Twitter Data Analytics

CASE I- Report

**9/29/2014**

**DS 501**

**Kafley, Bir**

**Lama, Suman**

**Ayub, Mohammed**

**Introduction:**

This report elaborates the reason behind the choice of the topic, how data was collected, discuss the results and try to draw some conclusion after looking at the processed data.

Since, this was the first project which involved data processing, and using API’s it was very critical that we had a nice outline before the project was started. The lecture, slides and the sample code present in the course website were used to refer back and as a guide while working for the project.

**Topic Choice:**

Ebola, was chosen as the topic for the project. Ebola is a new endemic disease which originated from West Africa and has stirred global attention due to its spread and with no cure yet found. It created more stir in the US because two of the US health workers serving in Africa were contaminated.

There have been recent outbreaks of several diseases but Ebola has got more global attention. With global attention towards the outbreak of disease, it could affect several other sectors such as business, security. On top of that Ebola is associated with religious values and beliefs. It looks like the transmissions of disease have been occurring while loved ones are compelled to conduct the ritual rites for the dead bodies. The data also showed use of religious word and ‘god’ frequently. Moreover causes of enough awareness and knowledge about the Ebola, there have been several incidents. Local people have been chasing health workers out and in some cases killed several people.

Despite, having chosen a topic for the project; the code has been designed to handle other inputs. During code design flexibility and performance of code was taken into consideration. For the design flexibility the program is designed to take any user input values for processing, as it would make program suitable for long period of time and give more flexibility to user. The syntax for it is: var = raw\_input(“Search Word”); where var is the name of variable where the user input value would be stored. With this, any user can enter their topic of interest and do analytics involved in it.

Code used in the program:

search\_word = raw\_input ("Enter the topic of your choice: ")

Data Collection:

This project is about the twitter feeds, so data was collected via creating twitter application. Python was the choice language as python is object oriented as well as script based language. Python also has built in packages for twitter. There are so many other built in packages available for processing of data and packages to create diagrams/tables to analyze the data in user friendly way.

Historical as well as live streaming data were collected so that we could understand the trends, as Ebola has been for a long time. Also, Ebola is a global topic data was collected on different time so that data had better representation of all different time zones used around the world. Language filter was used to see if there was anything we could find. Arabic was the choice language as it is the most common language in Africa. Similarly, data was collected without specifying any language filter so that we could include various different languages that are used in twitter.

Raw Data was collected in the .txt format and saved so it could be better analyzed. Functions were written to read the raw data from the file and plot the graphs and tables as requested. Line Histograms were created as they are able to clearly show the results in pictorial diagram. Tables were also created to list the most common words used in the status.

**Data Analysis:**

Two most common regions that came along with Ebola were US and Africa. The other point worth mentioning here is that the use of word ‘HolyQuaran’, which shows the religious connections associated. ‘Survival’ seems to be the other word that came often and makes sense. There were some other words like ‘sexy’ mentioned 109 times when tweets were collected during Africa evening time. This word wouldn’t make lot of sense to Ebola.

US came frequently than any other region. This could be taken in several ways, and one of them could be that number of twitter users in the US could be greater. Also, the recent news regarding two US health workers being associated with Ebola could have triggered it. Analysis of data collected during various time amplifies noticeable difference of mentioning of ‘US’. Data collected during the evening in US; showed US as one of the top frequently used word. On the other hand data collected during the office hours in the US, or mid-day in West Africa didn’t had any significant number of ‘US’ mentioned.

Despite having seen the occurrence of ‘US’ in several tweets it was compared with the tweets collected during the office hours in the US, and midday in West Africa. ‘US’ didn’t even make it to the top 30’s in the list of most frequently word used for the office hours data. But, Africa was one of the most frequently word. This could be assumed to be correct as Ebola has affected Africa than any other part of world at this point.

The data collected right after the announcement made by World Health Organization to build Ebola center headquarter in Ghana, had ‘Ghana’ mentioned than any other region. For anyone who didn’t read the announcement of WHO regarding Ebola center could make a good guess after seeing the data coming from twitter.

Language filter was used and feeds in Spanish and Arabic were searched. There were not enough data to collect and associate with results. This shows that English as the most dominant language being used in twitter to express even across various regions. There were some tweets in other languages such as ‘Japanese’; ‘Chinese’ which were captured when there was no language filter in the query. It leads us to believe that people in various regions are aware of the region and are expressing their concern. The better approach would have been to look at the geo-location of the origin of the tweets. But, on the other hand that could also mean people of English origin who are spread all over the world are expressing their concerns. It was assumed that tweets presented in various languages would give more validity to the global reach of the topic.

Majority of the tweets were retweeted. There were other connecting words which came on the most frequently used list but don’t add up to the meaning. ‘Outbreak’, ‘Village’, and ‘survives’ were the other three words that came frequently. These three words provide more meaning to Ebola, as they are all related to Ebola.

Overall, tweets were able to reflect the current story and information obtained after processing the raw data was very meaningful.

**Results:**

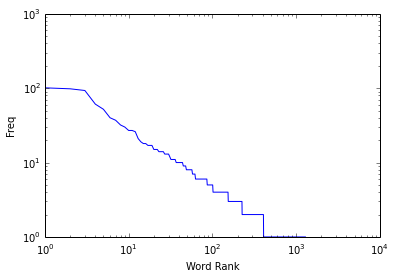
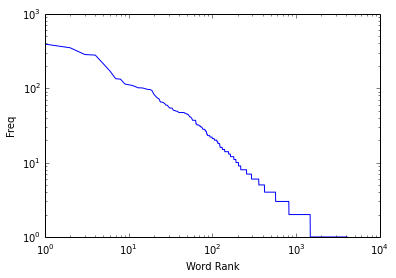
 

Fig 1: Small Set (200 tweets) of Evening Data. Fig 2: Big Set (1000 tweets) of Evening Data.

The two figures above clearly reflect that there is not any much different between two pools of data. The number of tweets collected and analyzed doesn’t seem to impact the outcome.

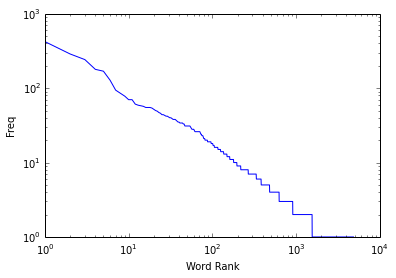
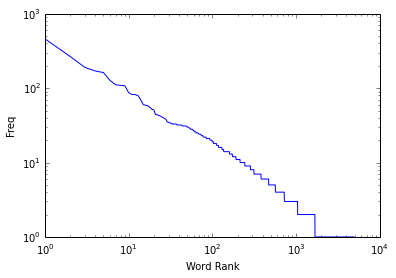
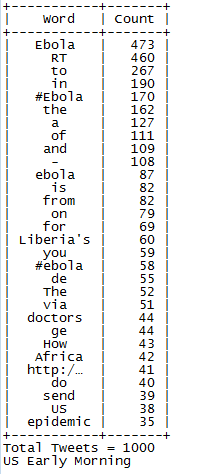
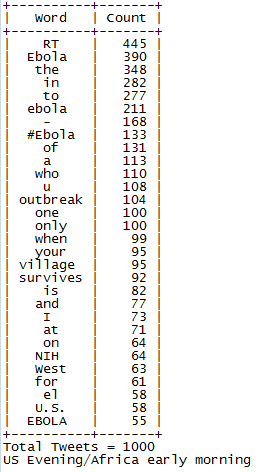
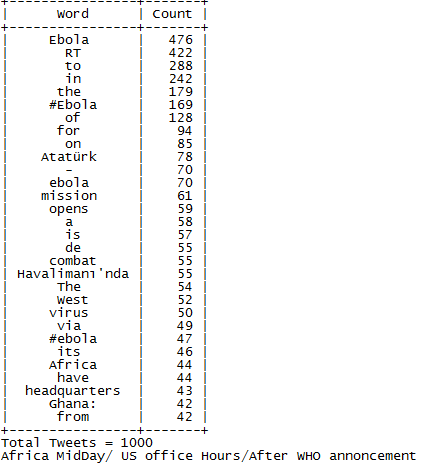


Fig III- US Early Morning Tweets Fig IV- US Office Hours/Africa Midday Tweets

The above two graphs when looked closely, resemble that there is slight difference how the line falls down. In Fig III there is decrease in the slop and seems to try to rise again after some time. But, fig IV shows that there is no steep decrease in the slope. On the other hand, the overall graph looks similar.

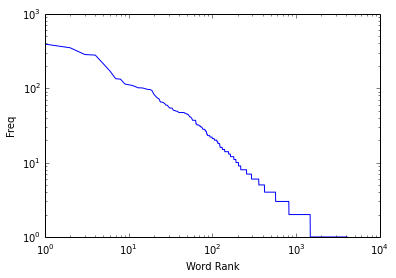
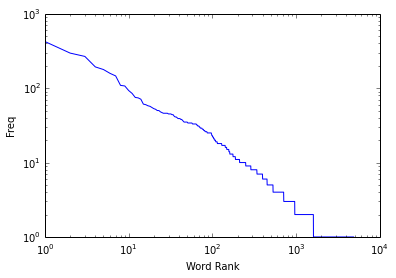
****

Fig V – US Evening Tweets Distribution Fig VI – Africa Evening Tweets Distribution

Looking at the above two diagrams we can analyze that there is not similar transitions between the curves. The two curves show that how use of words in various regions varies.

**Conclusion**:

The tweets collected from the twitter were difficult to comprehend initially, but with the help of packages available in python and twitter api, information were generated from RAW data. The informative data were then used to draw the figures and tables so that data could be easily grasped by any user.

Various data collected revealed several facts, such as tweets with most use of word ‘US’ showed the number of twitter users, or global impact of US. Data collected after a recent announcement showed that people expressing their feelings after the announcement. Most retweets reflected how people are openly sharing the information.

Since, it was the first project and we were not equipped with more information on data analytics tools at times it felt that there was a roadblock either due to coding or due to lack of proper experience. But, in the end this project made us believe that there is vast amount of data floating around which could be analyzed and turned into meaningful data.

**Appendix**:

1. Code Attached
2. Table Attached
3. DS -501 Slides link:

<http://users.wpi.edu/~xkong/course/ds501/f14e/index.html>

1. Twitter API Documentation

<https://dev.twitter.com/overview/documentation>

1. Python Documentation

<https://docs.python.org/2/>