**SENTIMENT ANALYSIS, TEXT DATA VISUALIZATION AND GEO-MAP VISUALIZATION**

PROJECT OPTION-3

BY- ANAMIKA SINGH

STUDENT ID – 109450027

TEXT DATA ANALYTICS

PROF ONOOK OH

12/05/2020

**EXECUTIVE SUMMARY**

The goal of the assignment is to perform sentiment analysis on the data extracted from Twitter using Twitter API. The API helped to extract the Twitter data in a structured format. The data was further cleaned to perform sentiment analysis. The analysis is performed on the Samsung and Apple products. The aim is to get an overall sense of the tone of tweets on Samsung and Apple products. Sentiment analysis helps to determine the positive and negative sentiments behind the tweets. Sentiment analysis is performed using bing, nrc and afinn dictionaries. These dictionaries help to obtain the positive and negative sentiments about Samsung and Apple products by comparing each individual word against the lexicon, it also helped to extract the sentiment scores from Twitter feeds. Word cloud is used to visualize all the words together that are used for Samsung and Apple products. Bigrams are also created to understand some of the top words used together as bigrams for Samsung and Apple. For further analysis word network is prepared to observe the relationship between the words. To find out the words used for Samsung and Apple products comparison cloud is prepared. Some data also contains the geo location of the users who wished to share their location on Twitter. By collecting those data geo-maps are created to show their location on world map. Separate world maps are created to show the countries from where tweets are posted on Twitter regarding Samsung and Apple products. The results of the sentiment analysis are visualized by using different text data visualization options.

**INTRODUCTION**

The purpose of the assignment is to determine the sentiment of the people on Samsung products and Apple products as expressed on Twitter by their tweets. Twitter users post tweets to share their views, suggestions, and interests. Analyzing each post and understanding the sentiments associated with the post helps to understand the positive and negative sentiments of the people. The assignment compares the sentiments related to the Samsung products and Apple products by analyzing Twitter tweets over time and by visualizing the most frequent words used for Samsung and Apple products.

Sentiment analysis of Samsung products and Apple products is important because by performing sentiment analysis companies can understand how customers feel about their products. Tweets posted by people on Twitter can be very expressive and help companies to understand the emotions and sentiments of customers. It can also help companies better analyze customer feedback and opinion and optimize their strategy according to the reviews provided by customers. Analyzing the tweets helps company to find out what people are talking about their products and if the sentiments related to their products are positive or negative. Sentiment analysis provides insights into customers behavior, what they want, their likes and dislikes about the products which is very valuable for the companies.

**DATA COLLECTION**

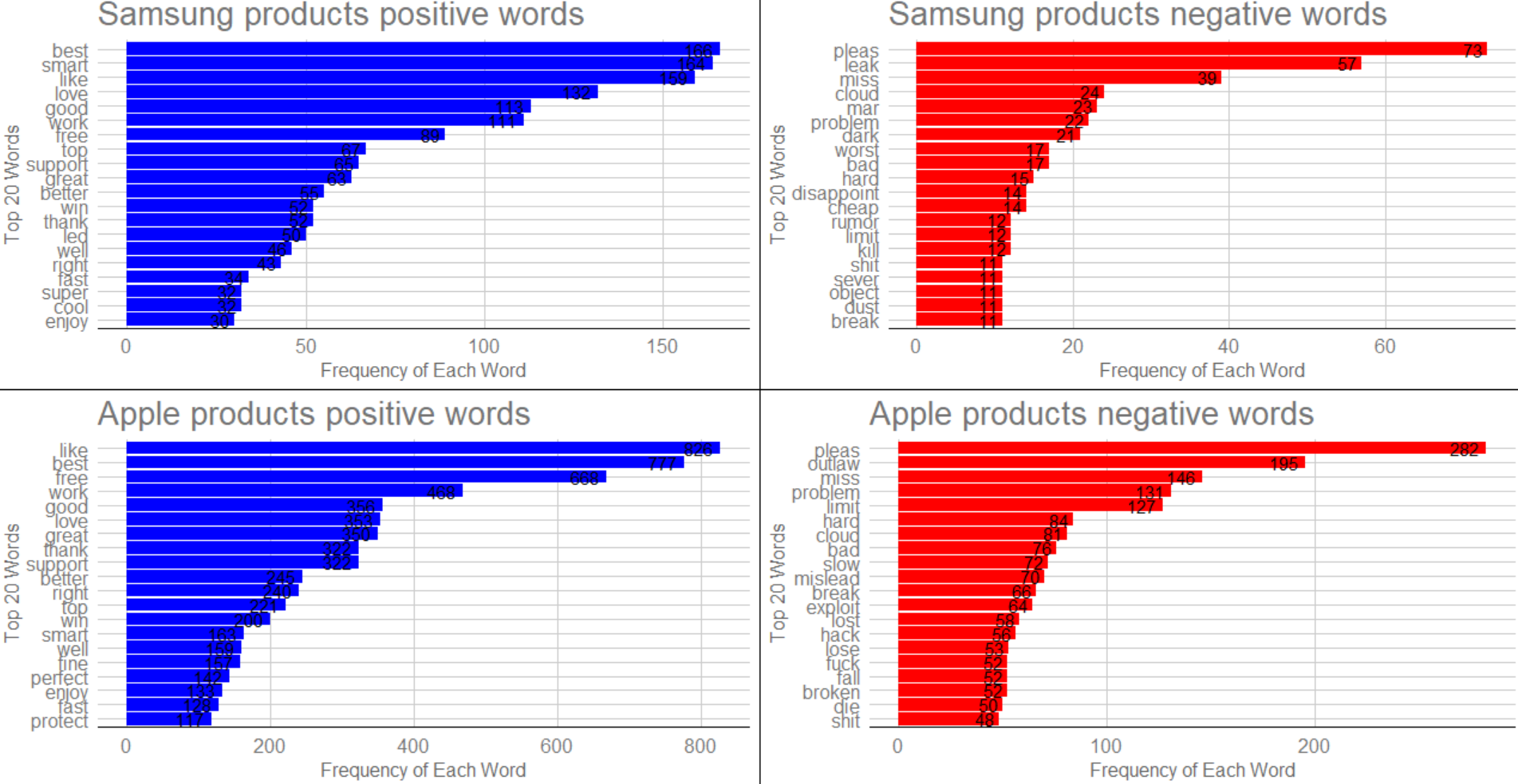
Data is collected from Twitter by creating a Twitter account. Using Twitter account application is created. Twitter provides API key, API secret, access token and access secret token. These tokens and keys are used to extract data from Twitter in R. Twitter API requires authentication to retrieve tweets. Required packages are installed in R to extract data from Twitter. By using hashtags related to Samsung products and Apple products data is downloaded in R. Using the search tweets function in R, thirty-five thousand tweets were searched. The data collected is in structured format. Twitter data contains tweets, location, geo coordinates of the twitter user if they wish to share their location. Country data will be useful to locate user’s country on world map.

The data collected from Twitter is exported to the comma separated file. The csv file is used for sentiment analysis on Apple and Samsung products. The unique data is collected by removing duplicate data based on the tweets. The tweets are further cleaned to remove hashtags, Twitter handlers, URL, Unicode, and ampersand from text data. The dataset is separated using hashtags column for Samsung and Apple products to build corpus and perform sentiment analysis and visualization.

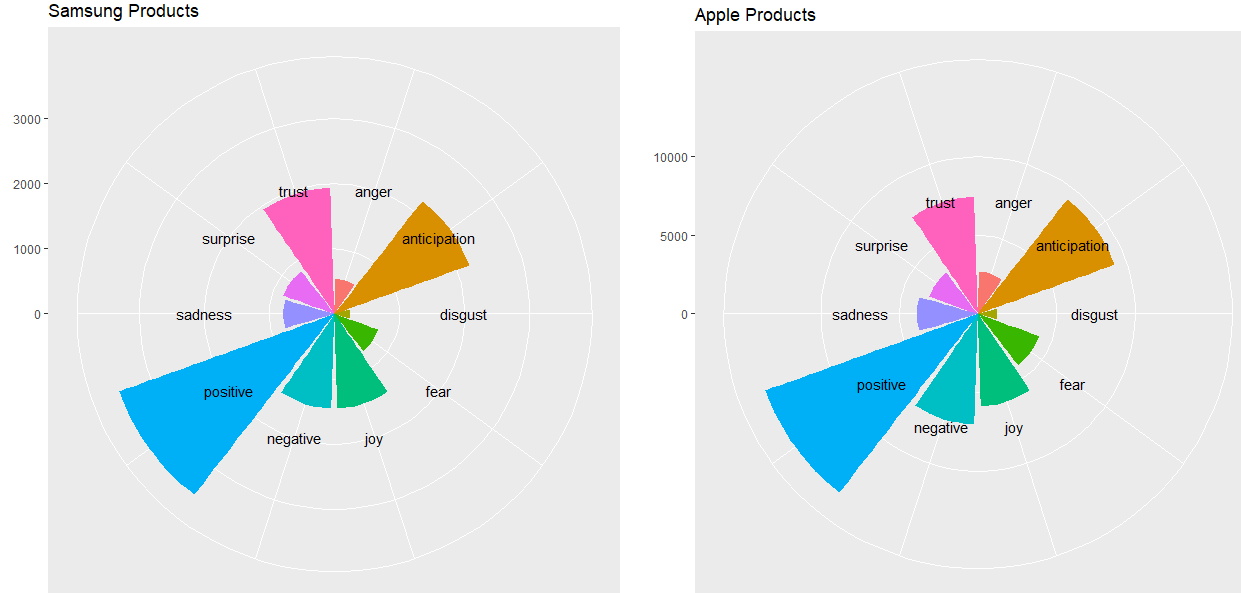
|  |  |
| --- | --- |
| **Sample size collected** | 32162 observations and 90 variables |
| **Sample size used** | 25177 observations and 6 variables |
| **Variable names** | text, hashtags, country, geo\_coords, coords\_coords, bbox\_coords, products |
| **Description of variable names** |  |
| text | Tweets posted by users on Twitter |
| hashtags | Hashtags used by user’s |
| country | User’s country |
| geo\_coords, coords\_coords, bbox\_coords | location of the users from where tweets are posted |

**DATA ANALYSIS**

After the initial data cleaning process corpus is created for Apple and Samsung product tweets using tm package. The corpus needs pre-processing that includes changing letters to lower case, removing punctuation, removing numbers and white spaces. Term document matrix is build using the corpus. Bar graph and word clouds are prepared to get most frequent words on Twitter used for Samsung and Apple products [Exhibit A]. Sentiment analysis is performed by calling bing, nrc and afinn dictionaries. With the help of bing dictionary positive and negative words related to Samsung and Apple can be visualized using bar plot and word cloud. we can see that best, like and love are some of the top positive words used for Samsung products and please, problem, are some of the top negative words used. Like, best, good are some of the positive words used for Apple products and please, limit and break are some of the negative words used.



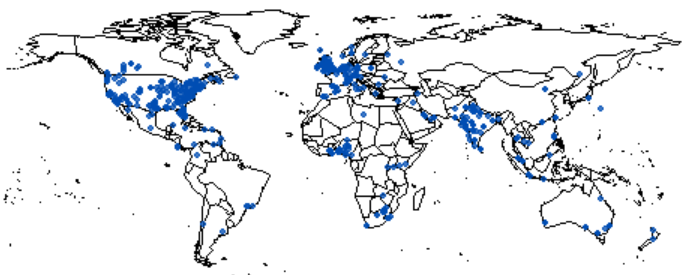
nrc dictionary is used to plot all the emotions present in the tweets by joining the nrc dictionary to Samsung and Apple dataset. Positive, trust and anticipation are the emotions high for both the company’s products.



To calculate sentiment score afinn dictionary is used. The polarity of each tweet is calculated, and the result is visualized using bar plot. Positive sentiment score is more than the negative sentiment score for Samsung and Apple products. [Exhibit B]

Bigrams are visualized using bar plots to understand which words are used together. Word network graph using igraph is used to check the words used with “Samsung. Fan shaped dendrogram is also prepared using Euclidean method by installing ape package and visualized using plot function. Plotting bigrams on bar plot shows- new phone, deal Friday are some of the top words used for Samsung products and black deal, app best are some words used together for Apple products. Dendrogram also shows the similar results as bigrams for words that are used together. Using package plotrix word that are common in both the datasets are visualized. Love, like, leak and support are some common words used for both the products [Exhibit C]. Comparison word cloud is built using the Samsung and Apple dataset to compare the words used for Samsung Apple products. It shows the words used frequently for Samsung and Apple products. Love and smart are the words used most frequently for Samsung products and free and like are the most frequent words used for Apple products [Exhibit D].

Geo map visualization is done using world map. With the help of geo coordinates present in the Twitter for 140 users, points were marked showing the location of the users on the world map. The Twitter data is also merged with in built world map data to visualize the country from where tweets are posted for Samsung and Apple products. Plotting country data on world map shows USA, UK, India are the locations where most of the tweets are posted for Samsung products. USA, UK are the locations from where most tweets are posted for Apple products. [Exhibit E]



**CONCLUSION**

Positive sentiments are high for both the companies when compared with all the other sentiments. Disgust, anger, fear sentiments are the lowest. Samsung products have fewer negative sentiments in comparison to Apple products. Like and best are the positive words used for both the companies. Please and problem are the most used negative words used for both companies. People have great trust in Apple and Samsung products. Love and support are the common words frequently used for both the company’s products. Most of the tweets are posted from Eastern USA for Samsung and Apple products. More tweets are posted for Apple products from USA in comparison to Samsung products, result is based on the users who wished to share their country on Twitter [Exhibit F].

**APPENDIX**

Exhibit A. Bar plot of top 20 words most frequently used for Samsung and Apple products.

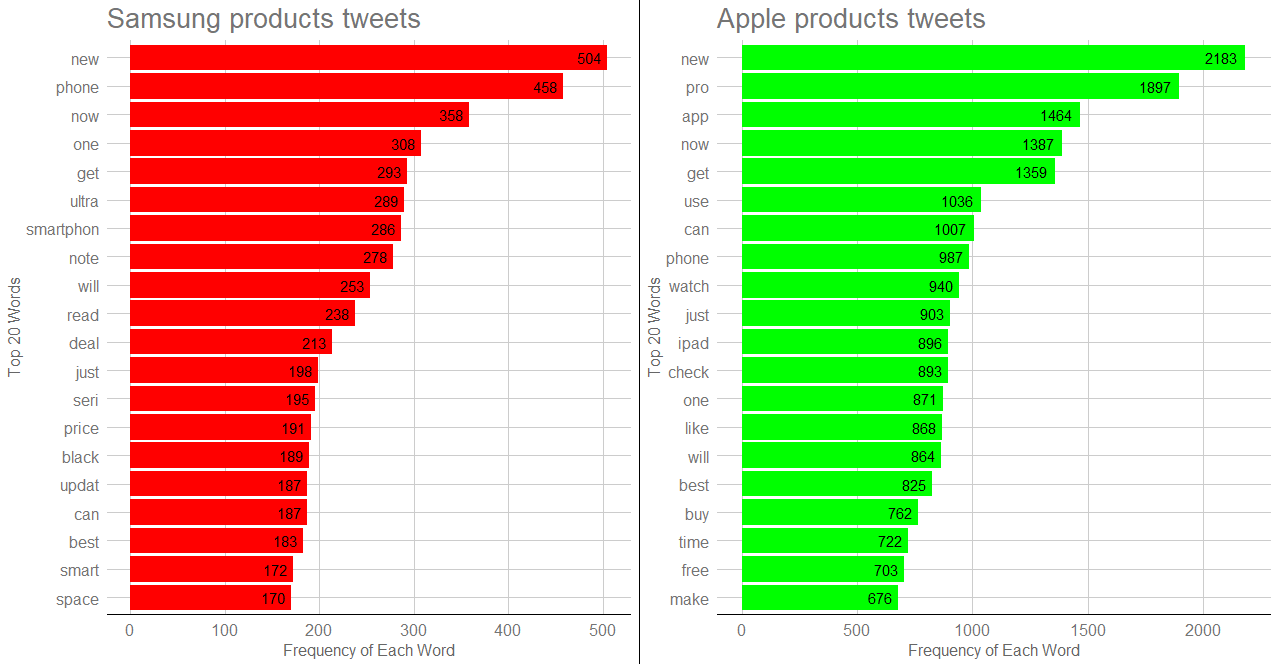


Exhibit B. Sentiment score for Samsung and Apple products

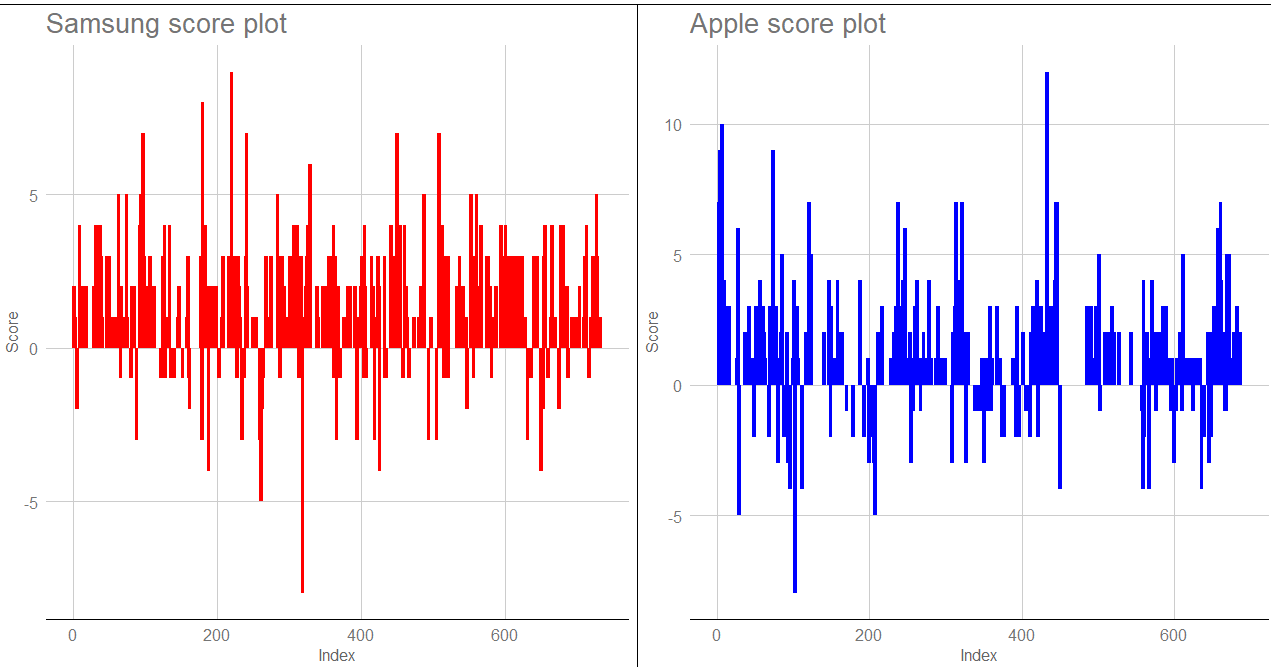


Exhibit C. Common words

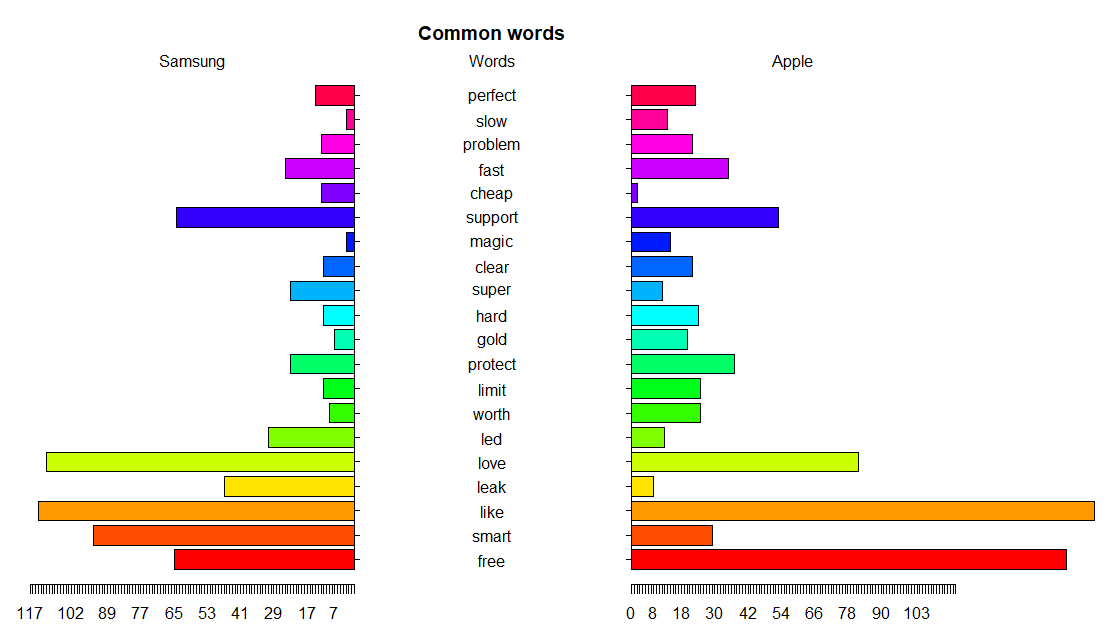


Exhibit D. Comparison Cloud

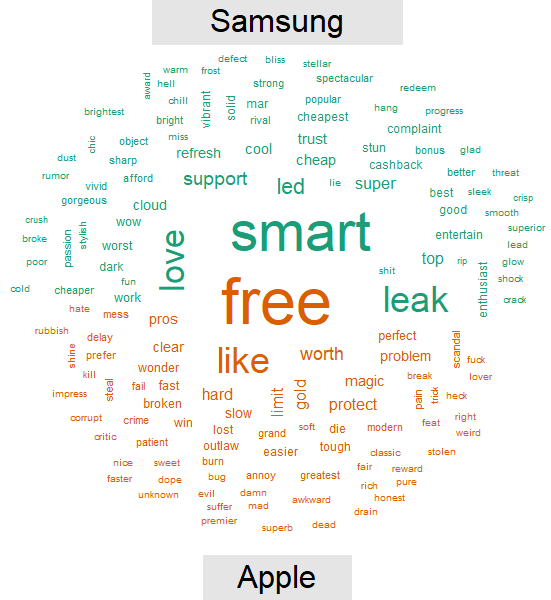
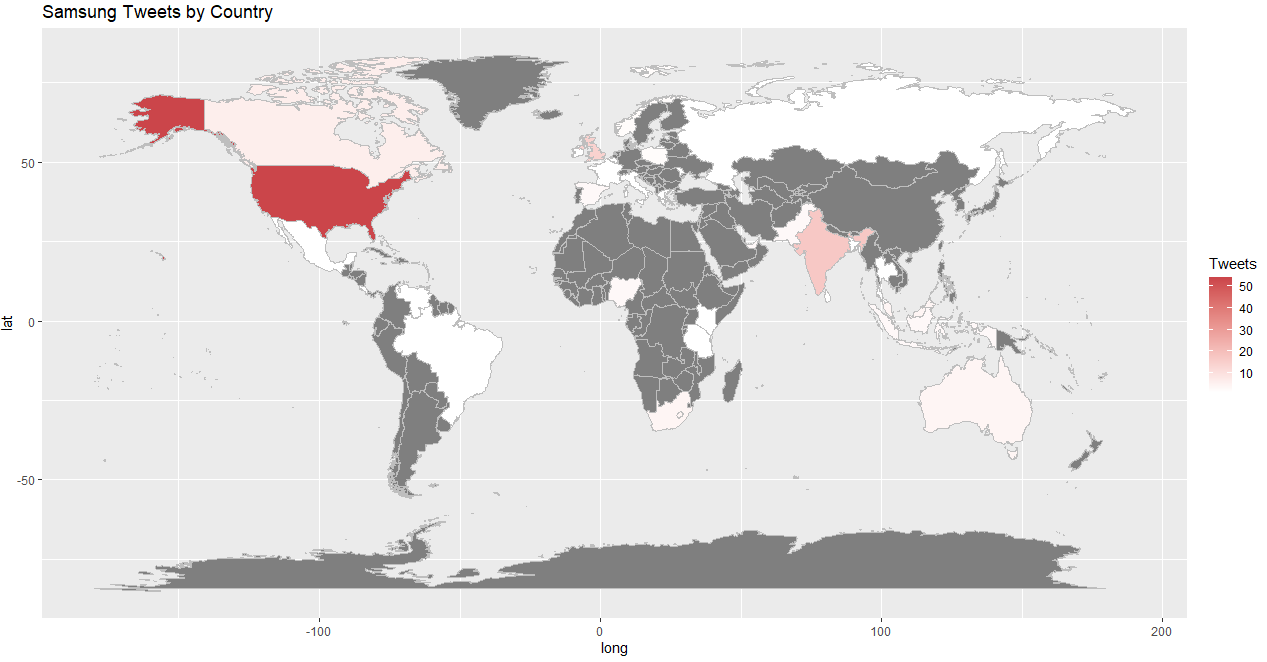


Exhibit E. Tweets location for Samsung and Apple products



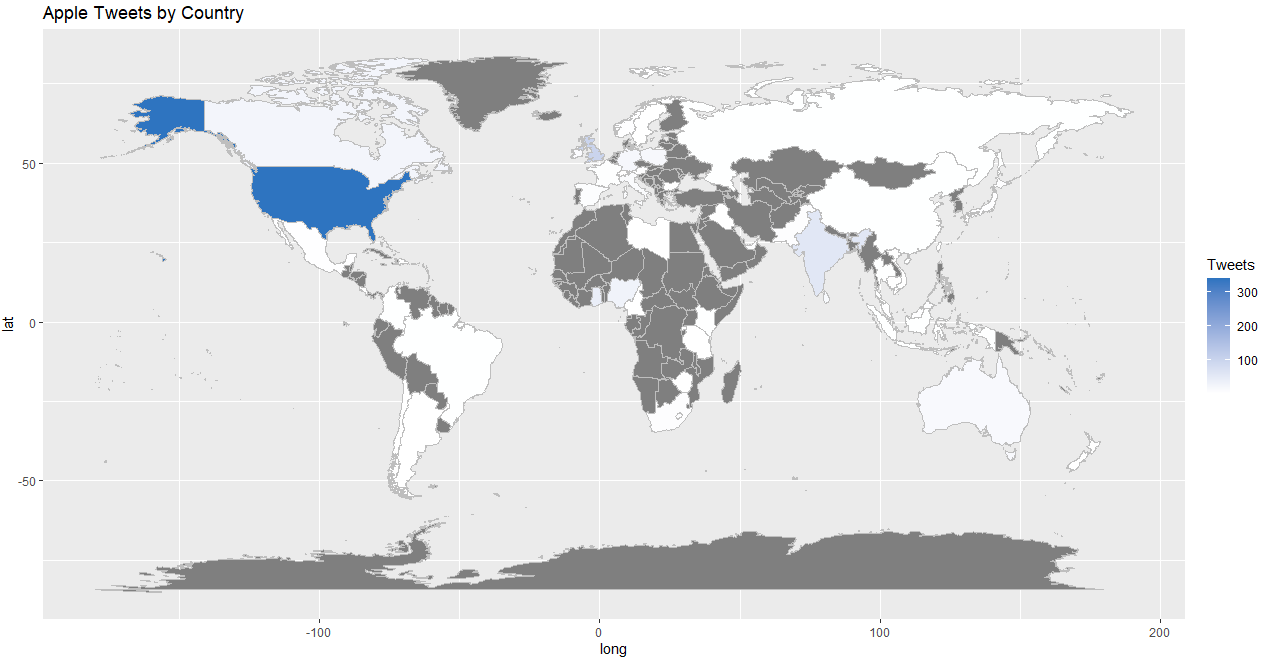


Exhibit F. Difference in number of tweets by country

