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

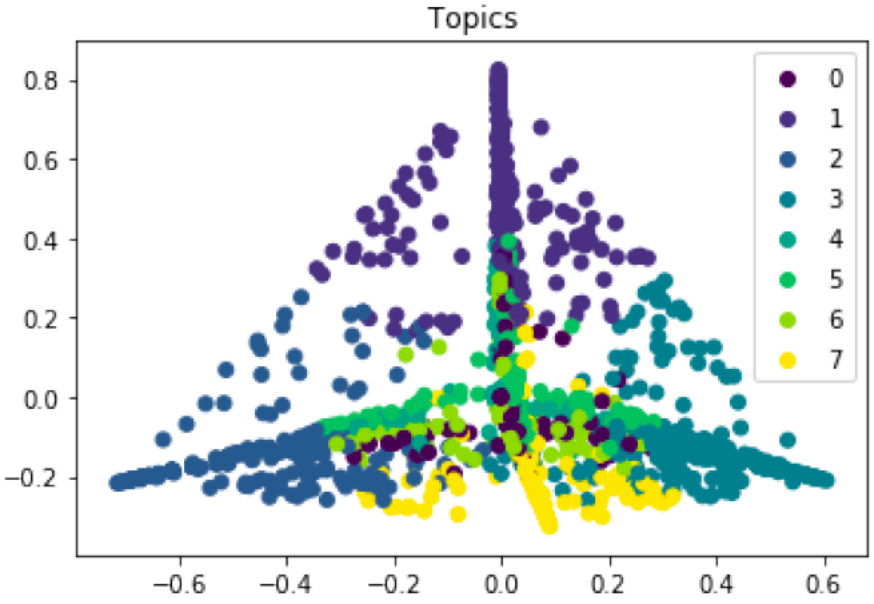
Rachael Soh

16 April 2020

Task 1: Your list of stop words

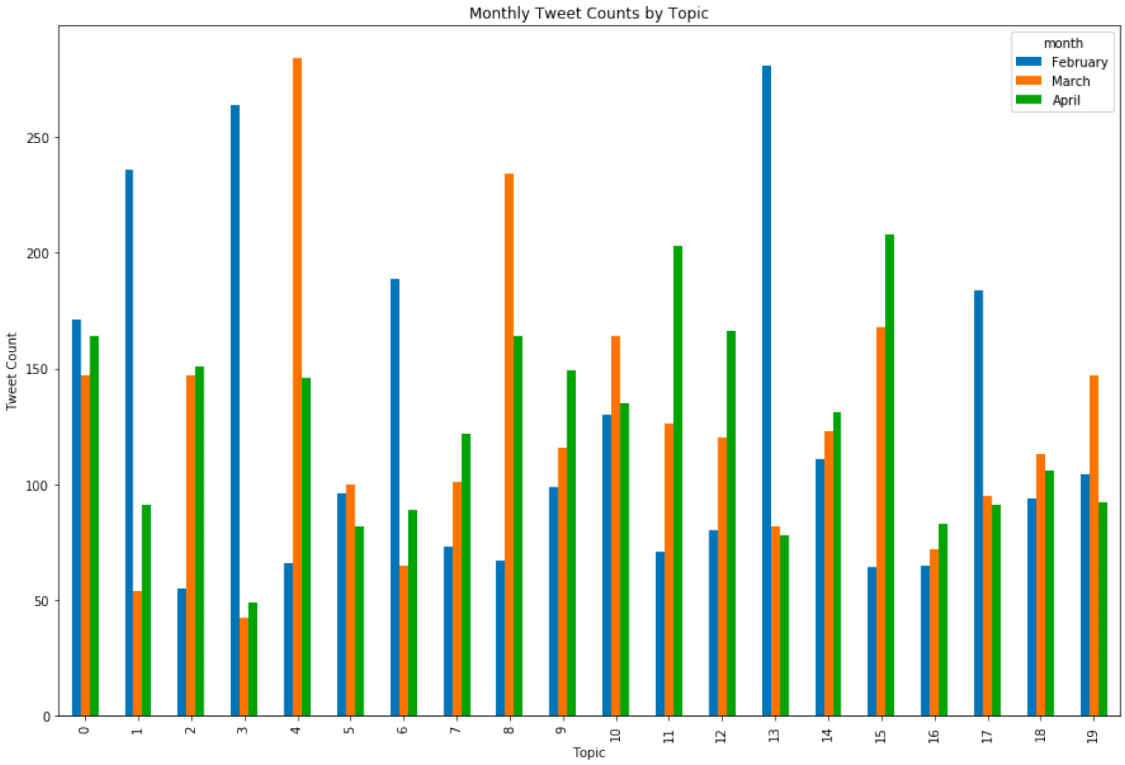
['corona', 'https', 'http', 'com', 'covid','19','covid_19','covid-19','covid－19','coronavirus','virus','covid19','www']

Tasks 2 and 3: An image of your PCA results on the tweet topics. In addition to the image, discuss what this graph shows us about the topics (are they more structured or unstructured and what does this mean?).



This graph tells us that our topics are more structured than unstructured. In the structured data visualization we are sampling from a distribution with a high likelihood of a (1.0,0,0,…,0) distribution, so it gives us this strangle-like shape.

Task 4: Image of 'Monthly Tweet Counts by Topic' bar graph



Task 5: Discussion of results (what you need to discuss is explained in Jupyter Notebook)

In February, we see the majority of the tweets were on topic 13: ['2020', 'fight', 'amp', 'uk', 'government', 'video', 'news', 'lockdown', 'pandemic', 'public', 'amid', 'queen']. These words seem to be related to'UK's response to the coronavirus outbreak' We know that this was when they first announced that the covid-19 was a pandemic and we see the numbers in UK rising as well as the term 'lockdown' used when cities had to close its borders. We see words like 'government, news, video, public' as we videos of hospitals in China emerged and governments asking the public to stay home. It makes sense that there was such a sharp decrease in tweets on the UK since the US became the new epicenter of the pandemic.

In March, we see that the majority of the tweets were on topic 4: ['china', 'japan', 'infected', 'new', '000', 'cases', 'total', 'death', 'cruise', 'ship', 'passengers', 'princess']. I would say this topic may have something to do with the cases on the Diamond Princess. I recall that sometime in March, information on the Diamond Princess cruise ship carrying infected passengers and them not being able to dock in Japan became pretty rampant. Meanwhile numbers in China becan increasing as well. These two places became some of the highest case counts at the time. We see very few tweets on them in February which makes sense as information on the diamond

princess had not emerged, while cases in China have decreased as well. In April, we see that the majority of the tweets were on topic 15:[‘people’, ‘flu’, ‘new’, ‘amp’, ’deaths’, ‘death’, ‘york’, ‘think’, ‘country’, ‘march’, ‘rate’, ‘times’]. I would say this topic is about cases in New York City. New York City became the new pandemic epicenter as death counts spiked. Trump also addresses this issue multiple times in press conferences. This makes sense for the numbers to increasing from February to March as the first case appeared in the first of march and only began to increase from there.

In April, we see that the majority of the tweets were on topic 15:[‘china’, ‘says’, ‘chinese’, ‘outbreak’, ‘global’, ‘apple’, ‘masks’, ‘supply’, ‘medical’, ‘amid’, ‘production’, ‘meet’]. This topic seems to be about “Medical supply shortage” as we see countries struggle to meet with the demand for medical supplies. We see many governments and doctors talk about this topic, asking for donations and governments to provide these medical supplies and masks to frontliners.