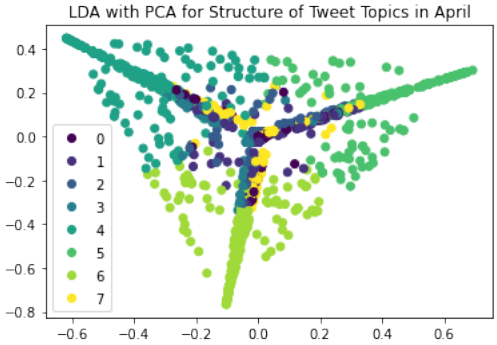
Task 1 – Stop Words

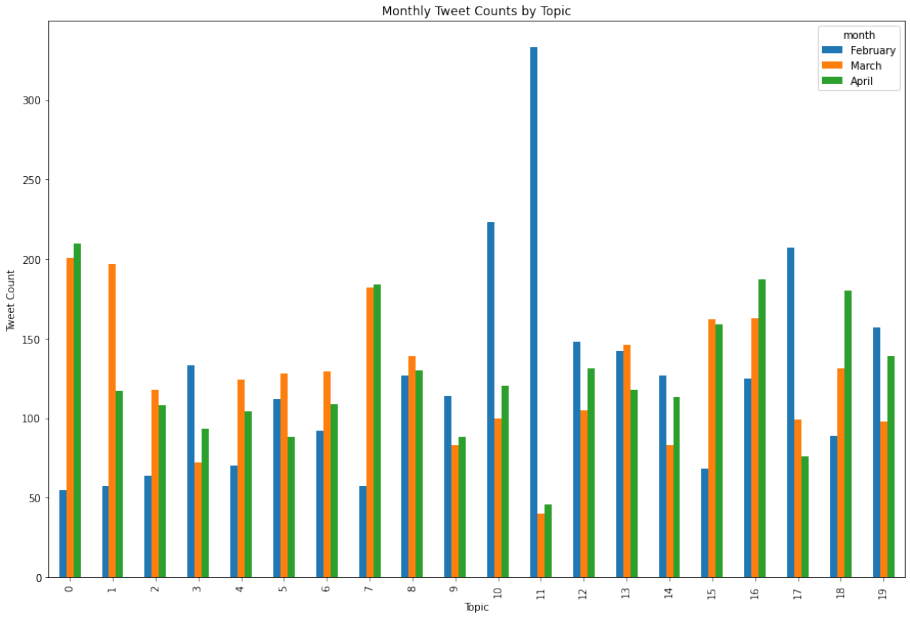
my_stop_words = ['covid', 'corona', 'coronavirus', 'covid19', '19', 'virus', 'com', 'https', 'www', 'pandemic', 'novel', 'covid_19', 'http', '2020', 'll', 've', '000', '04', 'covid-19', 'covid—19', '2019', 'covid2019', 'ly', 'bit', '5g', 'fuck', 'shit', 'pic', 'rs', 'ws', 'march', '03', 'html', 'reut', 'al', 'amp', 'twitter', 'crisis', 'gov','isn', 'don', 'doesn', 'april', 'cm', 'mar', 'just', 'like']

Task 2 and 3: Implementing PCA and Visualizing Results of LDA with PCA



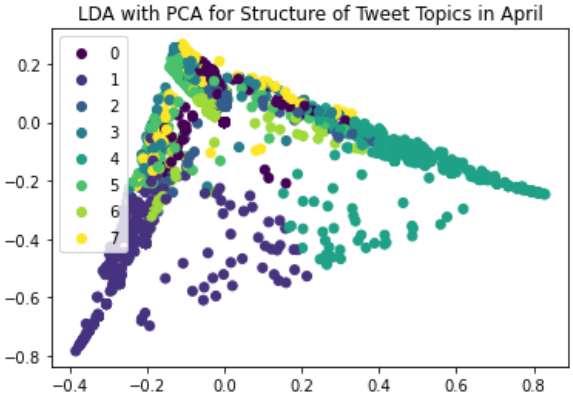
Note that structured data refers to clearly delineated data types with a clear pattern, while unstructured data has no particular organizational model. Tweets are a form of unstructured data. Since PCA is unsupervised learning, the layering of the scatterplot is not sufficient enough to make this claim strong. Nonetheless, it appears that this graph shows us that topics 4, 5, and 6 might do a good job spreading out the data, revealing some degree of a hidden structure in the data. Furthermore, comparing this plot to the simulated unstructured and structured plots rendered in Task 2, we can see that the tweet plot partially resembles the structured plot, meaning we can more confidently draw somewhat meaningful conclusions from the topic model.

Task 4 – Bar Graph

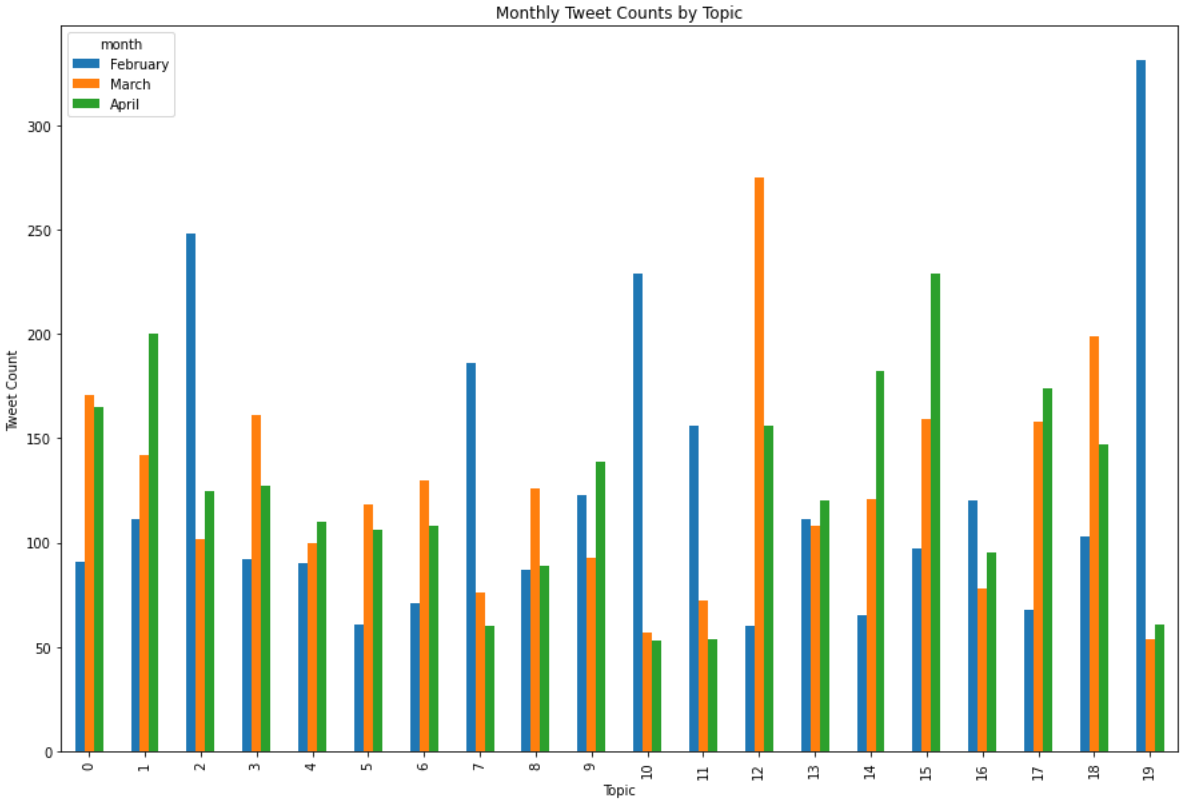


From this graph, we can see that the peak in February was Topic 10, which represents the vector [‘china’, ‘japan’, ‘outbreak’, ‘cruise’, ‘ship’, ‘quarantine’, ‘passengers’, ‘apple’, ‘latest’, ‘diamond’, ‘princess’, ‘home’]. We can “translate” this vector into a topic “Diamond Princess COVID-19 Outbreak”. The results on the bar graph makes sense, as the outbreak on the Diamond Princess had its first confirmed case on February 4th in Japan which garnered a lot of attention from the media. The ship sat in a Japanese port for approximately 2 weeks as the government there placed everyone aboard under quarantine. The peak of tweets for both March and February corresponded to Topic 0, the vector ['need', 'people', 'help', 'schools', 'closed', 'public', 'order', 'emergency', 'medical', 'support', 'relief', 'ik_saviourofnation']. This can be assigned the semantic meaning “Government Responses to COVID-19”. The results in the bar graph kind of make sense, as many local state governments began issuing stay-at-home orders, declaring public emergencies, and calling for PPE/medical supplies. For example, by March 17, 48 states had declared states of emergency in response to the coronavirus outbreak. However, I’d expect that as states have decreased various governmental orders and the curves are flattening, this topic would decrease in April, which is not evident in the graphs. Lastly, a topic consistently present in all 3 months is Topic 8. We can assign the semantic meaning “Chinese Response to COVID-19” for the corresponding vector ['china', 'good', 'people', 'sick', 'chinese', 'travel', 'lockdown', 'days', 'ago', 'doing', 'leave', 'weeks']. This makes sense, as China’s response to COVID-19 has been continually cited these past 3 months, but for different reasons. In February, China captured media attention for the immediate impact on its population (deaths, new cases, etc). In March and April, as China no longer leads the world in COVID-19 cases, the references to the nation center around their draconian measures and whether such measures are necessary in the other countries.

From an intuitional standpoint, I reviewed all the topic vectors and I wonder if the stops I used were too stringent. For one, I didn’t see any topics that could be easily summarized as “Testing” which is peculiar given the emphasis on testing in the media. Additionally, looking at the topic vectors, many of the vectors contained words that seemed randomly agglomerated. Therefore, I repeated the entire assignment with a new list of stop words: ['covid', 'corona', 'coronavirus', 'covid19', '19', 'virus', 'com', 'https', 'www', 'pandemic', 'covid_19', 'http', '2020', 'll', 've']. This gave me the following results:



By a similar line of reasoning detailed in Task 2-3, the PCA visualization seems to show that topics 1 and 3 might do a good job spreading out the data, and there is somewhat of a structure to the data. The corresponding bar graph was:



For February, the most popular topic vector was ['hospital', 'died', 'wuhan', 'china', 'chinese', 'outbreak', 'government', 'video', 'city', 'media', 'scientists', 'director'], which can be assigned the semantic meaning "Wuhan COVID-19 Outbreak". This makes sense, as the novel coronavirus hit China pretty hard in February. For March, the most popular topic vector was ['says', 'twitter', 'just', 'white', 'pic', 'trump', 'negative', 'house', 'tested', 'tests', 'test', 'president'], which could be assigned the semantic label "United States COVID-19 Testing". The popularity of this topic through March makes sense as Trump faced a lot of backlash for the lack of widespread testing in the United States. It is interesting that the tweets regarding testing decreased in April, as many news articles still seem to focus a lot of attention on testing shortages and flaws in the United States. For April, the most popular topic vector was ['patients', 'people', 'help', 'stay', 'possible', 'going', 'don', 'spread', 'doing', 'home', 'person', 'protect'], which can be assigned the semantic label "COVID-19 Social Measures" because all of the words directly or indirectly involve stay-at-home orders and other spread-mitigating measures. If we take this semantic meaning to the topic, the high frequency of tweets on this topic makes sense, as many social media sites are filled with social calls for folks to stay at home for the sake of their protection and fulfilling their civic duty.