DATA ANALYSIS REPORT

Majed Bawarshi, Hasan Sidawi

B1605.090072, B1705.090059

A) Project Name: Sentiment Analysis on corona hashtag tweets.

Your Data Won’t Speak Unless You Ask It The Right Data Analysis Questions.

0) Why did you choose this subject?

Answer : Lately, the world is facing a high crisis regarding Coronavirus (COVID-19). The virus has badly affected people's health due to its fast-spreading ability. However, the virus's affection on people's sentiment has not been measured yet, therefore, we thought to analyze people's sentiments regarding the spread of news about the virus. Using tweets sentiments analysis under both Corona and COVID-19 hashtags in the most common global media twitter accounts, such as **CNN**, **FoxNews**, **BBC**, **Ajazeera**, and **WHO**. And that will be achieved by comparing the people sentiments throughout their tweets at the beginning of the virus spread until it became a global pandemic.

1) What exactly do you want to find out?

Answer : We want to find out how the spread of the virus affected the people's sentiments. We will compare people's sentiments using their tweets that have the hashtag of the virus that were published between **January** and **April** 2020.

2) What standard parameters (features) will you use that can help?

Anwer : We will use sentiment analysis libraries in Python so that we measure the positivity and negativity of the tweets in the range (0-100).

3) Where will your data come from?

Anwer : Our data will come from twitter tweets using the Python library “twint” to help us scrap old tweets between given dates.

4) How can you ensure data quality?

Anwer : We will ensure the quality of our data by cleaning it up from useless/random terms such as emojis, redundant words, punctuation, and stopwords.

5) Which statistical analysis techniques do you want to apply?

Anwer : We want to find the regression, mean, standard deviation, min, max of sentiments based on the time interval and we want to model them as well.

6) Who are the final users of your analysis results?

Anwer : Data analytics scientists, psychologists, psychiatrists and who are curious about this type of researches.

7) What else do I need to know? If I want to have more meaningfull results.

Anwer : Extracting the relation between people's sentiments and the spread of the virus globally.

8) What data visualizations should you choose?

Anwer : Bar graphs, Regression, Pie Charts, Cloud chart(for the most bad word used).

B) 15 Questions related with data and subject

1) What is the percentage of **positivity** rate in the public sentiments regarding coronavirus?

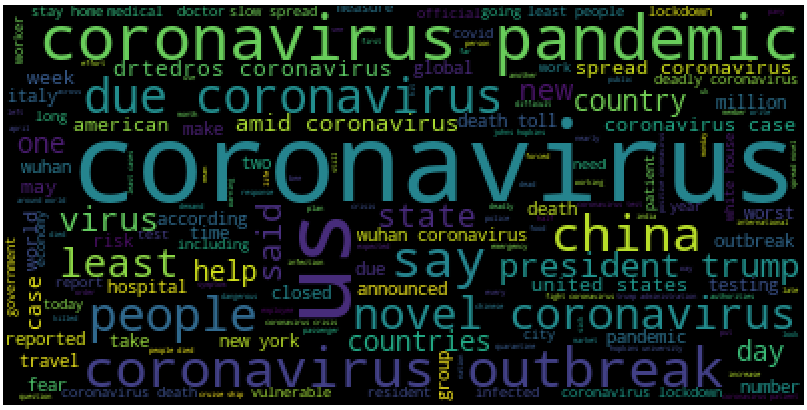
   Positve tweets percentage: 46.82%

2) What is the percentage of **negativity** rate in the public sentiments regarding coronavirus?

   Negative tweets percentage: 20.12%

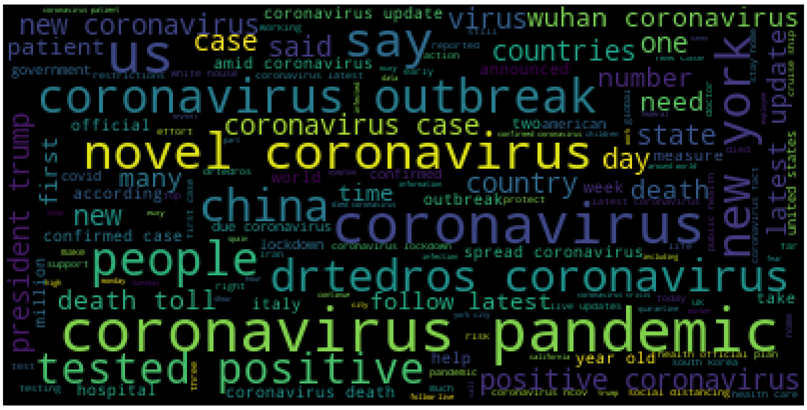
3) What is the most used **good** word describing what people think about the virus?

   The most occurred word in positive tweets is: [('coronavirus', 4695)]



4) What is the most used **bad** word describing what people think about the virus?

   The most occurred word in negative tweets is: [('coronavirus', 1999)]



5) What is the net count of **positive** tweets?

   Positive tweets net count: 464,882

6) What is the net count of **negative** tweets?

   Negative tweets net count: 199,863

7) Which month has the people engaged about the virus the **most**?

The month that the people has engaged about the virus the most is March.

8) Which month has the people engaged about the virus the **least**?

The month that the people has engaged about the virus the least is January.

9) What is the total amount of engagements that people had with this subject (like, comment, retweet)?
Likes: 8,148,965
Comments: 983,168
Retweets: 3,120,120
Total engagements: 12,252,253

10) What is the difference between the number of people that have tweeted in january 20 and April 2020?
Tweets and engagements in January: 567,102
Tweets and engagements in April: 4,788,658
Difference = 4,221,556

D) Dataset Features

Will update this section after we start implementing the project

Table count: *1*

Row count: *9,708*

Feature count: *34*

Is there any empty entries? Yes there is.

Dataset address link: ***The dataset has been generated by our code (we'll include the csv file with the code when we upload it to the system).***