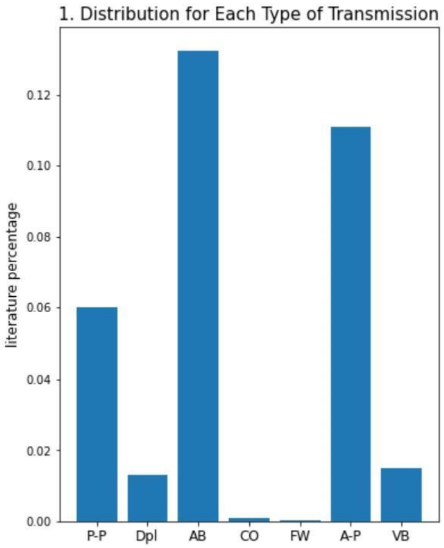
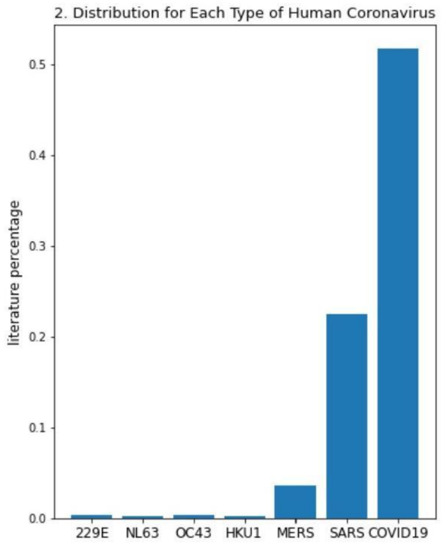
COVID 19 Research Literature Clustering: Insight on Preventing the Spread

Zhi Xin Zhu
#1002117112

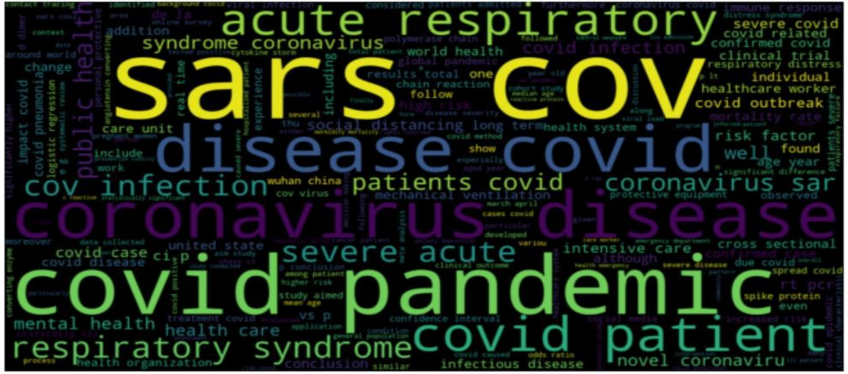
Exploratory Data Analysis

1. Distribution of Transmission Type
   a. Frequently mentioned transmission type is: airborne (13%), animal-to-person (11.5%), and person-to-person (6%).
   b. Covid19 is believed to be originated from bats (A-to-P) and is widely spread between humans through close contacts and airborne.
2. Distribution of Human Coronaviruses
   a. - The three most frequently mentioned human coronavirus type is: COVID19 (51.8%), SARS (22.4%), and MERS (3.6%). Common human coronaviruses including 229E, NL63, OC43, and HKU1 usually cause mild to moderate symptoms and respiratory illness, therefore, it is reasonable that majority of the researches in the dataset put the focus on more severe coronaviruses like MERS, SARS, and COVID19. Especially for COVID19 that is currently spreading around the world.
   b. The ranking of distribution of coronavirus type is in line with the rankings of the transmission. The primary transmission way of these three diseases matches the top three transmission type identified in the previous visualization.
3. Keyword of COVID19 Related Research Literatures
   a. The most frequent words from the word clouds are the alias of COVID19, such as covid, sars(cov-2), and coronavirus
   b. No obvious transmission related words in the wordcloud.Possible reason for that is currently the primary way of transmission are known for most of the public population. The research may have shifted to treatment and vaccination field (keyword: clinical trial).
   c. Some interesting word like 'public health', 'covid outbreak', 'healthcare worker', 'covid outbreak' can be seen in the word cloud. These keywords suggest that the virus has caused significant stress on the healthcare system in all countries. The frontline healthcare workers are facing pandemic burnout and serious work overload.


1. Distribution for Each Type of Transmission


2. Distribution for Each Type of Human Coronavirus

3. Keywords of COVID19 Related Reaseach Literatures



Model Implementation

- TF-IDF Feature Transformation

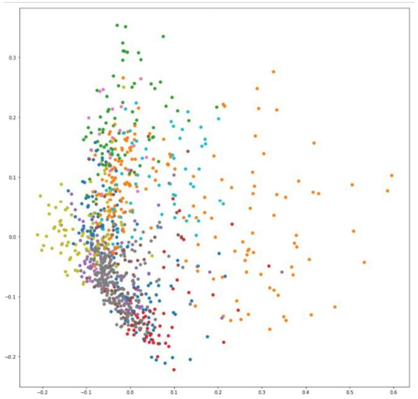
Model 1: Mini-batch K-Means Clustering

- ❑ K-Means clustering is a clustering algorithm that involves partitioning the dataset into a predefined number of clusters in an effort to minimize the variance within each cluster.
- ❑ Mini-batch K-Means is a modified version of k-means that makes updates to the cluster centroids using mini-batches of samples rather than the whole dataset. It is faster for large dataset and is more robust to statistical noise.
- ❑ Optimal Hyperparameters:
  - ❑ batch-size: 1000
  - ❑ n_clusters: 20
- ❑ Based on the result of the Elbow method, the impact of different batch-size is minor.

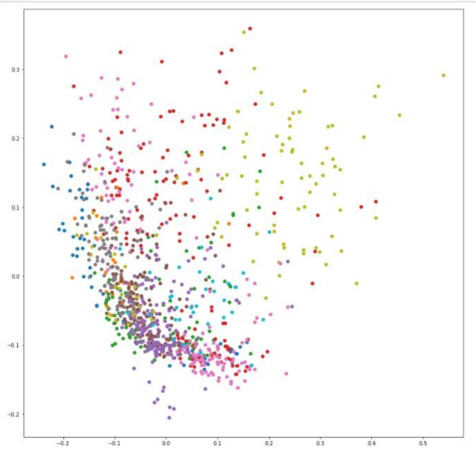
Model 2: Gaussian Mixture Clustering

- ❑ A Gaussian mixture model summarized a multivariate probability density function with a mixture of Gaussian probability. It is a flexible probabilistic method.
- ❑ Optimal Hyperparameters:
  - ❑ n_components (number of clusters): 20

PCA Cluster Prediction of Mini-batch K-Mean (Subset)



PCA Cluster Prediction of GMM(Subset)



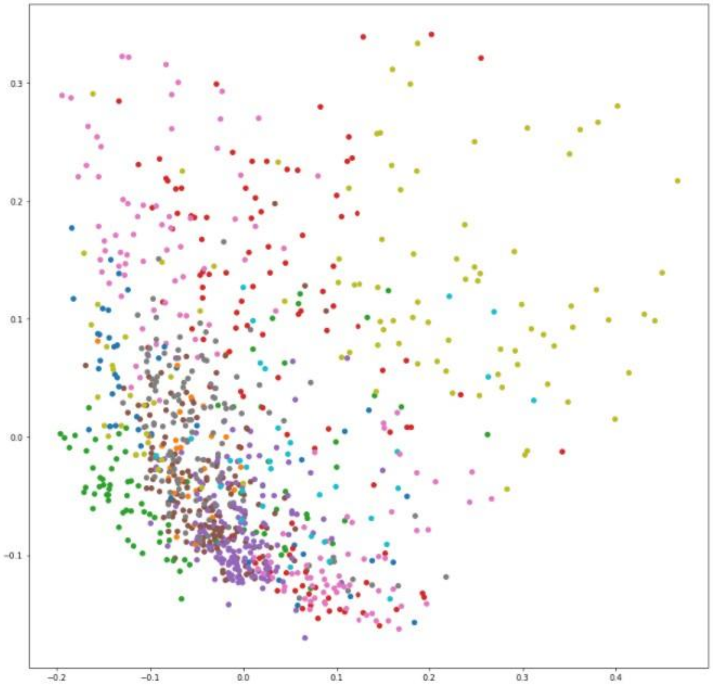
Proposed Model: K-Means Initialized GMM

- K-mean is used to initialized the clusters of GMM

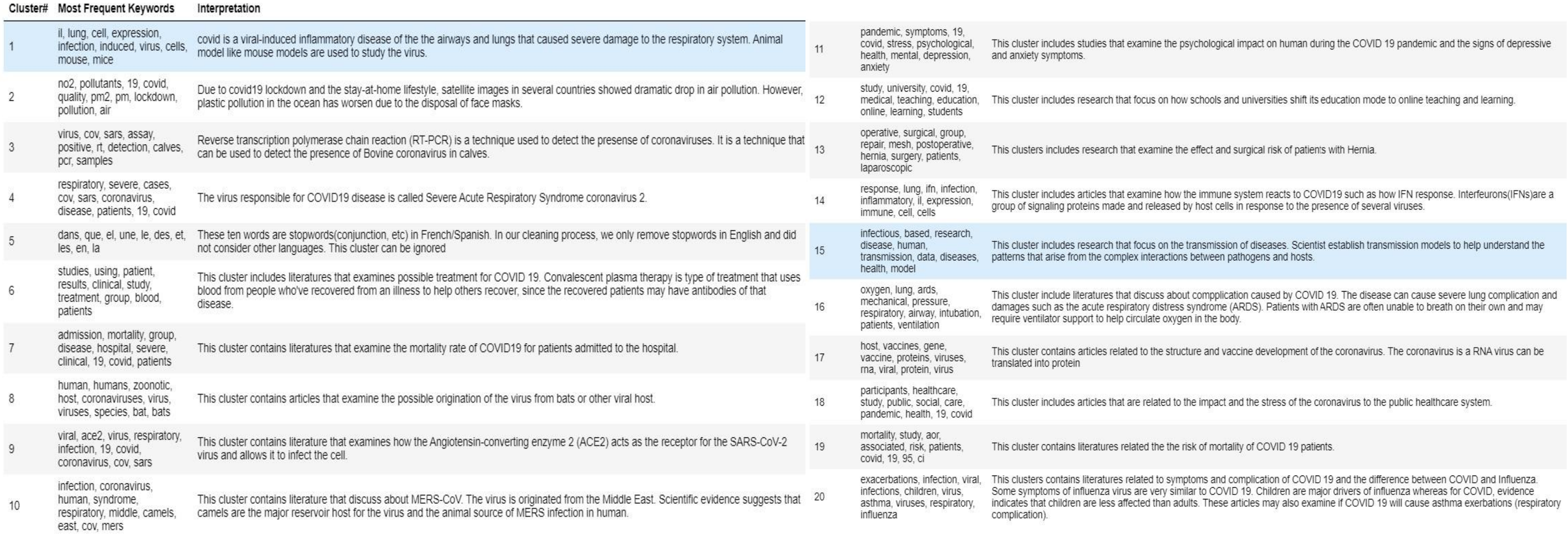
Model Selection Justification:

- K-Mean drawbacks:
  - Poor performance when the distribution of points is not in a circular form.
- Gaussian Mixture drawbacks
  - GMM tends to be slower than K-Means and it can also quickly converge to a local minimum that might not be a very optimal solution.
- The K-means initialized GMM balanced the drawbacks of both models, therefore it is selected.

PCA Cluster Prediction of K-Mean Initialized GMM (Subset)



Results



Cluster#	Most Frequent Keywords	Interpretation
1	il, lung, cell, expression, infection, induced, virus, cells, mouse, mice	covid is a viral-induced inflammatory disease of the the airways and lungs that caused severe damage to the respiratory system. Animal model like mouse models are used to study the virus.
2	no2, pollutants, 19, covid, quality, pm2, pm, lockdown, pollution, air	Due to covid19 lockdown and the stay-at-home lifestyle, satellite images in several countries showed dramatic drop in air pollution. However, plastic pollution in the ocean has worsen due to the disposal of face masks.
3	virus, cov, sars, assay, positive, rt, detection, calves, pcr, samples	Reverse transcription polymerase chain reaction (RT-PCR) is a technique used to detect the presense of coronaviruses. It is a technique that can be used to detect the presence of Bovine coronavirus in calves.
4	respiratory, severe, cases, cov, sars, coronavirus, disease, patients, 19, covid	The virus responsible for COVID19 disease is called Severe Acute Respiratory Syndrome coronavirus 2.
5	dans, que, el, une, le, des, et, les, en, la	These ten words are stopwords(conjunction, etc) in French/Spanish. In our cleaning process, we only remove stopwords in English and did not consider other languages. This cluster can be ignored
6	studies, using, patient, results, clinical, study, treatment, group, blood, patients	This cluster includes literatures that examines possible treatment for COVID 19. Convalescent plasma therapy is type of treatment that uses blood from people who've recovered from an illness to help others recover, since the recovered patients may have antibodies of that disease.
7	admission, mortality, group, disease, hospital, severe, clinical, 19, covid, patients	This cluster contains literatures that examine the mortality rate of COVID19 for patients admitted to the hospital.
8	human, humans, zoonotic, host, coronaviruses, virus, viruses, species, bat, bats	This cluster contains articles that examine the possible origination of the virus from bats or other viral host.
9	viral, ace2, virus, respiratory, infection, 19, covid, coronavirus, cov, sars	This cluster contains literature that examines how the Angiotensin-converting enzyme 2 (ACE2) acts as the receptor for the SARS-CoV-2 virus and allows it to infect the cell.
10	infection, coronavirus, human, syndrome, respiratory, middle, camels, east, cov, mers	This cluster contains literature that discuss about MERS-CoV. The virus is originated from the Middle East. Scientific evidence suggests that camels are the major reservoir host for the virus and the animal source of MERS infection in human.
11	pandemic, symptoms, 19, covid, stress, psychological, health, mental, depression, anxiety	This cluster includes studies that examine the psychological impact on human during the COVID 19 pandemic and the signs of depressive and anxiety symptoms.
12	study, university, covid, 19, medical, teaching, education, online, learning, students	This cluster includes research that focus on how schools and universities shift its education mode to online teaching and learning.
13	operative, surgical, group, repair, mesh, postoperative, hernia, surgery, patients, laparoscopic	This clusters includes research that examine the effect and surgical risk of patients with Hernia.
14	response, lung, ifn, infection, inflammatory, il, expression, immune, cell, cells	This cluster includes articles that examine how the immune system reacts to COVID19 such as how IFN response. Interfeurons(IFNs)are a group of signaling proteins made and released by host cells in response to the presence of several viruses.
15	infectious, based, research, disease, human, transmission, data, diseases, health, model	This cluster includes research that focus on the transmission of diseases. Scientist establish transmission models to help understand the patterns that arise from the complex interactions between pathogens and hosts.
16	oxygen, lung, ards, mechanical, pressure, respiratory, airway, intubation, patients, ventilation	This cluster include literatures that discuss about compplication caused by COVID 19. The disease can cause severe lung complication and damages such as the acute respiratory distress syndrome (ARDS). Patients with ARDS are often unable to breath on their own and may require ventilator support to help circulate oxygen in the body.
17	host, vaccines, gene, vaccine, proteins, viruses, rna, viral, protein, virus	This cluster contains articles related to the structure and vaccine development of the coronavirus. The coronavirus is a RNA virus can be translated into protein
18	participants, healthcare, study, public, social, care, pandemic, health, 19, covid	This cluster includes articles that are related to the impact and the stress of the coronavirus to the public healthcare system.
19	mortality, study, aor, associated, risk, patients, covid, 19, 95, ci	This cluster contains literatures related the the risk of mortality of COVID 19 patients.
20	exacerbations, infection, viral, infections, children, virus, asthma, viruses, respiratory, influenza	This clusters contains literatures related to symptoms and complication of COVID 19 and the difference between COVID and Influenza. Some symptoms of influenza virus are very similar to COVID 19. Children are major drivers of influenza whereas for COVID, evidence indicates that children are less affected than adults. These articles may also examine if COVID 19 will cause asthma exerbations (respiratory complication).

Insight on Current Policy and Regulation

Scientists and Researchers:

1. Research on COVID 19 complications on patients with pre-existing respiratory illness. Patients with pre-existing respiratory illness suffer more severe complication and symptoms
2. Post COVID 19 Conditions: Researchers should also put effort in studying possible permanent damages and sequelae of COVID 19.

Healthcare professionals and Workers:

1. Use all available personal protective equipment (PPE). Frontline healthcare workers and their families should be prioritize for vaccinations.

Governments:

1. Strict Regulations on Spread of COVID 19. Most people are still not protected by the vaccine, the government should put all its effort in preventing the public from gathering at this time.

General Public:

1. General public should continue to follow government regulations of stay-at-home order and protections to stop the spread of the virus. In Canada, evidence show that general public are slowly letting their guard down and the spread of COVID has bounced back to a significantly high level.

2. The results of our model (cluster 11) and research studies suggest that people are suffering mental health conditions due to the pandemic. People should find health and safe ways to cope with the stress.

School/Universities:

1. During the pandemic, schools and universities shift to remote teaching to prevent the spread of the virus(cluster 12). Schools should carefully evaluate the local epidemic conditions before shifting back to offline teaching.