Data Buddies DataFest 2020 Submission

Markelle Kelly   mkelly23@calpoly.edu
Jenna Landy   jlandy@calpoly.edu
Luke Reckard   lreckard@calpoly.edu
Ashley Jacobson   apjacobs@calpoly.edu

We are a team of fourth year students at Cal Poly San Luis Obispo. Our team consists of statistics major Markelle Kelly, software engineering major Luke Reckard, statistics major Jenna Landy, and statistics major Ashley Jacobson. We built an interactive user interface using the Shiny R package, and looked at the frequency of user-specified terms over time from Google and Twitter. Our program utilizes the Google trends API and a repository of pre-collected tweets related to investigating how social media trends and Google searches have been affected by COVID-19.

Since tweets and Google searches are worldwide, we compare these trends to worldwide COVID-19 cases gathered by Johns Hopkins. We investigate how these international trends line up with specific dates related to the spread of COVID-19 in the U.S, starting at the introduction of COVID-19 to the U.S. on January 21, 2020. Other dates include the day a national emergency was announced in the U.S., and the first day when shelter in place initiatives began in the U.S. By default, the plots show the trends of the term "coronavirus" on both Twitter and Google against the number of worldwide cases. All values are scaled from 0-1 to be visible on the same scale. This means exact y-values between trends and cases, as well as between google and twitter, cannot be compared directly. However, we can compare general trends and the frequency of terms within twitter and google.

The number of cases, under "cases", exponentially increases from the start of January, while the term "coronavirus" peaks before the first shelter in place for Google trends and in February for tweets. As our first example, we added "covid" as a trend to our graph. The use of the term "covid" begins to surpass "coronavirus" in Twitter mentions after the first shelter in place is enacted for the U.S. However, Google trends for "covid" peaked and then began to decrease around this same time, never going above "coronavirus".

We then added the term "mask" to the plots. The majority of tweets about masks occurred immediately after the first case of COVID-19 in the U.S. It was not until the beginning of April that people really began to make Google searches about masks. Tweets related to "trump" peaked around the date of the first death in the U.S., which is also when he referred to the virus as a "democratic hoax", while Google searches spiked for searching about Trump when the U.S. declared a national emergency. For "unemployment", google searches start to increase after the first shelter in place order, when unemployment starts to become a problem in a multitude of countries. Finally, Wuhan is considered the epicenter of the pandemic but the term "wuhan" was only a popular topic on twitter until February.

Data Sources
- Emily Chen's repository of pre-collected tweets related to COVID19: https://github.com/echen102/COVID-19-TweetIDs
- COVID-19 Data Repository by the CSSE at Johns Hopkins University: https://github.com/CSSEGISandData/COVID-19 number of daily worldwide cases
- Google Trends: https://trends.google.com/trends/
    - Accessed through the gtrendsR package for R: https://www.rdocumentation.org/packages/gtrendsR/versions/1.4.4