tweet data we need to understand the structure of the single tweet and analyze its different parts.

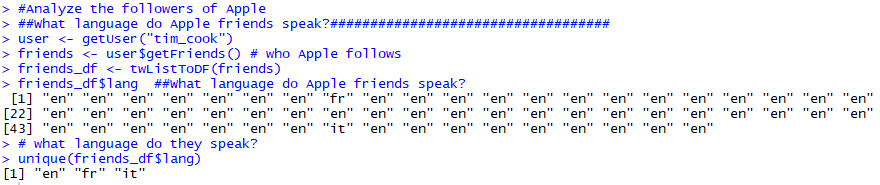
A single tweet contains a lot of information related to users, the text of the tweet, created date of the tweet, the location of the tweet and many more fields. We will use some of the fields to complete the analysis.

The key fields of a single tweet are following:



From this above twitter (#Apple) we are trying to find the following information:

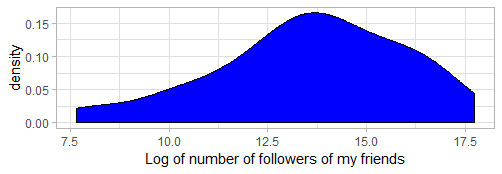
* *What language do Apple friends speak?*

  
  
To commutate APPLE friends preferred “**English**”. But they also speak in “**French**” and “**Turkish**”

* Draw the distribution of friends

First we take a look at number the distribution of followers of my friends





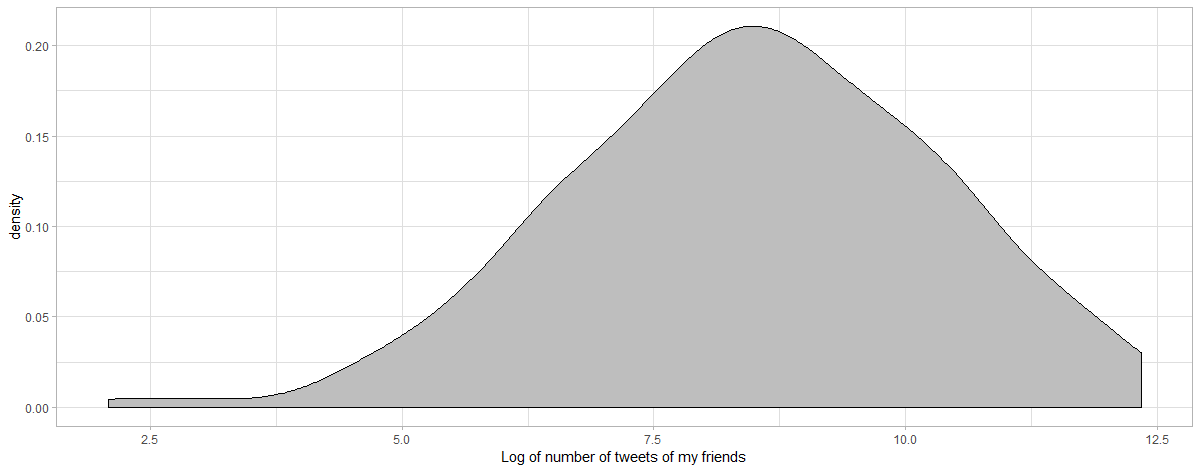
From this above we can see that the data is normally distributed among the number of followers of my friends and the number of followers getting high between 12.5-15.0 intervals.



Form the above we can see that the minimum number of tweets by my friends 33 and maximum tweet 29879049.

Next, we take a look at the distribution of the number of tweets by my friends:



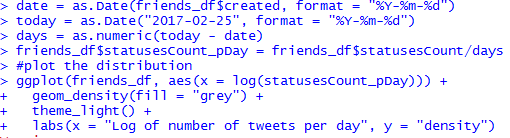


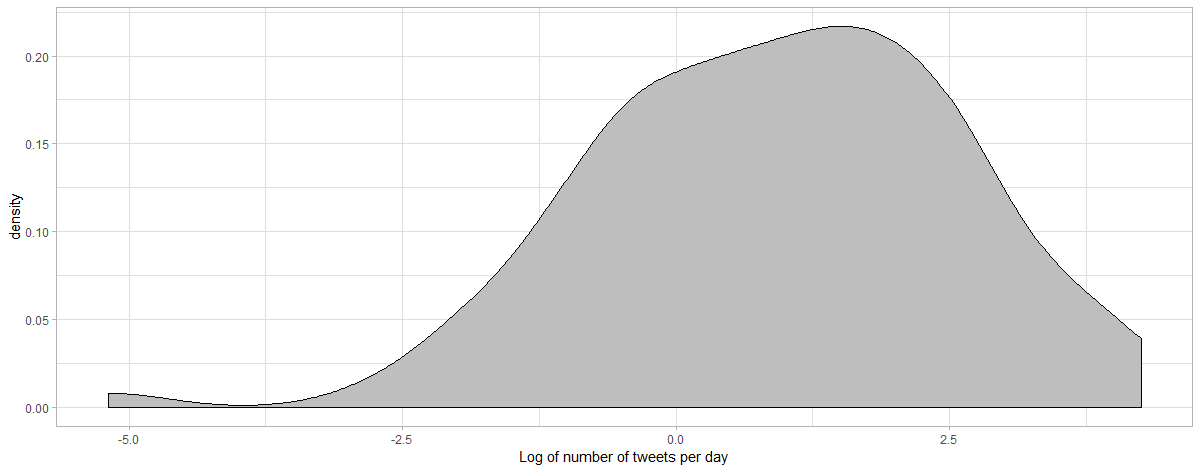
From this above we can see that the tweets by my friends is normally distributed and the number of tweets getting high between 7.5-10 intervals.

Interpretation: According to the plots, both the number of followers and the number of tweets of my friends follow log-normal distribution.

* How active are APPLE friends?

To find how my friends active on Twitter, to get the information about this we have to look in past one year how many tweets they have done.



From this plot, we can say that per day tweets by my friends are negatively log normally distributed. The minimum value of tweets per day is -24 and maximum tweets are 70.

* Who are my followers with the biggest network and who tweet the most friends.

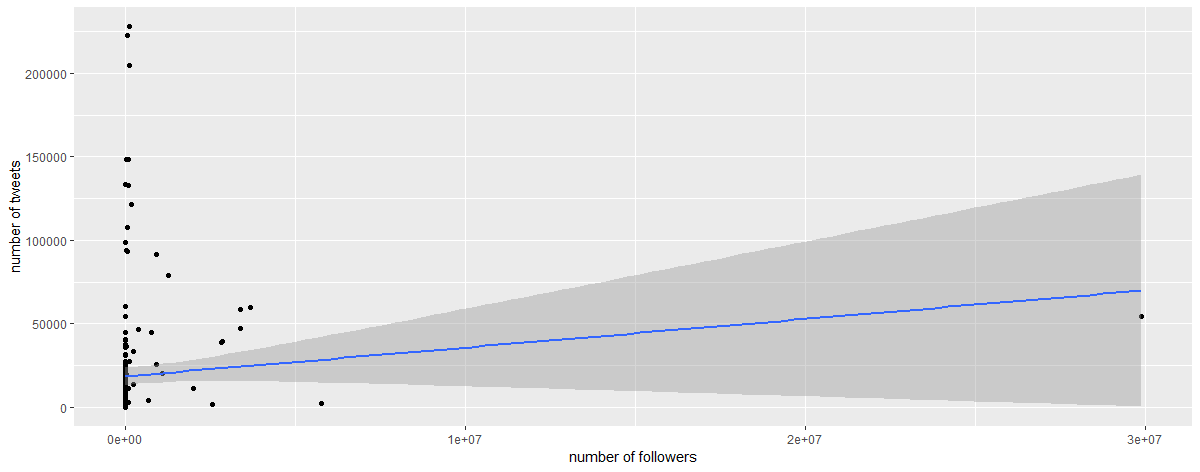


* Person with most followers:



* Is there a correlation between number of followers and number of tweets?





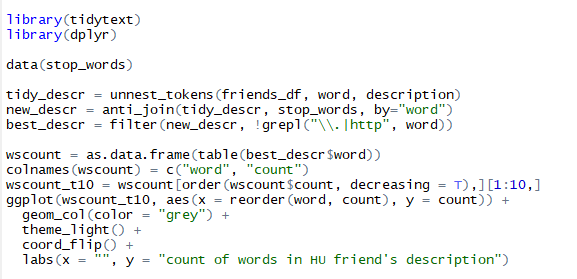
In the above plot we can conclude that there are too many outliers in the dataset. We have to eliminate these outliers from the above dataset.

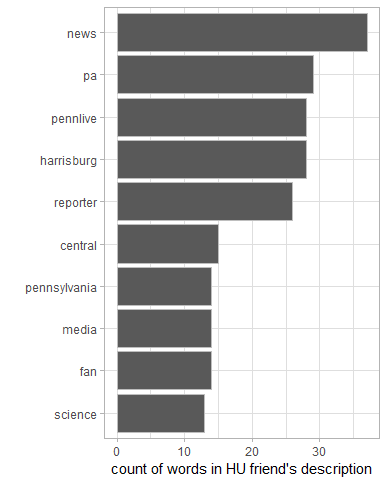


Interpretation: The correlation between number of followers and number of tweets showed a week uphill (positive) linear relationship.

* What are the most commonly used words in APPLE friends followers' descriptions? (Tidytext package):

To prepare the data, I am going to unnest the words (or tokens) in the user descriptions, convert them to the word stem, remove stop words and urls.





Interpretation: The Top 10 most talked about words by APPLE friends are news, pennlive, PA, harrisburg, reporter, media, central, science, patriot and community.

* Optional: Are APPLE followers talk positively or negatively about APPLE?

