Data Science for COVID19 Open Research

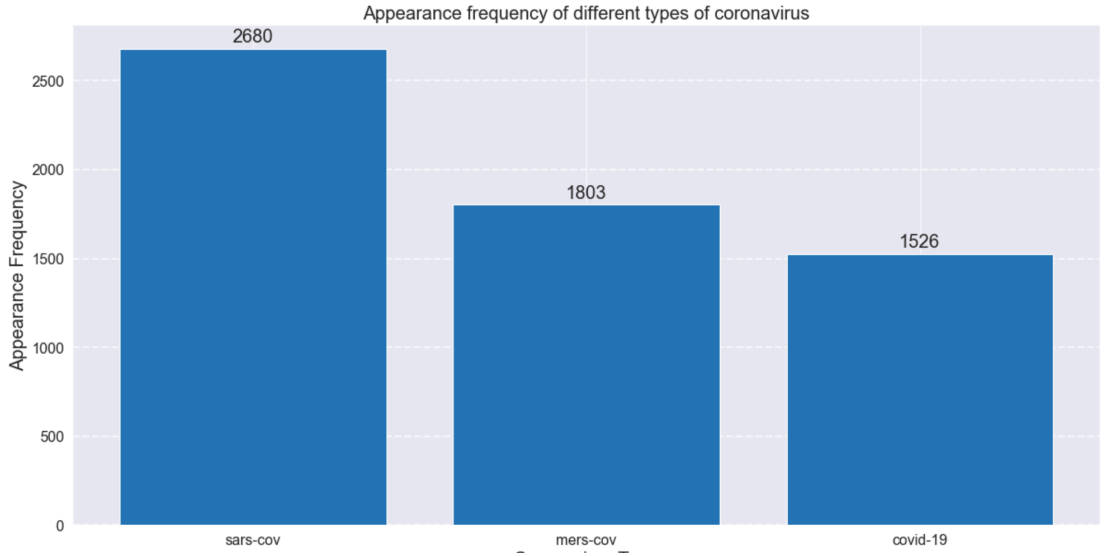
MIE 1624 Introduction to Data Science and Analytics

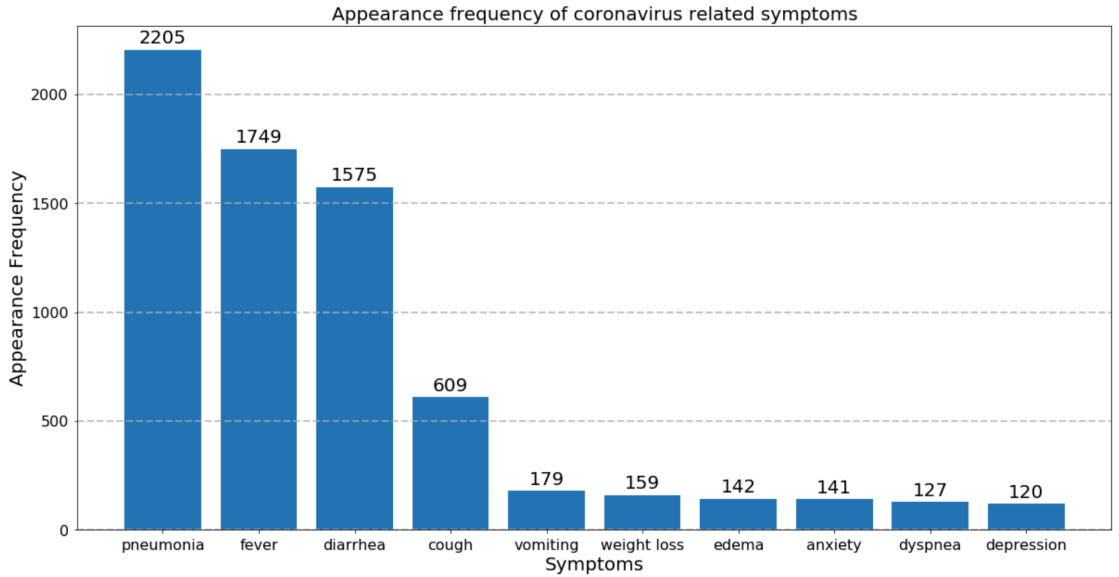
Name: Zichuan Wang
Student ID:1000474300

Exploratory Analysis

Out of the three outbreaks, the most frequently mentioned one is SARS with 2680 related papers, followed by MERS with 1803 papers, and lastly COVID19 with 1526 papers. The ranking make sense since COVID19 is really similar to SARS, therefore for papers which include covid19 is likely to include SARS as well. COVID19 just happened last year, that's why it has the last number of related research papers.

The most commonly mentioned symptom is "pneumonia" with 2205 related papers, followed by "fever" with 1749 related papers and "diarrhea" with 1575 related papers. Cough is mentioned in 609 related papers. All the rest of symptoms are mentioned under 200 related papers.

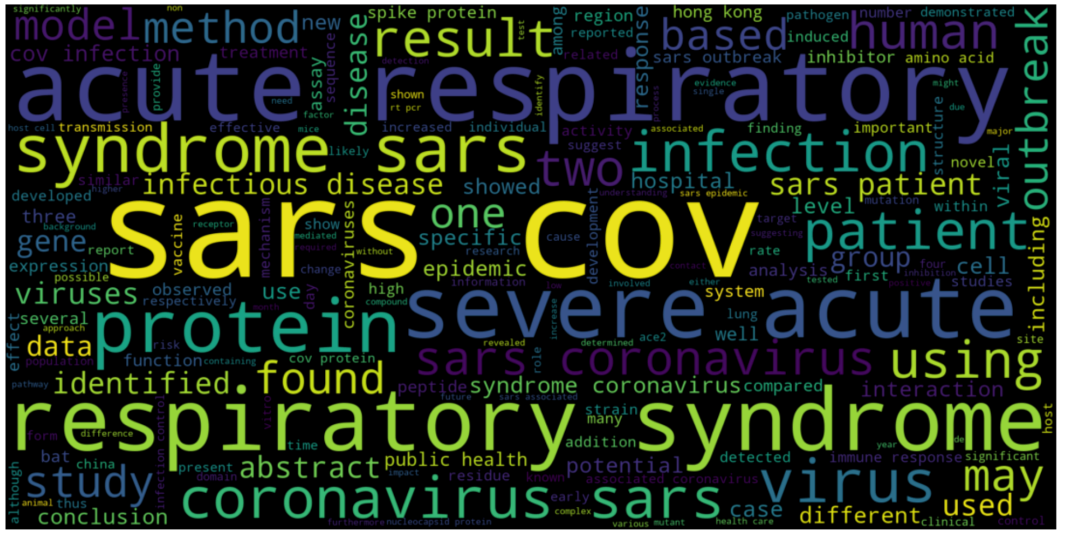




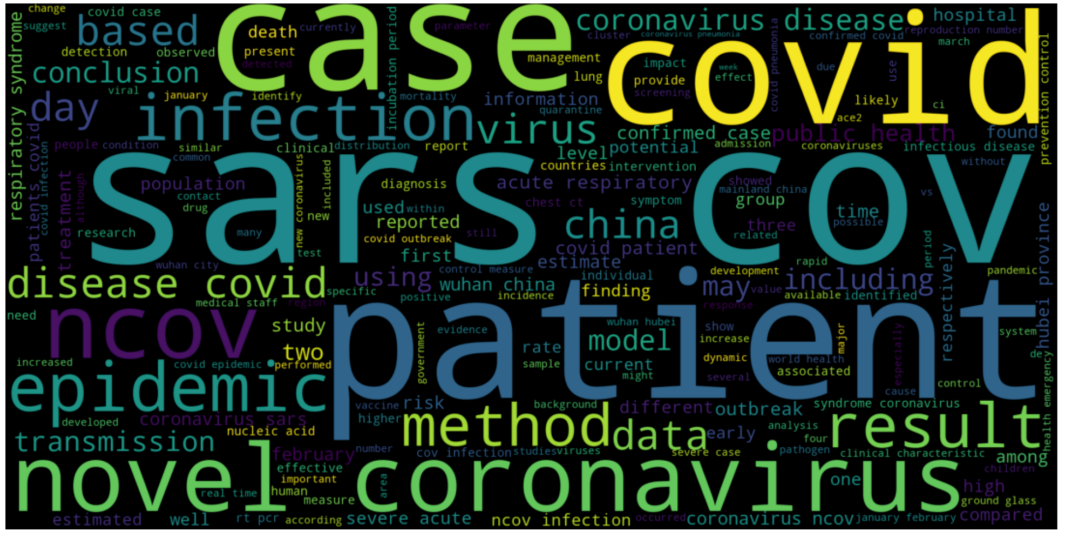
Exploratory Analysis

From the word cloud of SARS, we can see that related research papers have key words such as "acute respiratory" and "severe acute" meaning SARS is an acute illness, and "infection" meaning SARS is highly contagious.

From the word cloud of COVID19, we can see that related research papers have key words such as "SARS" meaning COVID19 is found to be closely related to SARS, and "novel" meaning COVID19 got discovered recently.



Word Cloud for SARS related research papers

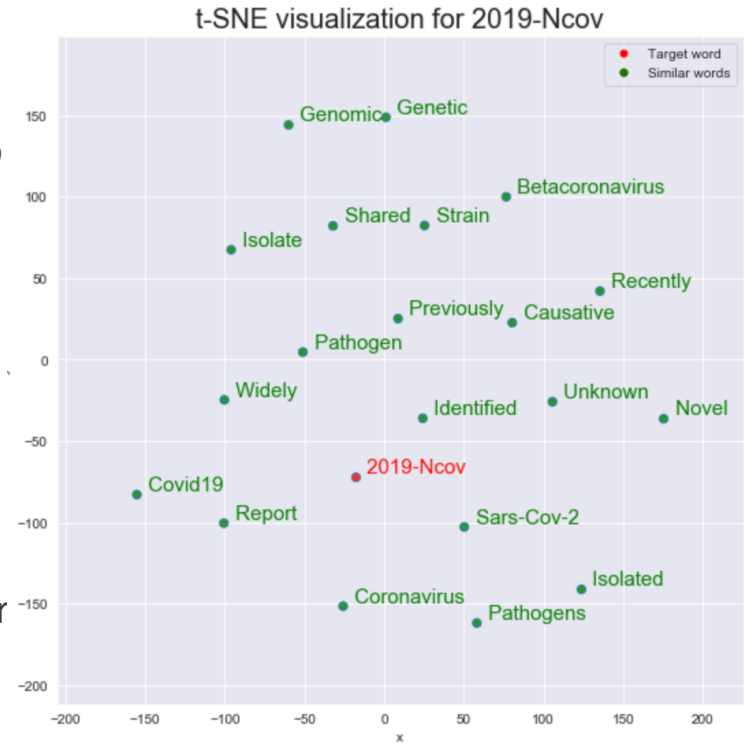


Word Cloud for COVID-19 related research papers

Model Feature Importance

TF-IDF:

The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today. That the reason why tf-idf is the main feature engineering technique used through out this assignment.



Word Embedding:

Word embedding represents words as vectors in multi-dimensional space. As an example, “2019-Ncov” is selected as the target word for COVID19 related research papers. Plot shows related words with reduced dimensionality. Since COVID19 is a novel coronavirus which is similar to SARS, it makes sense that it is surrounded by words like ”Novel” and “Sars-Cov-2”.

Model Results and Visualization

```
Model Training and Hyper-parameters Tuning for Logistic Regression

TRAIN PERFORMANCE:
The best score of model LogisticRegression through 5-cross validation is 0.9602948169281978, with the best hyper-para
meter {'C': 0.1}

Train accuracy of LogisticRegression model is 0.9959581550166429

Model LogisticRegression has test accuracy 0.956738768718802

Test result report:
```

	precision	recall	f1-score	support
0	0.98	0.96	0.97	524
1	0.94	0.97	0.95	774
2	0.96	0.93	0.95	505
accuracy			0.96	1803
macro avg	0.96	0.95	0.96	1803
weighted avg	0.96	0.96	0.96	1803

Model	train_acc	test_acc
Logistic Regression	99.6%	95.7%
Linaer SVC	100.0%	95.5%
Naive Bayes	90.9%	86.0%

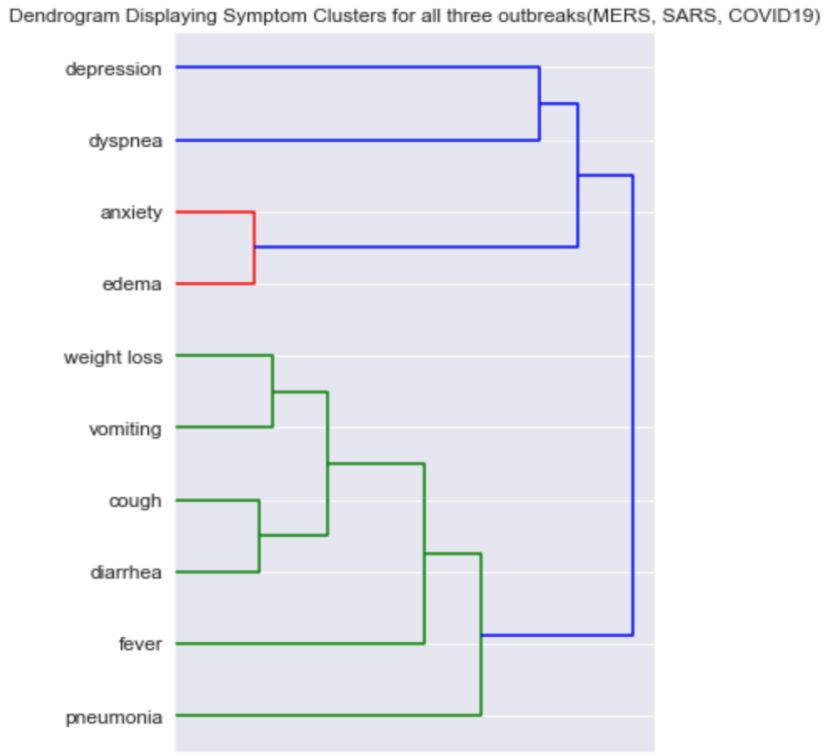
Supervised Learning:

3 different classifiers trained on labeled research papers are compared using grid search and cross validation. From the test accuracy, we can conclude that **Logistic Regression**, despite its simplicity, is the best classifier among all algorithms.

Unsupervised Learning (Hierarchical Clustering):

Symptoms are broken into 2 categories. First category includes edema, anxiety, dyspnea, depression, which is more closely related to the mental state of the patient.

For second category, the clusters are grouped according to the order of severity. Mild symptoms includes vomiting which often results in weight loss and cough which often comes with diarrhea. More serious illness can be developed such as fever or even pneumonia.



Policy and Guidance

1. Symptoms of COVID-19

According to our analysis, the most common symptoms of COVID-19 are pneumonia, fever, diarrhea, cough and tiredness. These symptoms are usually mild and begin gradually. Some people become infected but don’t develop any symptoms and don't feel unwell. But those with underlying medical problems like high blood pressure and bad habits like smoking are more likely to develop serious illness.

2. Self Protection and Spread Prevention

Some simple precautions can be taken right now reduce chances of being infected or spreading COVID-19:

1. Regularly and thoroughly clean your hands with an alcohol-based hand rub or wash them with soap and water.
2. Social distancing, maintain at least 2 meters distance between yourself and others.
3. Avoid touching eyes, nose and mouth without washing hands.
4. Avoid going to places with high density of people.
5. Wear face mask if have to perform outdoor activities.

3. After Infection

Self-care:
If you have mild symptoms, stay at home until you’ve recovered. You can relieve your symptoms if you:
1. rest and sleep to alleviate tiredness
2. drink plenty of liquids to alleviate cough and mild fever
3. use a room humidifier or take a hot shower to help ease a sore throat

Medical treatments:
If you develop a severe fever, cough, and have difficulty breathing, need to seek medical care immediately.