19OH01- SOCIAL AND ECONOMIC NETWORK ANALYSIS

BABITHA REDDY C H - 18Z211

M BHAVYASREE - 18Z214

CHERUKURI SHALINI - 18Z215

KEERTHANA.B -18Z226

KOPPETI JOSHNA - 18Z332

Group Report submitted in partial fulfilment of the

requirements for the degree of

BACHELOR OF ENGINEERING

Branch: COMPUTER SCIENCE AND ENGINEERING

Of Anna University APRIL 2021

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE – 641 004

Table of Contents

1. PROBLEM STATEMENT

A great activity concerning COVID-19 has emerged on Twitter when on March 11, the World Health Organization declared it a pandemic. As the social media platforms can help track natural disasters in real-time, the most preferred platform Twitter is used in this project to study discourse around the novel virus using some data science techniques.

Social media, in particular, plays an important role in successfully communicating risk to the larger public. Monitoring public discourse on social media during a pandemic situation is critical to evaluate the effectiveness of risk communication efforts. We research discourse on Twitter around coronavirus disease 2019 (COVID-19) which focuses on extracting themes from tweets mentioning “Coronavirus” and “COVID-19” and other hashtags. Tweets can be either a reply to another tweet and/or a quote. Otherwise, it is just a simple tweet. We use these properties to see whether debate or sharing explains an activity on Twitter around a specific topic. Hashtags and mentions in the tweets are taken into account as graphs and analyzed.

2. DATASET DESCRIPTION

According to the WHO (World Health Organization), risk communication is essential to help people understand the cause, how to prevent the cause, stop the spread of disease, and limit the social and economic impact of an outbreak. Social media, in particular, plays an important role in successfully communicating risk to the larger public. The dataset contains variables associated with Twitter: the text of various tweets and the accounts that tweeted them, the hashtags used sand the locations of the accounts.

Due to the large volume of Tweets, there may be some gaps for some hashtags (not all Tweets with a given hashtag may be captured). Because some hashtags are used less frequently than other hashtags, less frequently used hashtags may span a longer period of time (going back earlier) than more frequently used hashtags. The hashtag "#coronavirus" seems to be the most frequently used - despite scraping 500,000 Tweets, there was no overlap between Tweets with this hashtag in version 1 and version 5, therefore gaps remain. The retweets argument has been set to FALSE, so this dataset does not include retweets (although a count of retweets is provided as a variable). A dataset containing these tweets were considered in the middle of April 2020.

3. TOOLS USED

PYTHON:

Python is a widely used scripting language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Python consists of various libraries which help view and analyze various networks by plotting graphs.

Jupiter Notebook: The Jupiter Notebook has been used to execute PYTHON 3 code for dataset cleaning and analysis.

These are the following libraries that have been used in this project:

1. **Pandas:**

   pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. In particular, it offers data structures and operations for manipulating numerical tables and time series. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.

2. **Numpy:**

   NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

3. **Plotly:**

   plotly.py is an interactive, open-source, and browser-based graphing library for Python sparkles. Built on top of plotly.js, plotly.py is a high-level, declarative charting library. plotly.js ships with over 30 chart types, including scientific charts, 3D graphs, statistical charts, SVG maps, financial charts, and more.

4. **Seaborn:**

   Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

5. **NLTK:**

   Natural Language Processing with Python NLTK is one of the leading platforms for working with human language data and Python, the module NLTK is used for natural language processing. NLTK is literally an acronym for Natural Language Toolkit. It helps us to tokenize data by words and sentences.

6. **Wordcloud:**

   Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant

textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

7. **Advertools:**

   advertools is a Python package that manages, manipulates, visualizes, communicates, understands, and make decisions based on data.

8. **Matplot:**

   Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack

4. CHALLENGES FACED

- The dataset used in this project initially had a number of empty columns that were unnecessary. Those columns had to be removed and the dataset had to be preprocessed.

- The dataset contains language column that had various languages other than English. Those rows had to be removed and only English tweets were taken into account for further analysis.

- Due to the large volume of Tweets, there may be some gaps for some hashtags (not all Tweets with a given hashtag may be captured).

- Due to the large volume of the dataset, the execution process takes more time and thus, system may freeze at times.

5. CONTRIBUTION OF TEAM MEMBERS

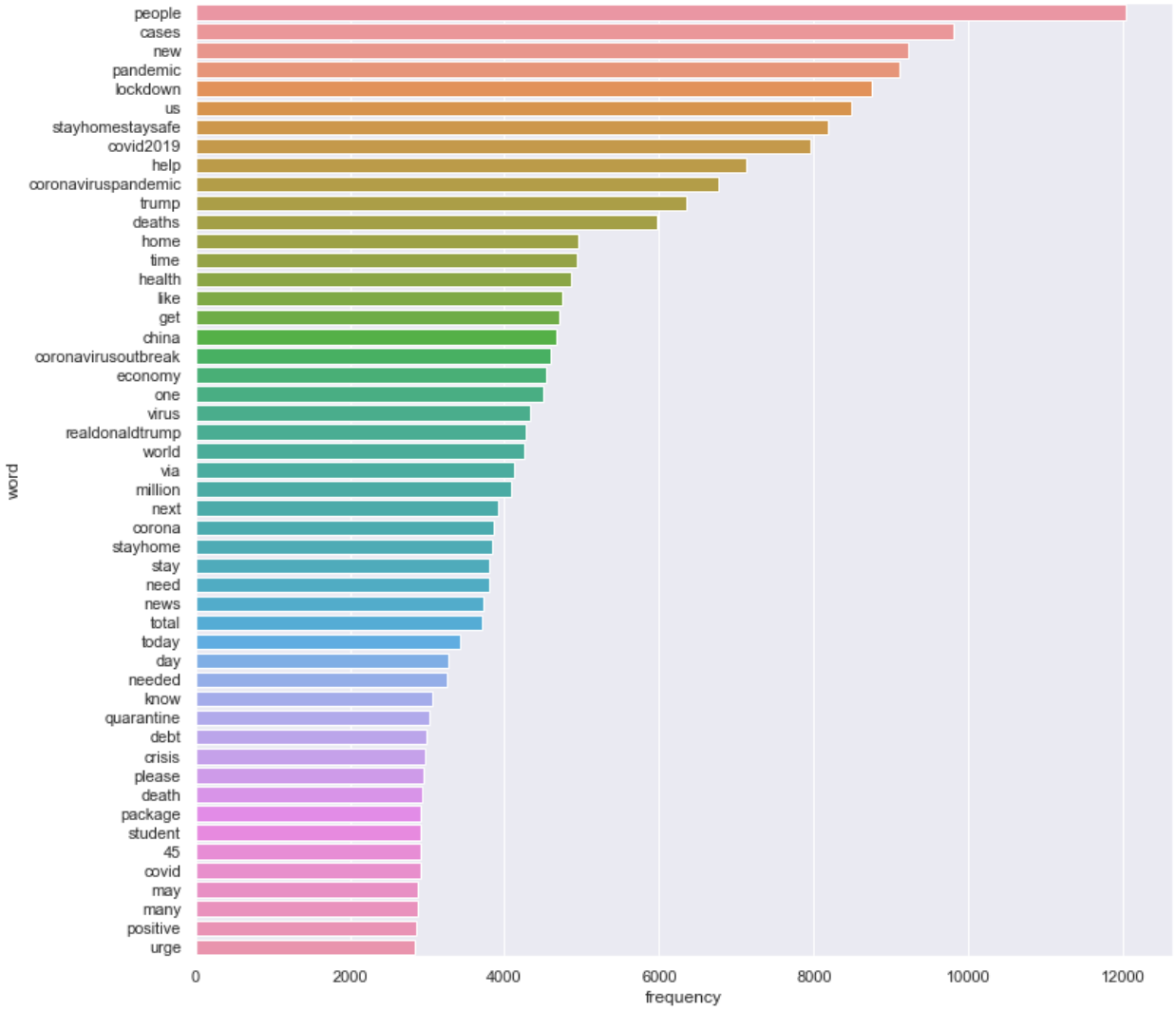
Roll no	Name	Contribution
18Z211	Babitha Reddy C H	Preprocessing and frequency of words
18Z214	M Bhavysree	Word cloud and sentiment analysis
18Z215	Cherukuri Shalini	Word cloud based on sentiment scores
18Z226	B Keerthana	Hashtag analysis
18Z332	Koppeti Joshna	Top hashtags and mention analysis

6. ANNEXURE I: CODE

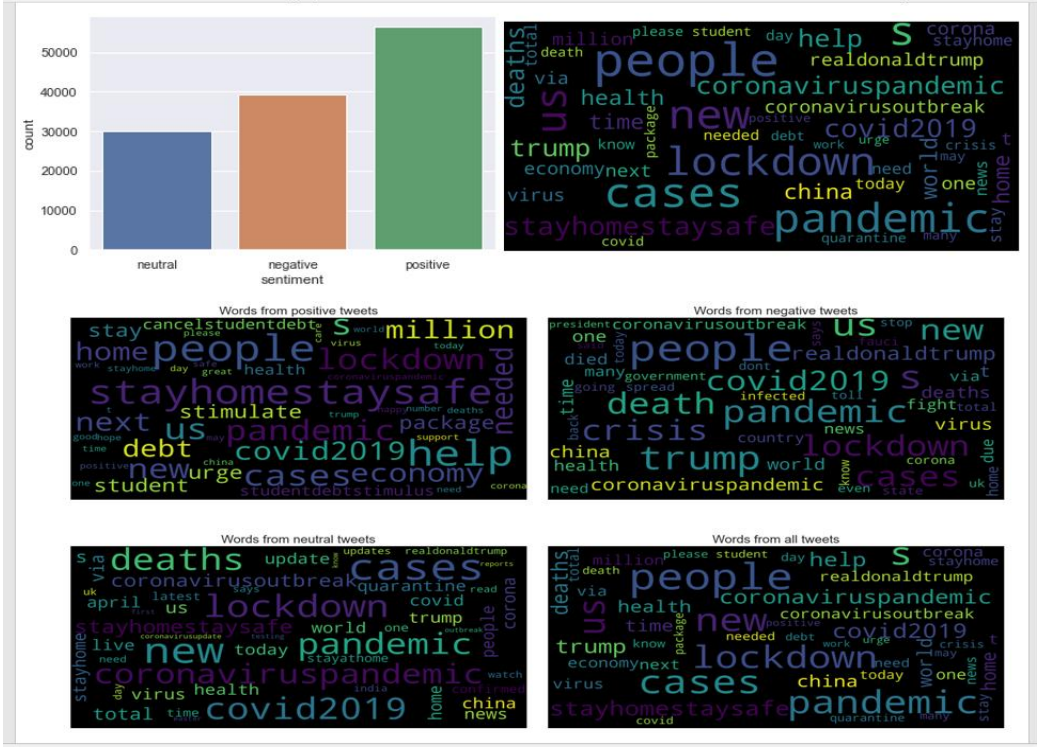
https://github.com/bhavyasree0108/Visualising-twitter-discourse-on-Covid-19

7. ANNEXURE II: SCREENSHOTS OF THE OUTPUT

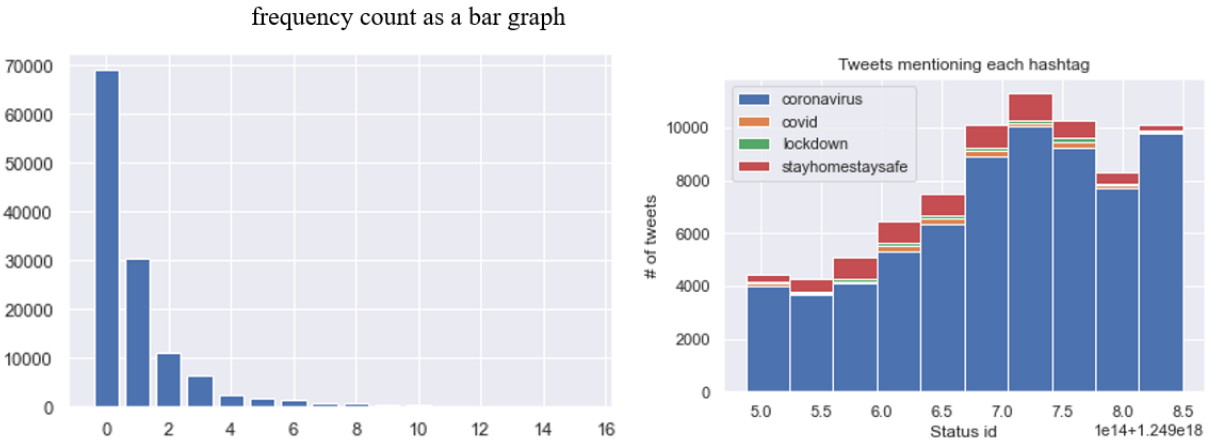
Frequency of the words:



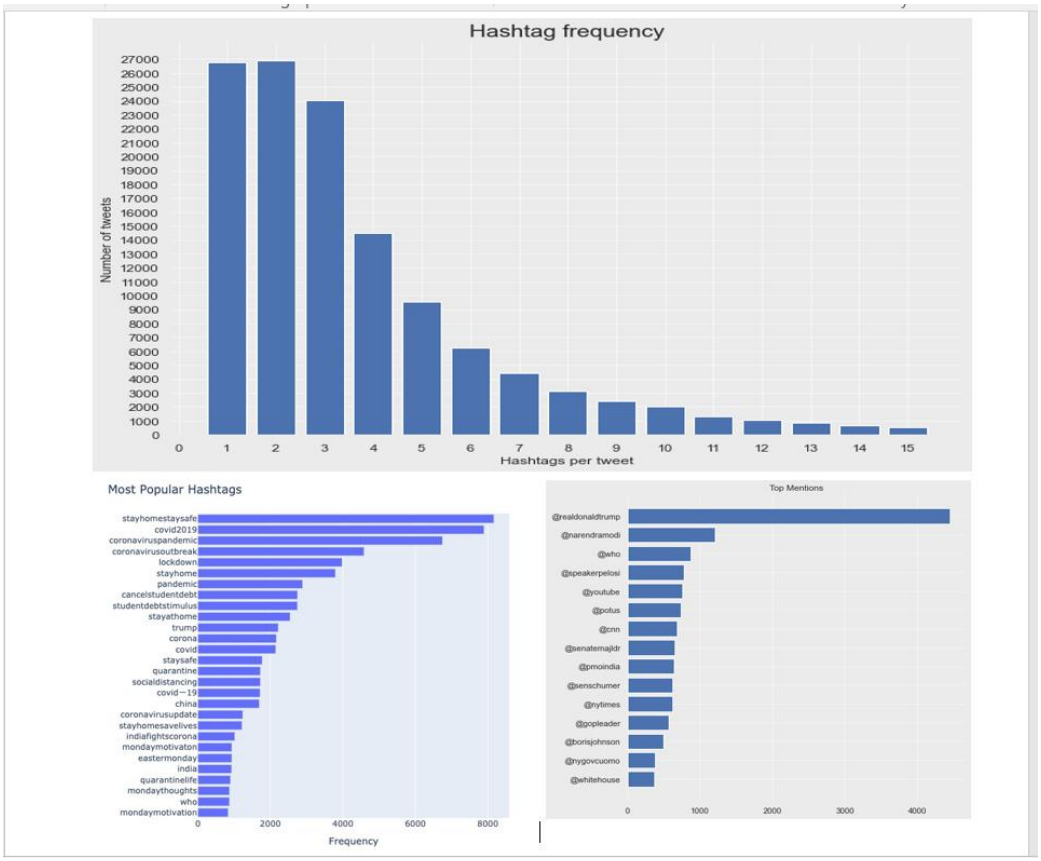
Word cloud and sentiment analysis:



Hashtag analysis – bar graph and histogram:



Hashtag frequency, Top hashtags and mentions:



9. REFERENCES

1. https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/
2. https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d
3. https://www.kaggle.com/jagangupta/wordcloud-of-tweets
4. https://realpython.com/python-nltk-sentiment-analysis/
5. https://www.datacamp.com/community/tutorials/wordcloud-python
6. https://www.cambridge.org/core/journals/disaster-medicine-and-public-health-preparedness/article/social-network-analysis-of-covid19-public-discourse-on-twitter-implications-for-risk-communication/5CF2824263F23693F9AEFB4A5E56880A
7. https://www.analyticsvidhya.com/blog/2021/02/sentiment-analysis-predicting-sentiment-of-covid-19-tweets/
8. https://www.makeschool.com/mediabook/oa/tutorials/tweet-generator--data-structures---probability-with-python/analyze-word-frequency-in-text/
9. https://towardsdatascience.com/covid-19-data-visualization-using-python-3c8bcfaeff5f
10. https://www.kaggle.com/smid80/coronavirus-covid19-tweets