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

Group name: DIC Omar + Weiran

September 27th, 2020

1 Problem Description

We're living in an era where data becomes the most valuable resource. At the beginning of 2020, an unprecedented pandemic named Covid-19 has hit the world. Nowadays, the trend of this pandemic is still widespread in some countries, which has brought significant financial and political effects as well as diverse public opinions all over the world. In our project, we seek to have a better understanding of the world's opinion towards the pandemic. Thus, we will obtain streaming data from tweets written in English that contain the global Covid-19; group those tweets by country and perform a sentiment analysis to try to understand people's perceptions per country and visualize it in a color map.

2 Tools

The tools utilized for this project will include:

- Scala/Java: Main development language and obtain data streaming of Tweets [1]
- Spark: Pre-processing data streaming and obtaining Tweets
- MLib [2]/Stanford CoreNLP [3]: Sentiment Analysis
- HBase: For main storage of data obtained from Spark
- Datawrapper [4]: For visualization people's perceptions per country in a color map

3 Data

First, we plan to use TwitterUtiles in Spark to stream data of latest tweets that contain the word Covid-19. TwitterUtiles contains multiple modifiers and types which create an input stream that returns tweets received from Twitter using Twitter4J's default OAuth authentication. Then we'll filter those Tweets with geolocation information to group by country.

4 Methodology and Algorithm

We propose a sentiment analysis algorithm which is used to classify each Tweet into negative, neutral and positive. It will then be streamed to Spark Streaming where it will be processed. Specifically speaking, each Tweet will be grouped by country and by class to show a graph where each country is colored depending on its output class. Finally, we would visualize the results.

References

[1] https://spark.apache.org/docs/1.0.0/api/java/org/apache/spark/streaming/twitter/TwitterUtils.html

[2] https://spark.apache.org/docs/1.0.0/mllib-naive-bayes.html

[3] https://nlp.stanford.edu/sentiment/code.html

[4] https://www.datawrapper.de/