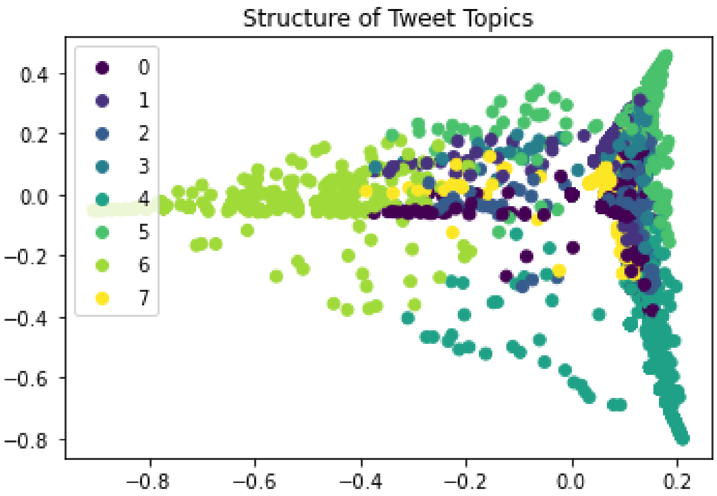
Homework 2

Task 1: List of Stop Words

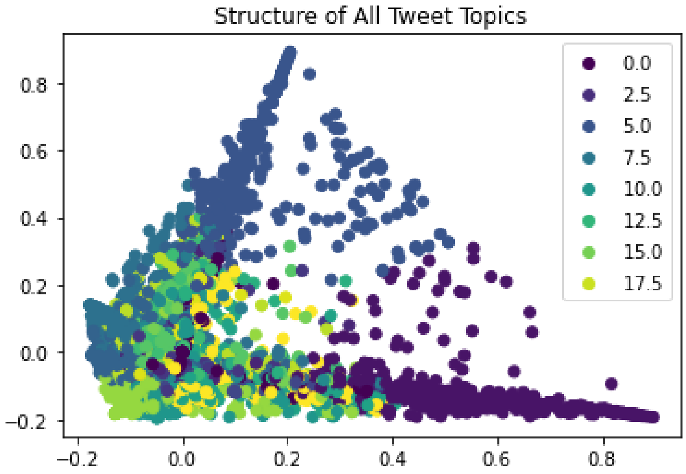
I added the following words to my stop list (21 additional words): ['com', 'https', 'www', 'http', 'll', 'just', 'fuck', 'say', 'says', 'said', 'really', 'like', 'liked', 'doesn', 'let', 'want', 'twitter', 'coronavirus', 'covid', '19', 'corona']

Task 2 and Task 3: PCA Results on Tweet Topics



Analysis:

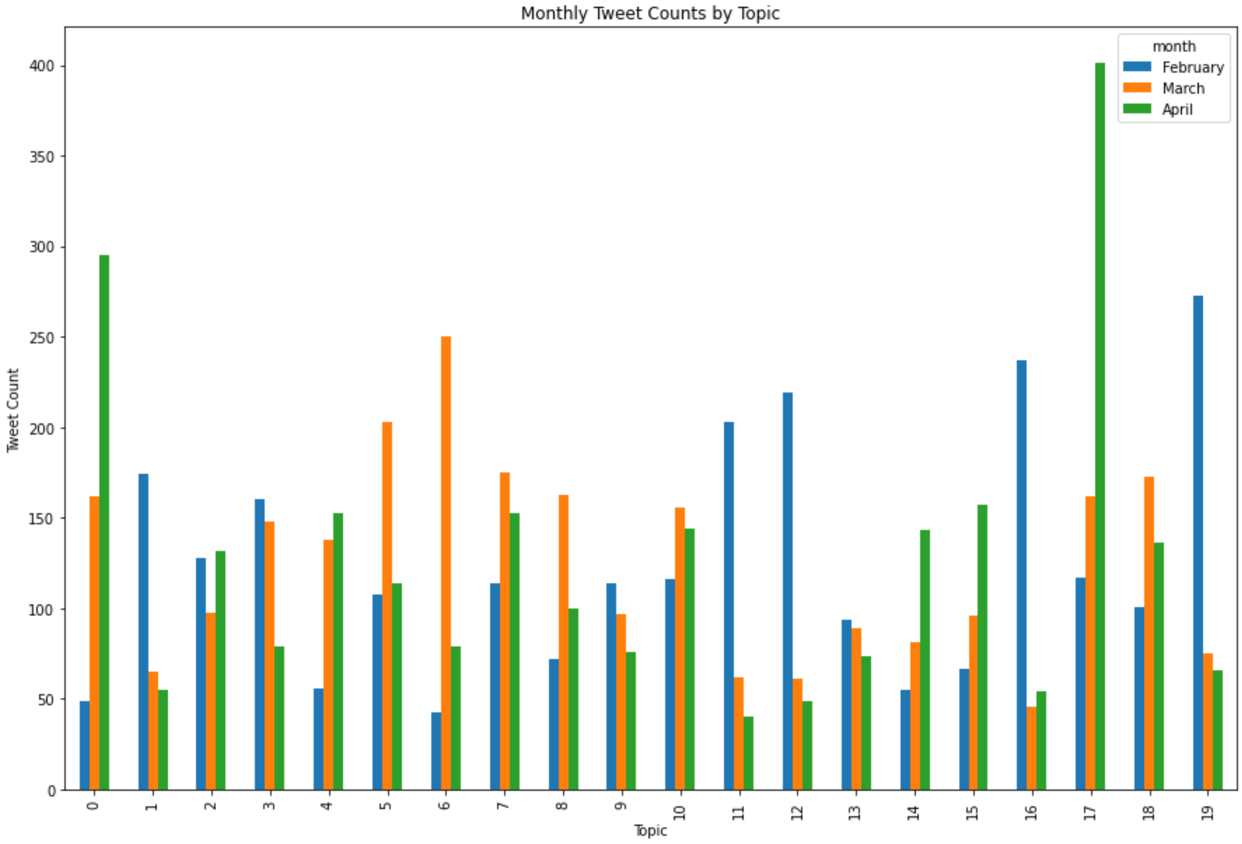
The data for just April seems to be somewhat structured as each of the topics seem to cluster around each other. There is some overlap between a few topics, but topics 6, 5, and 4 seem to be clustered around each other. A more structured PCA plot seems to suggest that topics have less overlap with their words. Therefore, from April we see that some topics overlap, but some are also independent. Ergo, this is somewhat structured.



Analysis:

The data for all tweets seems to have less structure than the April tweet space. There is a lot more overlap of topics in this PCA plot. Really only topic 0 and 5 seem to be independent. This makes sense, as you would expect a lot of overlap with words between topics for Covid-19 tweets, especially over a three-month period. For that reason, this graph is probably closer to being unstructured rather than structured.

Task 4: Monthly Tweet Counts by Topic



Task 5: Results

The topics with a majority from each month that I will be assigning meaning to are topic 19 (February), topic 6 (march), and topic 17 (April). The topic that seems to be present in all months that I will assign meaning to is topic 7.

Topic 19 (February): ['hospital', 'died', 'patients', 'infected', 'patient', 'death', 'hong', 'kong', 'got', 'symptoms', 'year', 'old']

Topic 6 (March): ['need', 'people', 'sick', 'work', 'stay', 'health', 'social', 'don', 'workers', 'home', 'lives', 'save']

Topic 17 (April): ['time', 'great', 'people', 'amp', 'help', 'day', 'response', 'cure', 'minister', 'prime', 'boris', 'johnson']

Topic 7 (All): ['fight', 'amp', 'help', 'outbreak', 'response', 'health', 'pandemic', 'trump', 'public', 'administration', 'medical', 'officials']

I will call topic 19 "Covid-19 Outbreak" because of how many words reflect the spread and travel of disease. Additionally, February was the month were the virus became more widespread and a problem around the world. It travelled to multiple countries and continents during this time. Furthermore, it became a much larger problem in the countries that had it before/during this month. Therefore, the semantic meaning to this topic makes sense – twitter should be talking a lot about hospitalizations, infections, symptoms, patients, deaths, etc.

I will call topic 6 "Reaction to Covid-19". This is because of the words need, people, and sick – they seem to be words one would use to describe requirements for treatment. Furthermore, workers and home pops up and this would be describing what many countries did as a reaction to Covid-19, which was lock down. This meaning makes sense, as many countries did lock down, and most of the world was still discussing how to go about combating the disease.

I will call topic 17 "Treatment of Covid-19". This is because of words like help, time, response, and cure. Additionally, we see politics enter the fold (prime, minister, boris, johnson), which would indicate countries are taking action against Covid-19. This is somewhat accurate, as companies are planning and producing vaccines and governments continue to debate what is the best way to move forward. However, it does value to encapsulate the economic problems the world is facing and how those are also receiving a lot of attention in the political world.

Lastly, I will call topic 7 "US reaction to Covid-19". This is because of how many words reflect treatment and planning against the virus. Furthermore, the mention of Trump probably specifies this to the US. This makes some sense to me, as the US has been criticized by many for their response to the pandemic for the past few months. Furthermore, their actions impact a lot of people with access to twitter, so it makes sense to me that their response would be a constant in the twitter world.