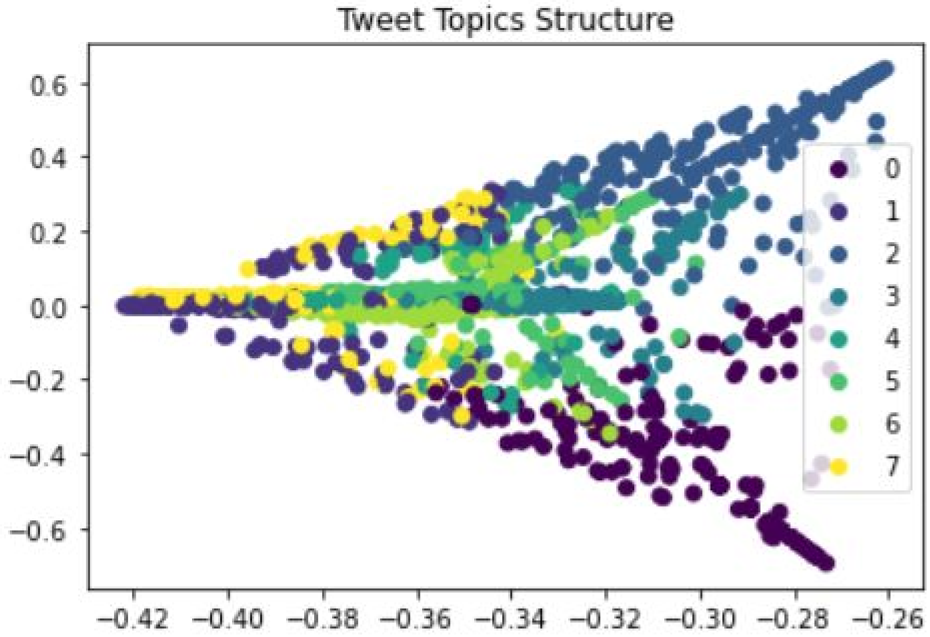
Shanni Lam
Math 189Z
HW2
16 April 2020

Task 1: List of Stop Words

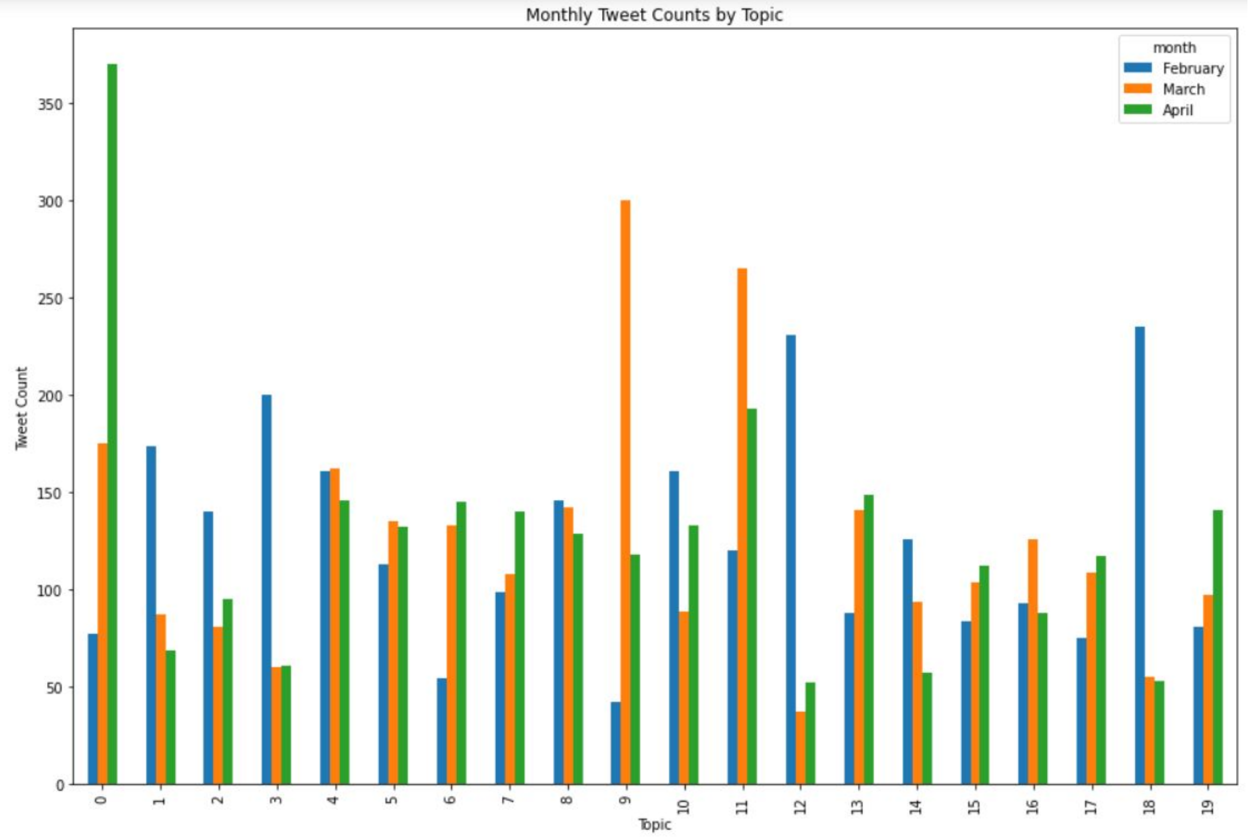
['https', 'com', 'corona', 'coronavirus', 'covid', '19', 'www', 'virus', 'covid19', 'covid_19', 'like', 'did', 'say', 'said', 'really', 'thank','just', 'says', 'twitter', 'time', 'breaking', '2020']

Tasks 2 and 3:



This graph shows aspects of structured and unstructured data (topics). There are some clumps of data at the points of this almost triangle, but there is also a lot of scatter, so the model cannot be too sure/assigns low probability to each topic because it is not sure which topic a group of words belong to. This is expected, as this is real-world data, and we can expect noise. There may still be irrelevant words that we should have put stop words on that can affect how topics come to be.

Task 4: Image of "Monthly Tweet Counts" by Topic Bar Graph



Task 5: Discussing Results

A topic with a clear majority in February is Topic 18. A topic with a clear majority in March is Topic 9. A topic with a clear majority in April is Topic 0.

4 topics that are present in all months are Topics 4, 5, 8, and 15.

Semantic meanings I assigned for each topic:
0: adjusting to life with COVID-19 via social distancing
4: COVID-19 starting in China and the hysteria causing toilet paper to be bought
5: UK government corruption not dealing with COVID-19 well
8: statistics on global confirmed cases, deaths
9: Donald Trump testing for COVID-19, his policies on testing people for COVID-19
15: biological COVID-19 news, where it spread, how to treat it
18: How COVID-19 affects China and Hong Kong

The bar graph for Topic 0 makes sense. We know that many public places and schools were still open in February but began to close in March and April once the disease spread more and the WHO declared a pandemic. The closure of facilities was done to promote social distancing.

The bar graph for Topic 4 makes sense. This is more of a general news topic, that China was the origin of the virus. People have been mass buying toilet paper for a while.

The bar graph for Topic 5 makes sense. If there is government corruption, people will likely continue talking about it. Boris Johnson got COVID-19 in March and is being treated hence why the March and April bars are slightly higher. People have generally been upset by the UK's slow response.

The bar graph for Topic 8 makes some sense. These are regular news reports of statistics, so the frequency of these tweets are likely similar but with higher numbers in each tweet. Oddly, there seems to be a decrease in the number of tweets. I would expect that as more people in more countries get sick or die, the number of reports would increase.

The bar graph for Topic 9 makes sense. It was a news item in mid-March that Donald Trump tested negative for coronavirus. News outlets will not keep reporting stuff that happened a month ago.

The bar graph for Topic 15 makes sense. As time passes, more scientific research is being conducted to treat the virus.

The bar graph for Topic 18 makes sense. There are more news topics about China and Hong Kong in February because then, the virus was still mostly contained there.