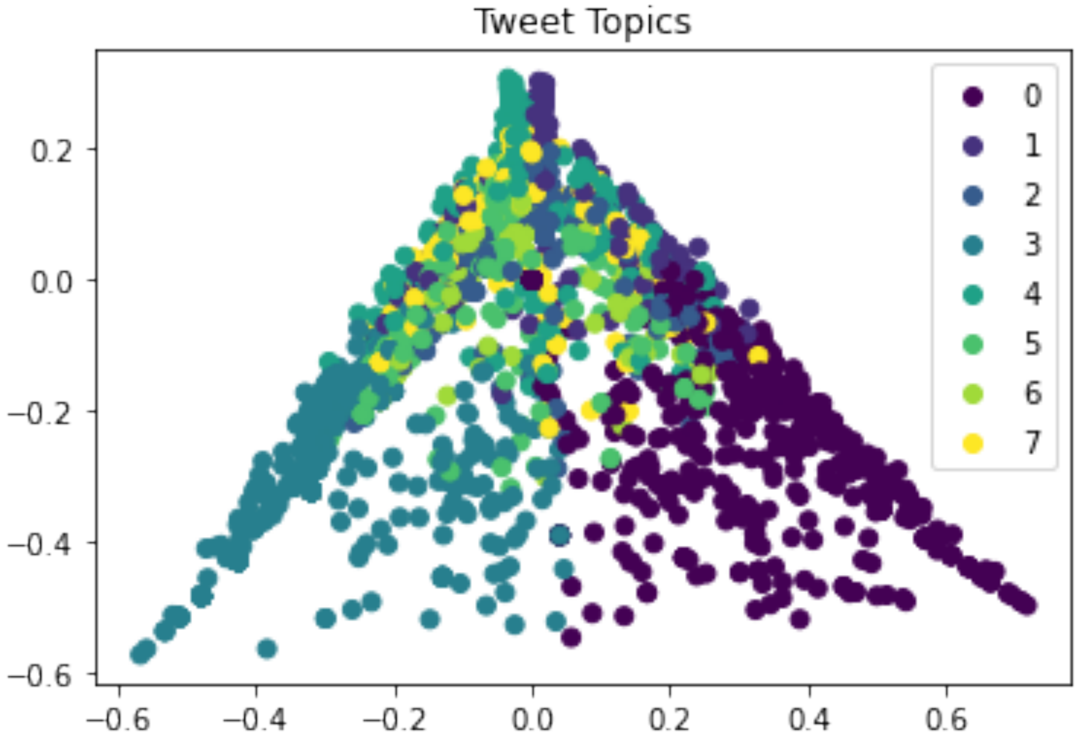
189Z Homework 2: Ethan Greenberg

Task 1: List of Stop Words

I made the following list of stop words iteratively after looking at some of the tweets and generating topic spaces:
'COVID-19','covid_19', 'let', 'want', 'today', 'like' ,'19', 'coronavirus', 'pandemic', 'virus', 'https', 'www', 'com', 'net', 'cases', 'breaking', 'news', 'story', 'infection', 'link', 'covid19', 'covid', 'man', 'covid—19', 'http', 'bit', 'ly'

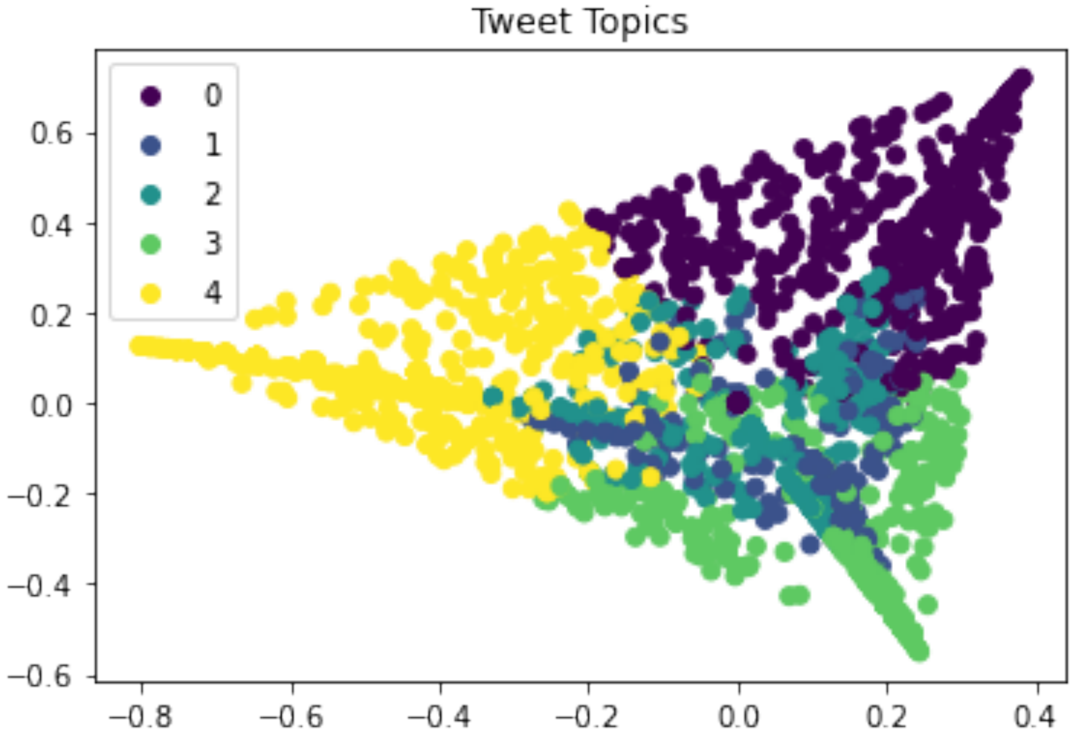
Tasks 2/3: Tweet Topic PCA

The following is an image of the PCA results from my tweet topics:



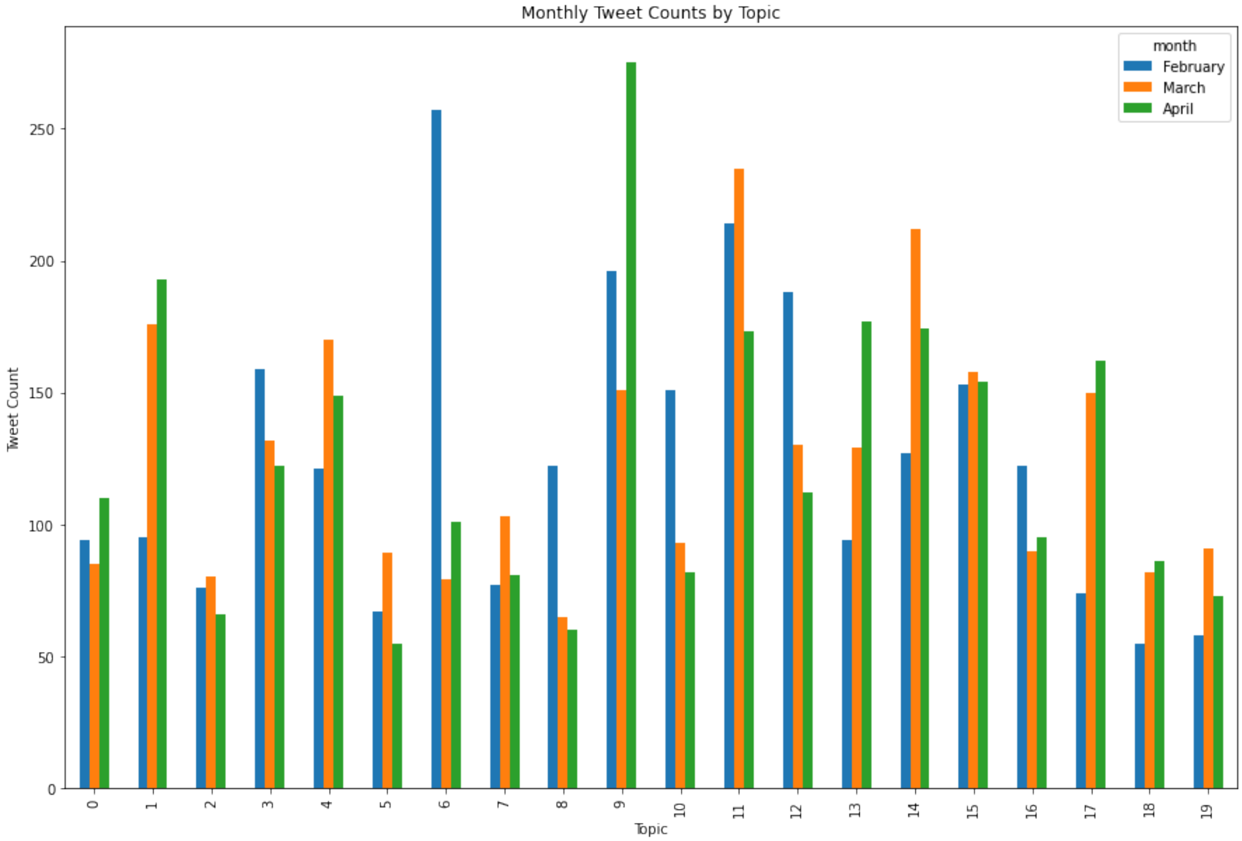
This plot suggests that there is some structure to the tweets, but they certainly are not an ideal, perfectly structured dataset. The edges of this plot look “cleaner” than the center space; and there is more overlap between topics. Perhaps adding more stop words would result in more structured topics.

I also LDA and generated a smaller topic space with 5 topics and used PCA to visualize the structure:



Tasks 4/5: Monthly Tweet Counts and Topics

Next, we ran LDA on tweets from February, March and April. We then generated a bar graph to show the number of tweets by topic per month that were sent.



I used this plot to explore the most popular topic from each month and also the topic that is most consistently tweeted about over all months.

February: Topic 6

The most tweeted-about topic from February was topic 6, with the following 12 most important words:

> 'china', 'amp', 'vaccine', 'world', 'outbreak', 'high', 'economy', 'impact', 'apple', 'supply', 'week', 'april'

Based on these 12 important words, topic 6 seems to be about coronavirus outbreak in China, and potential for a pandemic. This also makes sense with the timeline of the outbreak, as it was in early February that COVID-19 began spreading globally, but it was raging in China. The last handful of words are about the potential economic impact of the pandemic.

March: Topic 11

The most tweeted-about topic from March was topic 11, with the following 12 most important words:

> 'people', 'number', 'americans', 'confirmed', 'positive', 'disease', 'negative', 'tested', 'workers', '14', 'tests', 'test'

Based on these 12 important words, topic 11 seems to be about coronavirus in the US. This makes sense as US coronavirus cases really started picking up in March. The three mentions of "test" in this topic are also interesting. None of the other topics seems to include much about testing in the descriptor word set, and there were issues with tests in the US that were talked about. Thus, it is unclear if "test" is an important word simply because infection rates rose, or because there was a significant story about the testing itself.

April: Topic 9

The most tweeted-about topic from April was topic 9, with the following 12 most important words:

> 'china', 'fight', 'chinese', 'live', 'lockdown', 'corona', 'spread', 'doctors', 'thanks', 'think', 'real', 'party'

Based on these 12 important words, topic 9 seems to be about how coronavirus was being fought in China and about the lockdown. In early April, the lockdown on the Chinese city Wuhan was lifted. Also, China's case count had very clearly plateaued by this point, and there was certainly lots of talk about how China handled coronavirus.

Overall: Topics 11 and 15

Topics 11 and 15 are the two most tweeted-about topics across all three months. I already discussed topic 11, so here are the 12 most important words defining topic 15:

> 'hospital', 'away', 'man', 'time', 'people', 'just', 'chinese', 'new', 'york', 'city', 'food', 'local'

Based on these 12 important words, topic 15 doesn't really seem significant. I cannot see any single coherent story from these words. What I don't know is how one should respond to meaningless topics like this when analyzing a text. Is this an indication that I need to expand my stop words list? Or change the number of topics to search for?