**Brief description of dataset**

The data comprises of covid19 tweets, 42,000 in total, from Twitter and data from John Hopkins’s GitHub (<https://github.com/CSSEGISandData/COVID-19>).

**Twitter Data**

Tweets were scraped from Twitter using Twitter’s API. We scraped data between March 1 and March 14 (3,000 tweets per day, 42,000 observations/tweets total). The data that were scraped include: date, name, sources, isretweeted, text, favorites, and retweets. To better assess the effect of each tweet on retweet total, we selected a few keywords that were commonly used and created variables that tally their number of occurrences.

**John Hopkins’s Github**

Covid19 data were also cloned from John Hopkins’s GitHub repository. The two datasets that we utilized from the repository were the time\_series\_19\_covid-Confirmed and time\_series\_19\_covid-Deaths. These time series datasets include: Province/State, Country/Region, Lat, Long, and a time series between 1/22/2020 - 3/16/2020.

Since China, Italy and US are considered the epicenters of the crisis we grouped the two time series datasets by these countries. The grouped datasets were merged with our tweet data and 8 new variables were added.

In summary, our covid19 dataset (covid19march) has 42,000 observations/tweets and 36 variables. The data were sourced from covid19-related tweets from Twitter and John Hopkins’s GitHub repository. Majority of the variables are integers except for the following: date (POSIXct), name (character), sources (factor), isretweeted (logical), and text (character). Please see the appendix for in depth description of the dataset. (see Appendix for details explanation)