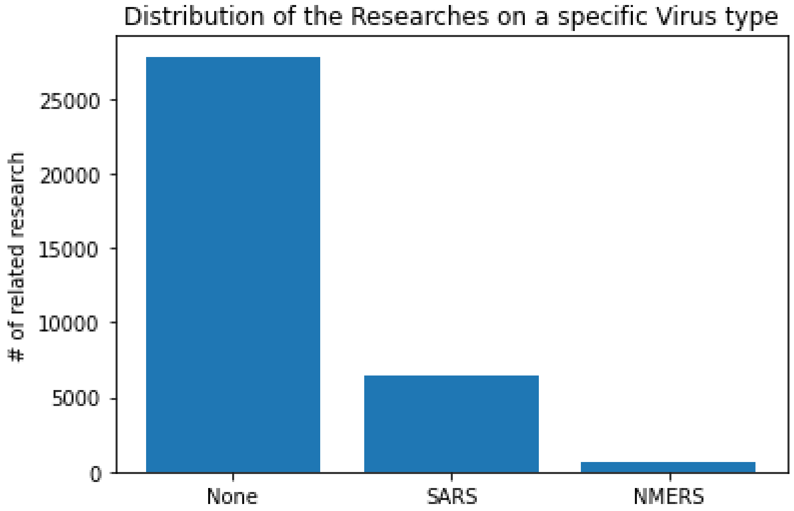
Recently, COVID-19 is spreading all over the world, and researchers are still working on finding an efficient way to solve the issues that bring by this disease. Although the treatment of COVID-19 is not yet discovered, the number of relevant research about coronavirus is large. Thus, it is important to find a way to extract useful information. It would be helpful if data analytic technology can help the experts in the medical and healthcare system to improve their work efficiency.

This project is mainly about how to find a more efficient way to sort the most relevant information from a huge dataset. First, we need to figure out what kind of information is more relevant to 'COVID-19'. To do this, we first need to do some research about the current types of coronavirus.

Based on the 349809 research papers, human coronavirus are mainly in the following types:

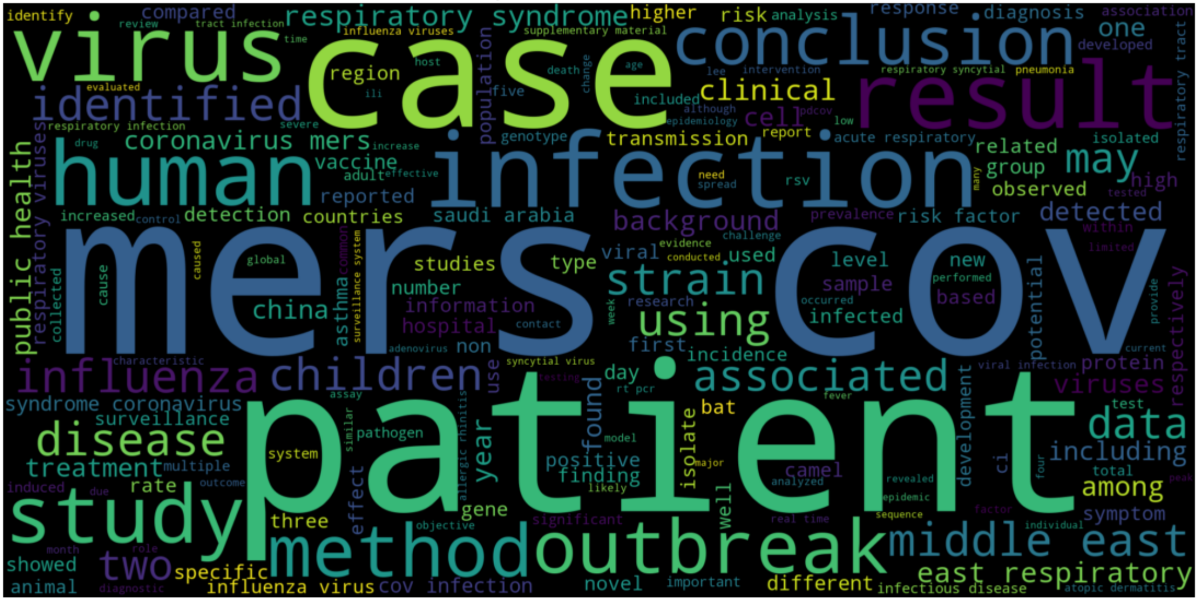
- Alphacoronavirus
- Betacoronavirus
- Middle East respiratory syndrome-related coronavirus (MERS-CoV)
- Severe acute respiratory syndrome coronavirus(SARS-CoV)

The first figure shows the distribution of the researches on a specific Virus type. According to the plot, we can find that the recent researches mainly focus on MERS-CoV and SARS-CoV. Moreover, there are fewer researches about MERS-CoV than SARS-CoV. The current research shows that the two species (MERS-CoV and SARS-CoV) produce potentially severe symptoms. Therefore, there might be more research focus on this two conornvirus specie.



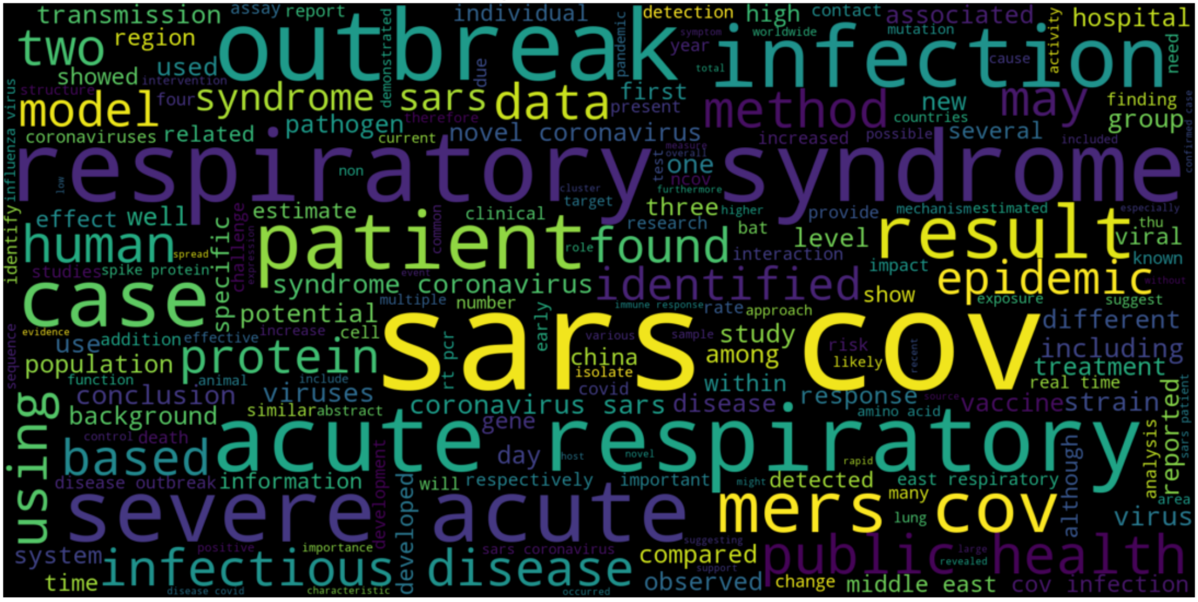
Secondly, we plot the word cloud to see the keyword for MERS-CoV and SARS-CoV virus. Below is the word Cloud for Keywords of Research about MERS. We can find that the key words includes MERS, COV, patient, infection, etc.

Keywords of Research about NMERS



Below is the word cloud for Keywords of Research about SARS. We can find keywords that include sars, cov, acute, respiratory, outbreak, etc.

Keywords of Research about SARS



Next, I will focus on the SARS-CoV since the disease COROV-19 is caused by a novel type of SARS-CoV virus.

I will first split the dataset. Then I will use the words in 'abstract' data to train multiple models. The model will predict the type of coronavirus that each research is mainly talking about. The goal is to find the best model that can help to select the researches that are more relevant to COVID-19.

A problem with modeling text is that it is messy, and techniques like machine learning algorithms prefer well defined fixed-length inputs and outputs. We cannot use the raw text as our training data. Moreover, some words or symbols with high frequency in the language but less value in machine learning such as ‘the’, ’#’, and ’is’...etc. Therefore, we need to remove them from our feature columns and split the sentences into the feature of word. We can choose to count Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF reflects how important a word is.

Model	Accuracy
Logistic Regression	70.39165%
K-NN	79.06732%
Naive Bayes	65.27818%
Decision Trees	70.64062%

From the previous result, we can observe that the Naive Bayes model performs worst in our case. Decision Trees and Logistic Regression model shows a similar performance. The k-NN model performs much better than other models in the TF-IDF dataset. Then we can run the grid search on the k-NN model to find the parameters that maximize the testing accuracy. The result for hyperparameters tuning in k-NN model is:

n_neighbors	weights	Accurcy
2	uniform	77.76501%
2	distance	71.32050%
5	uniform	79.06732%
5	distance	78.90453%
8	uniform	79.84296%
8	distance	79.61314%
10	uniform	79.96744%
10	distance	79.85253%
20	uniform	80.17811%
20	distance	80.14938%
30	uniform	80.22599%
30	distance	80.21641%

40	uniform	80.12065%
40	distance	80.20684%

In the previous step, we can find the model with the best performance is the k-NN model with parameters: n_neighbors=30, weights=uniform. Using this finding, we can easily find the research paper that is most related to the COVID-19 from their abstract.

In the previous step, we fitted the model with the research abstract data. Now, we can see how well the model is when applying the optimal model to the research title. This is the classification report for use the best model on the research paper’s title:

```
              precision    recall  f1-score   support

        MERS       0.00      0.00      0.00       200
        None       0.80      1.00      0.89      8349
        SARS       1.00      0.00      0.00      1894

    accuracy                           0.80     10443
   macro avg       0.60      0.33      0.30     10443
weighted avg       0.82      0.80      0.71     10443
```

The accuracy is 79.95% when we apply the optimal k-NN model to the title data and predict whether the research is related to the COVID-19. This result shows that the optimal model works well in this prediction. Thus, we can use this model as a tool to increase the efficiency of finding relative research information about COVID-19.

Considering the current situation of COROV-19, it is important to find an efficient way that can help us to find relative and useful information as soon as possible. Using the optimal K-NN model can help to achieve the goal. This model can take the natural language such as the research paper's abstract, full content, or title as input, then predict whether it is relevant to SARS (the virus causes COVID-19), MNERS, or another type of coronavirus.

Since the novel type of SARS is the virus that causes COVID-19, the previous research about SARS might be more helpful than other research about coronavirus. With the research papers which prediction result is SARS, we can pay more attention to these papers.

I hope this model could be useful in research about COVID-19. Moreover, the data are not limited to the title, abstract, or content of the research paper. Other resources like webpage can also apply this model. The model can also help increase the efficiency of searching relevant information about COVID-19.