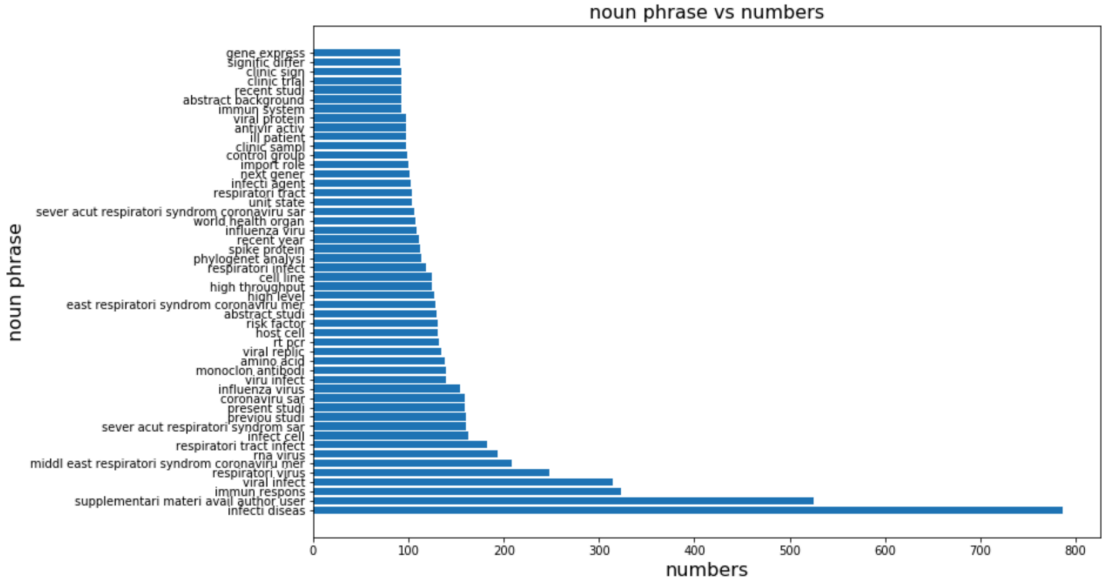
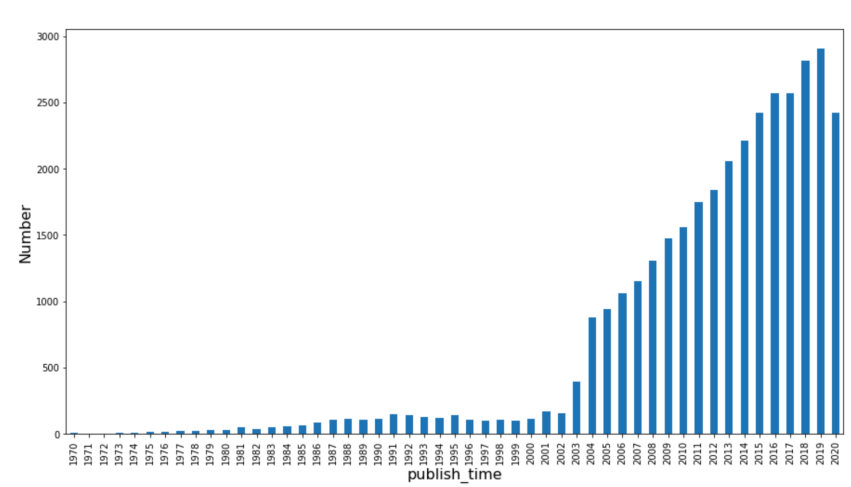
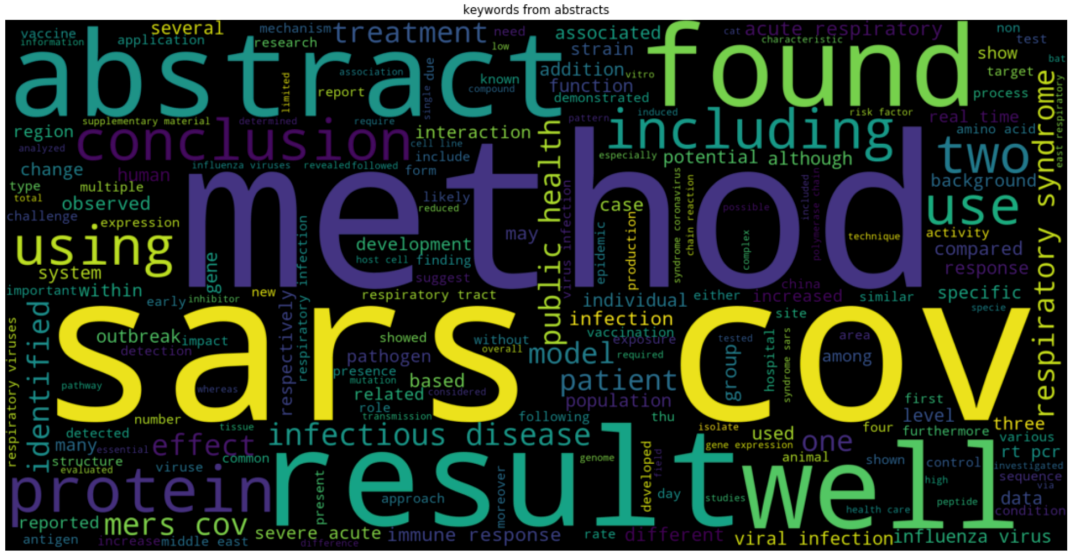
Exploratory data analysis





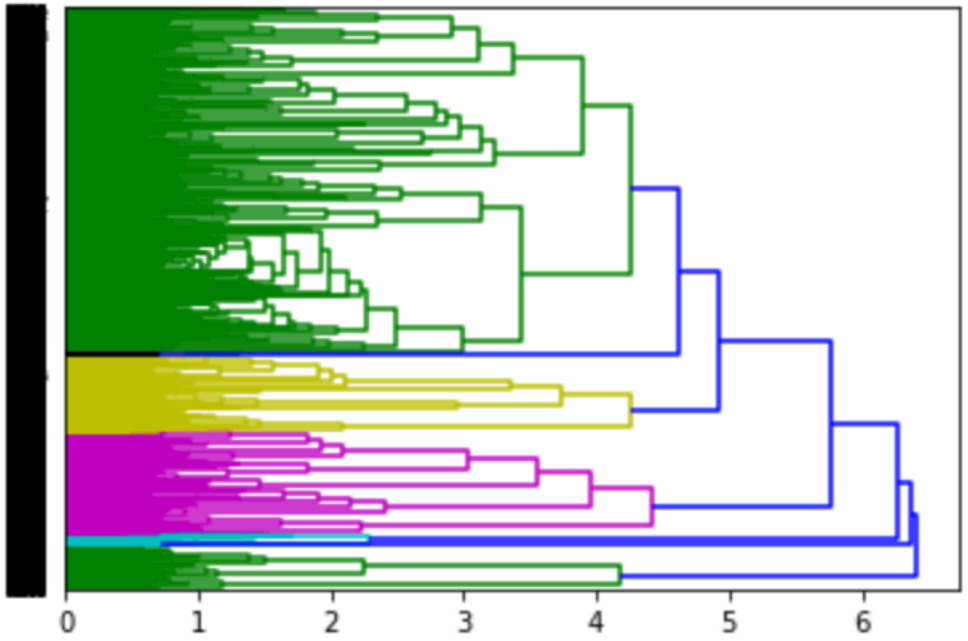
From the bar plot above, the literature increase rapidly after 2003. Because the SarS kills thousand of people. Therefore, to find more useful information, we can choose to use data after 2003. I like to use the most recent literature to look insight. The very old literature has very few value to refer.

I parse natural language biomedical literature data, and try to get adj+noun combination. From noun phase we can find more useful information from these articles. We can see which top is the most popular topic in these articles. First, the infectous disease; second, the immunity response to coronavirus. Then, we also can see the respiratory virus, rna virus, infect cell and previous studi. Therefore, these noun phrases can help us better to know what is included in these articles.

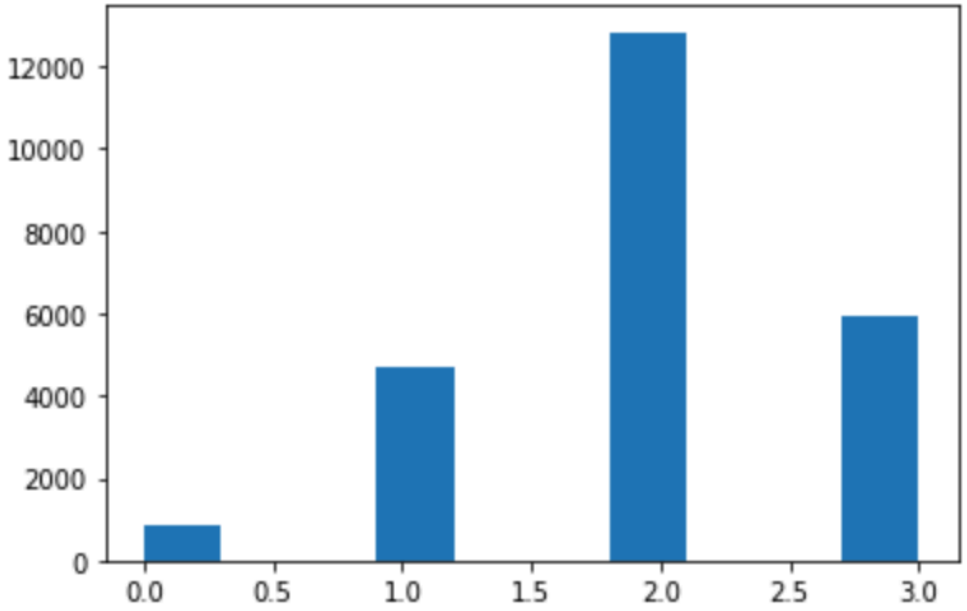


keywords from abstract make senses. we can find the keywords: method, treatment, infection, result. Besides, these article includes more talk about the SARS and MERS. The covid-19, MERS and SARS are same kind of virus. Therefore, in these articles, we can analyze the other coronavirus to get the some features of COVID-19, and also know something common with coronavirus.

Clustering the articles



Hierarchical clustering



K-Mean clustering

In this step, I use tfidf to get the top 500 words in the abstracts. Then, we can see how many clusters we have. Then based on the number of the clusters, we can separate the articles and generate the topics of each clusters.
Here, we need one more steps, I compared the tfidf and word frequency to see the different results of clustering. Within TDIFD, the clustering results looks more equally distributed. Within word frequency, the result doesn't make senses, because too many articles belong to one clusters. Therefore, here I would use tfidf, and the clustering results of tfidf
From the results, I believe these articles has 4 clusters from the graph of dendrogram. In other words, there are 4 topics of these articles.

From the k-mean clustering, when I use for loop to tune the best number of clusters. I find the 4 is the best choice. The results match the results of hierarchical clustering.

Results and interpret the findings of these models

3.4 Finding 1: Find the topic in each clusters of articles

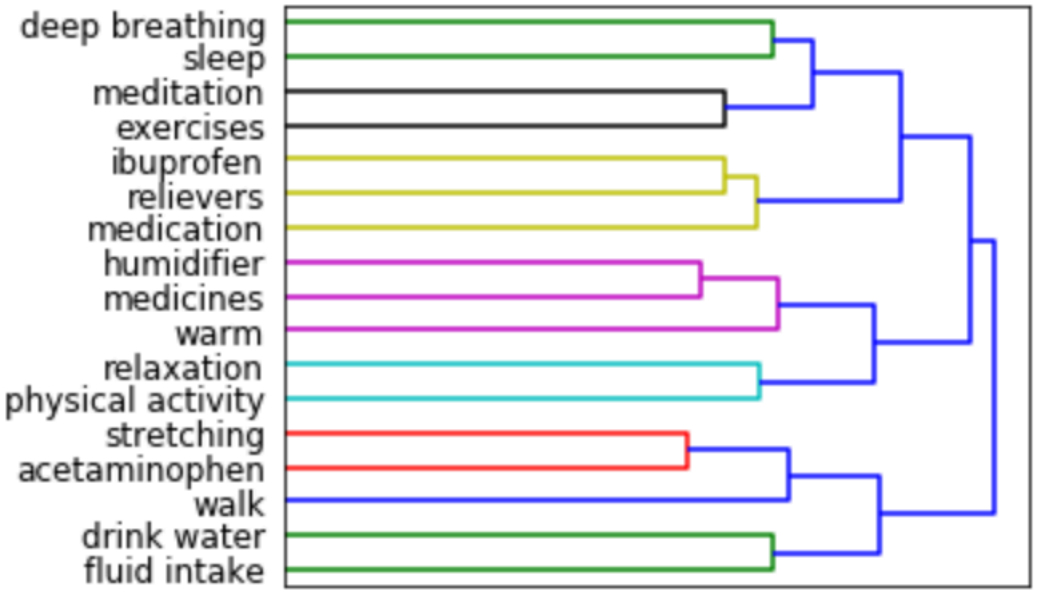
```
Topic 1:  cells virus infection protein viral cell expression proteins immune host
Topic 2:  virus viruses pcr samples viral detection respiratory study results using
Topic 3:  health disease data diseases influenza infectious public study control results
Topic 4:  patients respiratory infection cov sars study virus clinical children infections
```

Discuss and interpret the findings of these models

in the part 3, I use unsupervised clustering to cluster the articles, and then I can get the topics of each cluster. This can help doctor and government to sort these articles, and it is very convenient to look what includes in these articles, and where they should get their requiring information.
In summary, Topic 1 talks about how the virus infects host cell. Topic 2 talks about how people detect the virus and related to the respiratory study. Topic 3 talks about how to control these disease and protect public health. Topic 4 talks about study about coronavirus including sars and patients' study and children infections.

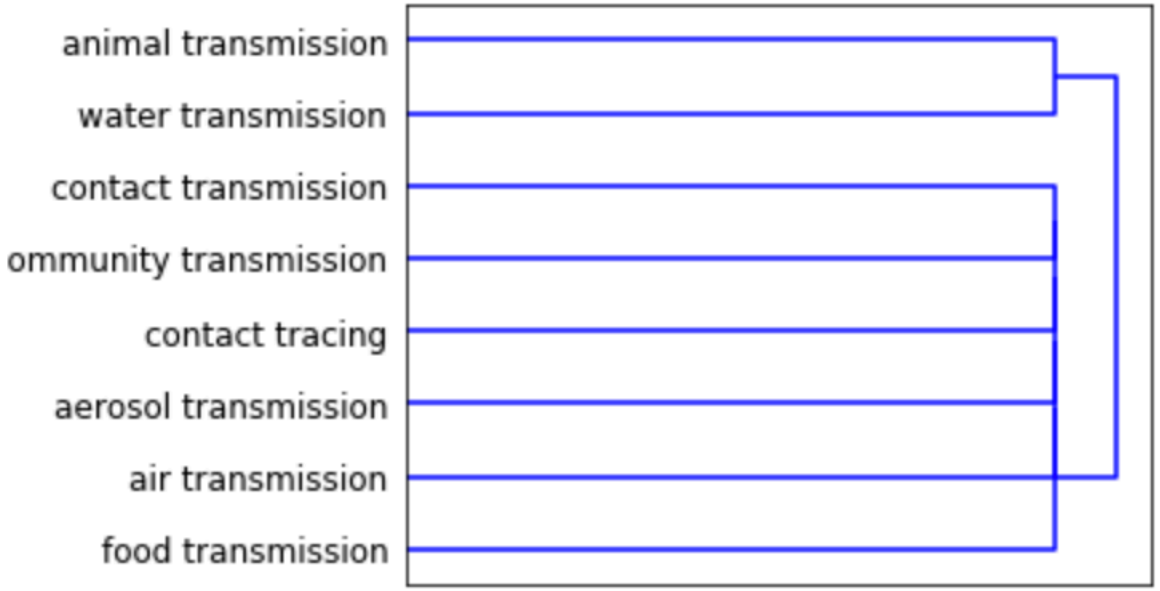
Results and interpret the findings of these models

3.5 Finding 2: Find the relation between each treatment



From the dendrogram, we can find the deep breathing and sleep is within a cluster. Ibuprofen, relievers and medication is within a cluster. Relaxation and physical activity is within a cluster. Drink water and fluid intake is within a cluster. Therefore, we can infer that there are four main treatment includes physical activity, drink water, take medication and sleep.¶

3.6 Finding 3: Find the transmission method



From the dendrogram, we can find there are two main clusters: non-air intermedium transmission and air transmission. Air transmission, community transmission, contact transmission, and food transmission has very closed relationship. It makes sense because people infect others through air. It also shows community transmission relies on the aerosol intermedium. Besides, animal and water can be another very important transmission method.

Deriving insights about policy and guidance to tackle the outbreak based on model findings

4.1 How scientists, doctors, nurses, healthcare professionals, industry and governments can best use the insights from your data science model to assist in the fight against the COVID-19 pandemic?

4.1.1 From 3.4 findinds, we know there are four topics obtained from 3.5 findinds. Topic 1 talks about how the virus infects host cell. Topic 2 talks about how people detect the virus, and related to the respiratory study. Topic 3 talks about how to control these disease, and protect public health. Topic 4 talks about study about coronavirus including sars and patients study and chlidren infections. healthcare professionals can use this classificaiton to find their required documents very quickly. Here, topic 3 is very useful for the government, because articles in topic 3 gives more research on how to control the spread, and how to protect public health

4.1.2 From 3.5 findings, we know there are about four treatment methods. We know currently there are not enough place in hospital to accept infectious people. Very light symptom patients need to self-treatment. Therefore, patients need to drink more water, keep enough sleep, keep warm, relaxation and do some entertainment when staying home is very helpful to recover. Therefore, I suggestion the healthcare professionals and the government can publish a guide for people who stay home and need self treatment, which also can release the public healthcare pressure.

4.1.3 From 3.6 findings, we know the main transmission way is through the air. Therefore, the community need to desease the contact of people. Government should encourage people stay home, and avoid the public gathering. When a place find a person infected by COVID-19, the whole place should be blocked to avoid virus spread. Besides, the water transmission is very dangerous. Government should try the best to keep the water supply is safe.

Proposed policies or action items

From 4.1, we can conclude some actions to avoid the virus spread. First of all, the government has take action to close the public places, which can avoid the air transmission between human. Second, government should encourage people wear masks. Because lots of people are asymptomatic carriers. Third, the shopping mall is encouraged to supply their goods by the shop assistant. Because people touch the goods in shopping mall may increase the cross infection.

Policy: 1.government should require people stay home and work from home. 2. Fine the people who appear in public park, and party with friends. 3. Force people to keep distance with others more than 2 meters.