Data-Intensive Computing Project Report
A Sentiment Analysis on Public Opinion of Covid-19

Group Name: DIC Omar+Weiran

1) What We Have Done

In September, we submitted a proposal named "A Sentiment Analysis on Public Opinion of Covid-19".

In October, we put our proposal into practice. We used a sentiment analysis algorithm which is used to classify each Tweet into very negative, negative, neutral, positive and very positive. It will then be streamed to Spark Streaming where it will be processed and stored in Cassandra. Specifically speaking, each tweet is grouped by country and by class and the final output determines the overall attitude of a certain country's citizens towards this pandemic. Finally, we visualized the results. More details are in the third subsection "The Method".

It's worth mentioning that there are two tiny differences between the proposal we submitted last month and the real process. First, in our proposal, we decided to use HBase as the main storage of data obtained from Spark. But the second laboratory gave us a hint, i.e., perhaps the Spark-Cassandra Integration might be better. Therefore, we switched from HBase to Cassandra. The second difference is that we used to think about using Datawrapper to visualize the result. However, during the process and after searching for some information, we eventually used Tableau because it has a better support on the global mapping-data function, which is more suitable for our project.

2) The Dataset

We applied for a Twitter developer account for the access to Twitter API. We extracted the tweet every second with key words "covid-19", "COVID-19", "covid19" "pandemic" and "coronavirus" and limited the language to English.

For the storage of Dataset, we used Cassandra. The following picture is the part of the Cassandra after the execution of catching tweet for an entire evening (The total data can be referred from the folder dataset in the submission zip file).

```
cqlsh:sentiment_keyspace> select * from sentiment_count;

 country                          | count | sentiment
----------------------------------+-------+---------------
                           Ghana2 |     1 |        Neural
                        Botswana1 |     4 |      Negative
                           Spain1 |     4 |      Negative
                      Seychelles2 |     1 |        Neural
                   United States4 |     9 | Very Positive
 Republic of the Philippines1     |    14 |      Negative
 Republic of the Philippines2     |     1 |        Neural
                           Italy1 |     1 |      Negative
               Republic of Korea2 |     2 |        Neural
                         Ireland3 |     6 |      Positive
                       Sri Lanka0 |     1 | Very Negative
                          Mexico1 |     3 |      Negative
              Antigua and Barbuda2 |     1 |        Neural
 Republic of the Philippines3     |     3 |      Positive
                          Uganda1 |     2 |      Negative
                      دولة الكويت |     1 |      Negative
                         Україна2 |     1 |        Neural
         Kingdom of Saudi Arabia2 |     1 |        Neural
                       Indonesia3 |     2 |      Positive
        المملكة العربية السعودية2 |     2 |        Neural
                          Panama2 |     1 |        Neural
                         Bermuda1 |     1 |      Negative
                         Belgium2 |     2 |        Neural
                            台灣3 |     1 |      Positive
                           Italy2 |     1 |        Neural
                          Kosovo3 |     1 |      Positive
                           Malta3 |     1 |      Positive
                        Bulgaria2 |     1 |        Neural
                          Kuwait2 |     3 |        Neural
                        Paraguay2 |     1 |        Neural
                          Mexico2 |     3 |        Neural
                         Austria1 |     1 |      Negative
            United Arab Emirates3 |     1 |      Positive
              Trinidad and Tobago2 |     2 |        Neural
                         Sverige1 |     2 |      Negative
                        Portugal1 |     1 |      Negative
                         Jamaica1 |     4 |      Negative
                   United States0 |     7 | Very Negative
                          Canada2 |     2 |        Neural
                           Egypt2 |     1 |        Neural
                         Bahamas2 |     1 |        Neural
                        Malaysia1 |     6 |      Negative
```


Fig 1. Part of Cassandra Storage Result

3) The Method

We utilized Stanford NLP, a deep learning algorithm for sentiment analysis which is used to classify each Tweet into very negative, negative, neutral and positive and very positive, five categories. Then we applied for a Twitter developer account so that we could have access to the Twitter API in Spark streaming process where tweet data can be processed. After we extracted the data and stored them in Cassandra in the format of three columns (country, sentiment, count), where the country stands for the location origin of the extracted tweet; sentiment stands for the category of result after using the classification algorithm and count is the total number of tweets from the same country that share the same category. Finally, we used Tableau, an interactive data visualization software, to visualize the data and exported five results ranging from very negative to very positive. The following shows the tools which we mainly used during this project.

- Scala: Main development language and obtain data streaming of Tweets.
- Spark: Pre-processing data streaming and obtaining Tweets
- Stanford CoreNLP: Sentiment analysis and classification
- Cassandra: For main storage of data obtained from Spark
- Tableau: For visualization people's perceptions per country per category in a colour map

4) The Results

The following pictures show the result, where the brown one stands for very negative; red stands for negative; yellow stands for neural; blue stands for positive and the dark blue one is very positive. (Unfortunately, we extracted very few numbers of very positive and very negative tweets but it's common sensible). The number on the country is the count of this category tweet.

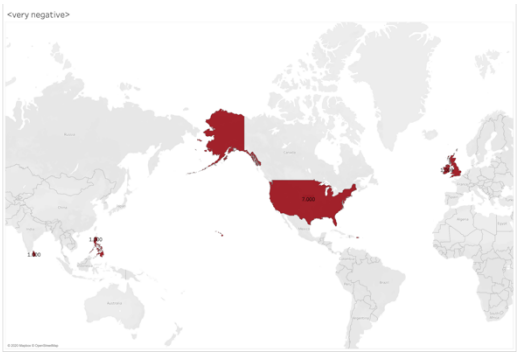

Fig 2. Very Negative

Fig 3. Negative

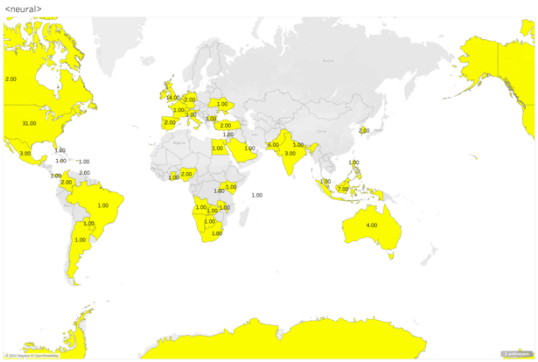

Fig 4. Neural

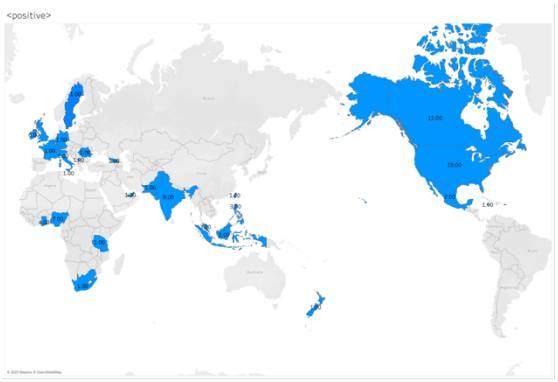

Fig 2. Very Positive

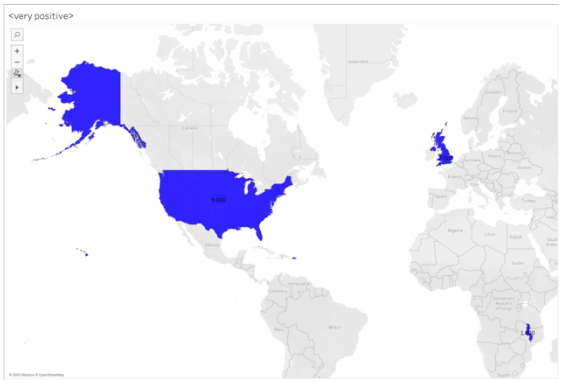

Fig 3. Positive

5) How to Run the Code:

First, in the terminal, we need to start Cassandra in the foreground with the command:

cassandra -f

Then, open a new tap and go to our project directory. Input the command and a tweet extraction process can be seen.

sbt run

Open another new tap and input command to enter Cassandra:

cqlsh

use sentiment_keyspace;

select * from sentiment_count;