Math 189Z Homework 2: Topic Modeling of COVID-19 Tweets Over Time (and PCA)

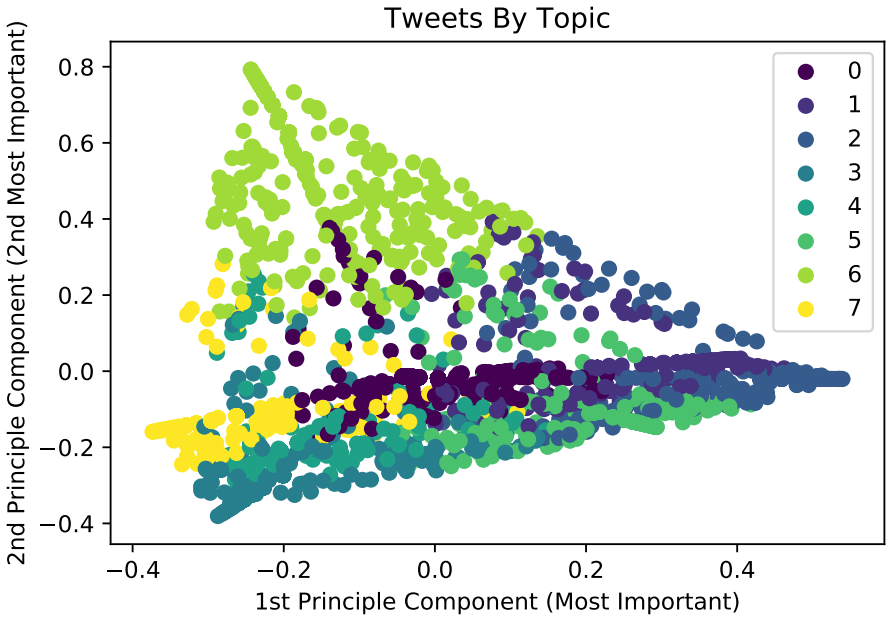
Daphne Poon

April 17 2020

1 List of Stop words

I added the following stop words: covid, coronavirus, covid-19, covid_19, virus, corona, sickness, viral, 19, https, http, com, org, www, outbreak, effect, sick, infected, infection, like, novel, new, news.

2 Tweet Topics (April)



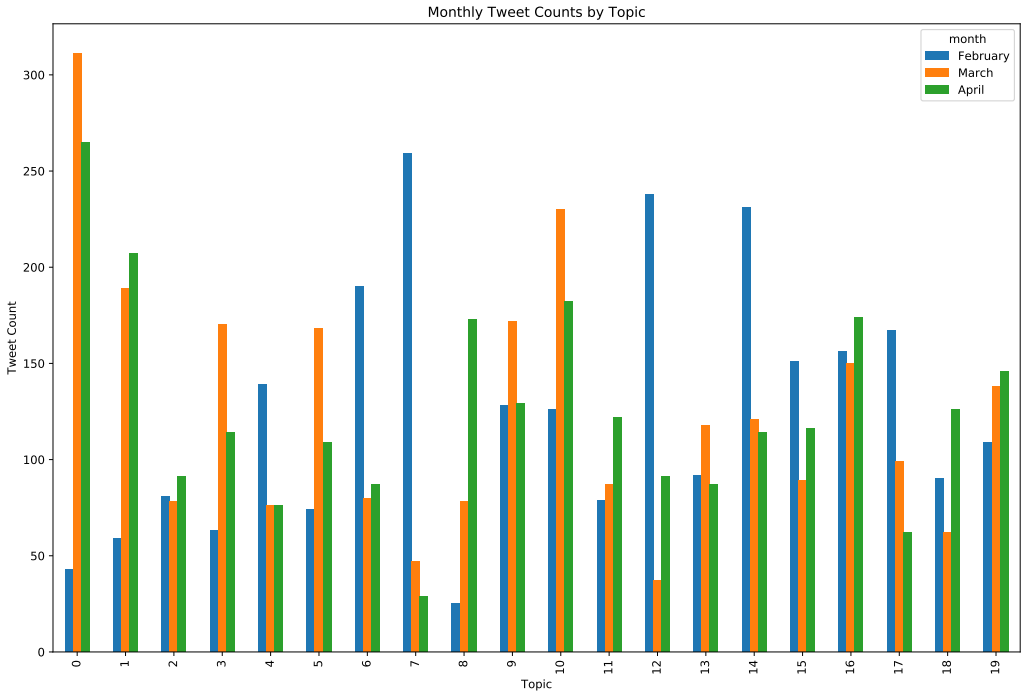
The graph above tells us that the tweets (from April, at least) are not very structured. There is a lot of overlap between colours (topics), and no topic is constrained to only a small section of the graph.

This means it may not be possible to separate topics in a clear way, and people may be

writing about several topics in one tweet. It might also mean that our dataset wasn't filtered well enough (maybe we needed to include more stop words).

There may also be fewer topics than we've set, where the topics would've been more dense around the bottom of the graph and less dense around the top.

3 Monthly Tweet Counts by Topic



From looking at the graph, I chose the four following topics to further analyze: 0, 1, 7, and 17.

- The words associated with topic 0 are: ['says', 'getting', 'positive', 'trump', 'negative', 'breaking', 'realdonaldtrump', 'tested', 'testing', 'tests', 'test', 'president']. Looking at that, it seems like the topic is **Testing for COVID-19**. That makes sense, considering how in March, when the COVID-19 situation in the US suddenly got a lot worse, many people were criticizing the government's response to the virus, and in particular how few tests were available. The situation has gotten better since then, so it would make sense that there are slightly fewer tweets about the topic (people don't tweet as much if they're not as angry).
- The words associated with topic 1 are: ['care', 'work', 'help', 'thank', 'health', 'staff', 'healthcare', 'emergency', 'workers', 'covid 19', 'lives', 'save']. From that, it seems like

the topic is **Thanking Healthcare Workers**. In the last month, a lot of attention has been drawn to those working on the frontlines combatting the virus (mainly essential workers and healthcare workers). It then makes sense that there have been an increased number of tweets thanking healthcare workers since mid-March.

- The words associated with topic 7 are: ['japan', 'quarantined', 'cruise', 'ship', 'positive', 'passengers', 'hong', 'kong', 'diamond', 'princess', 'reut', 'rs']. From that, it seems like the topic is **Diamond Princess**. The cruise ship was quarantined in Japan on the 4th of February, after an outbreak happened abroad. It was the centre of attention for a few weeks, but since March it has been discussed less, explaining the rapid drop in mentions since February.

- The words associated with topic 17 are: ['china', 'million', 'chinese', 'amp', '000', 'season', 'global', 'economy', 'impact', 'workers', '100', 'employees']. From that, it seems like the topic is **Impact on the Chinese Economy**. The Chinese economy's shutdown due to the coronavirus started at the end of January, and the economic effects (see: factory closures affecting global supply chains) were felt quickly across the world. While that is still very relevant, it's not as fresh of a topic (as opposed to, say, the US economy) so people are probably talking about it less, making it a declining topic relative to everything else.