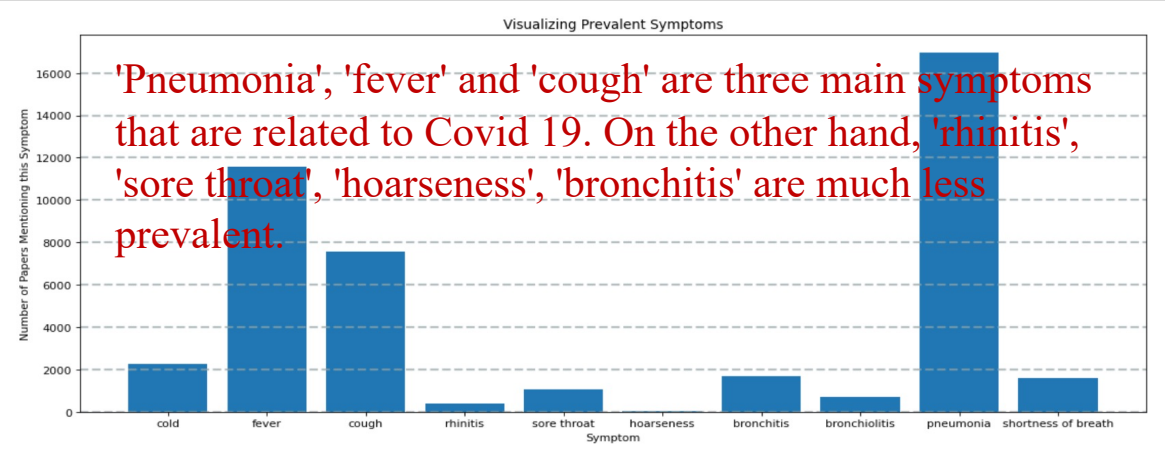
1624 Final Project

Exploratory Analysis
Model Feature Importance
Model Results
Visualization

Angela HSU 1003328874

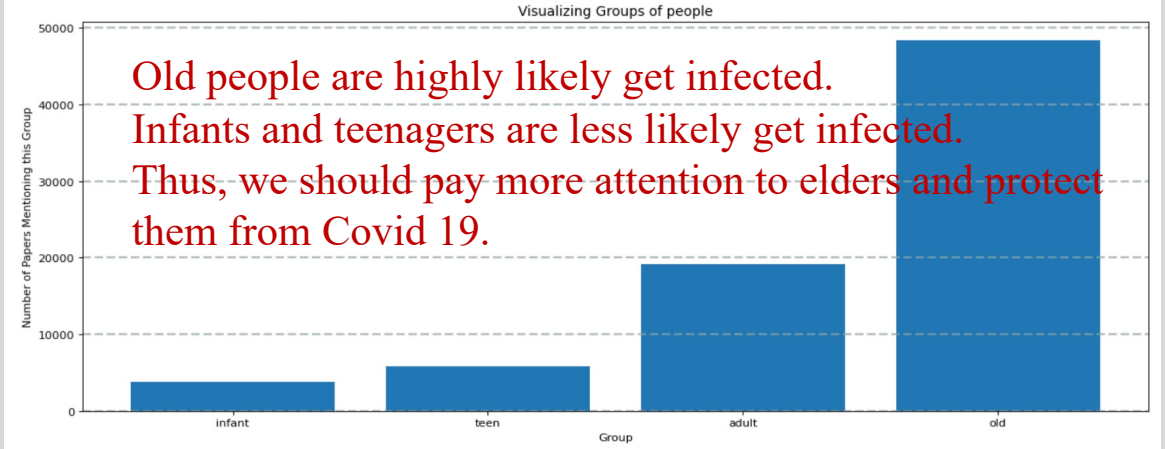
Exploratory Analysis

Visualize Prevalent Symptoms



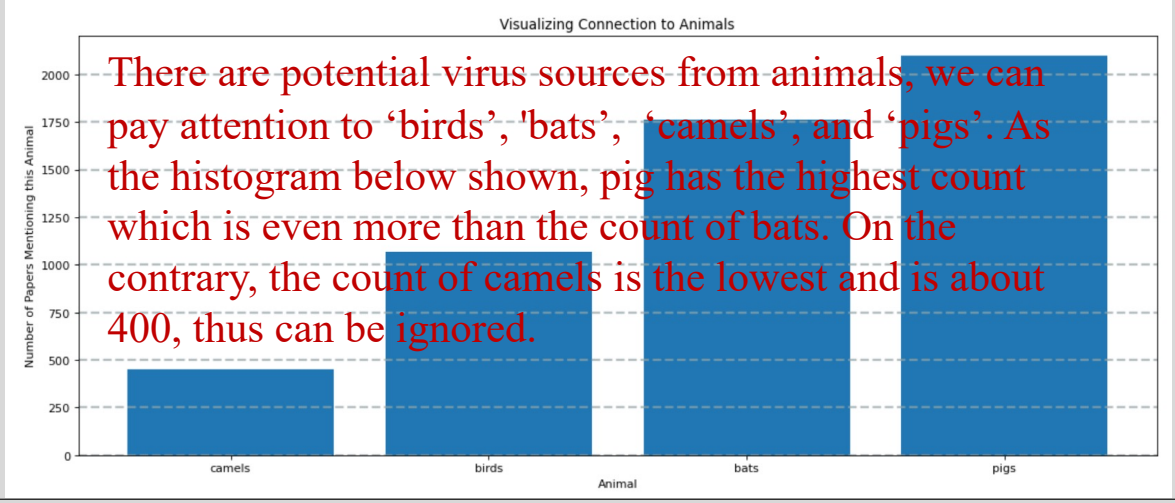
'Pneumonia', 'fever' and 'cough' are three main symptoms that are related to Covid 19. On the other hand, 'rhinitis', 'sore throat', 'hoarseness', 'bronchitis' are much less prevalent.

Visualize Age Groups



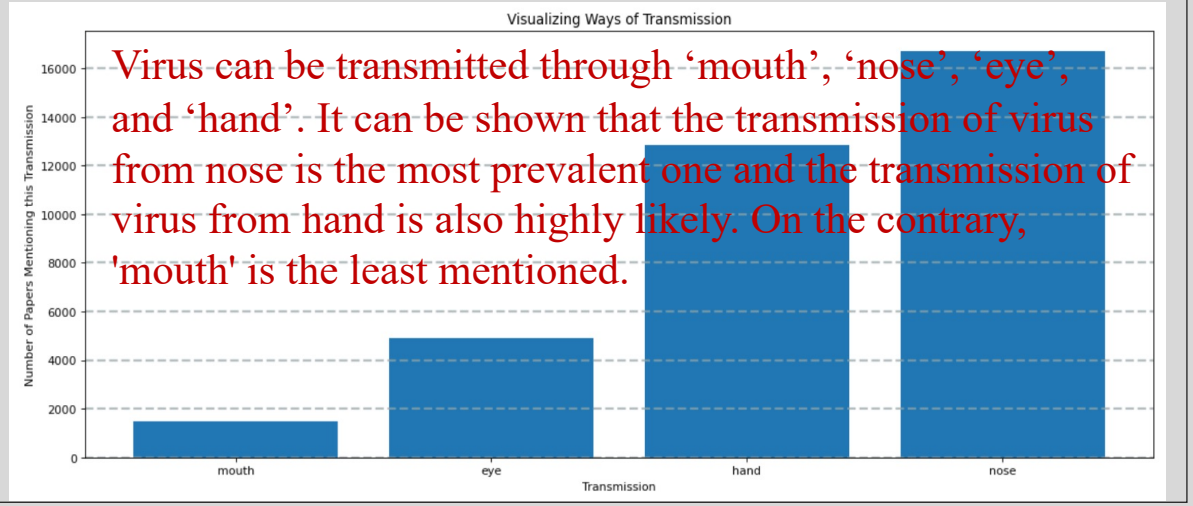
Old people are highly likely get infected.
Infants and teenagers are less likely get infected.
Thus, we should pay more attention to elders and protect them from Covid 19.

Visualize Connection to Animals



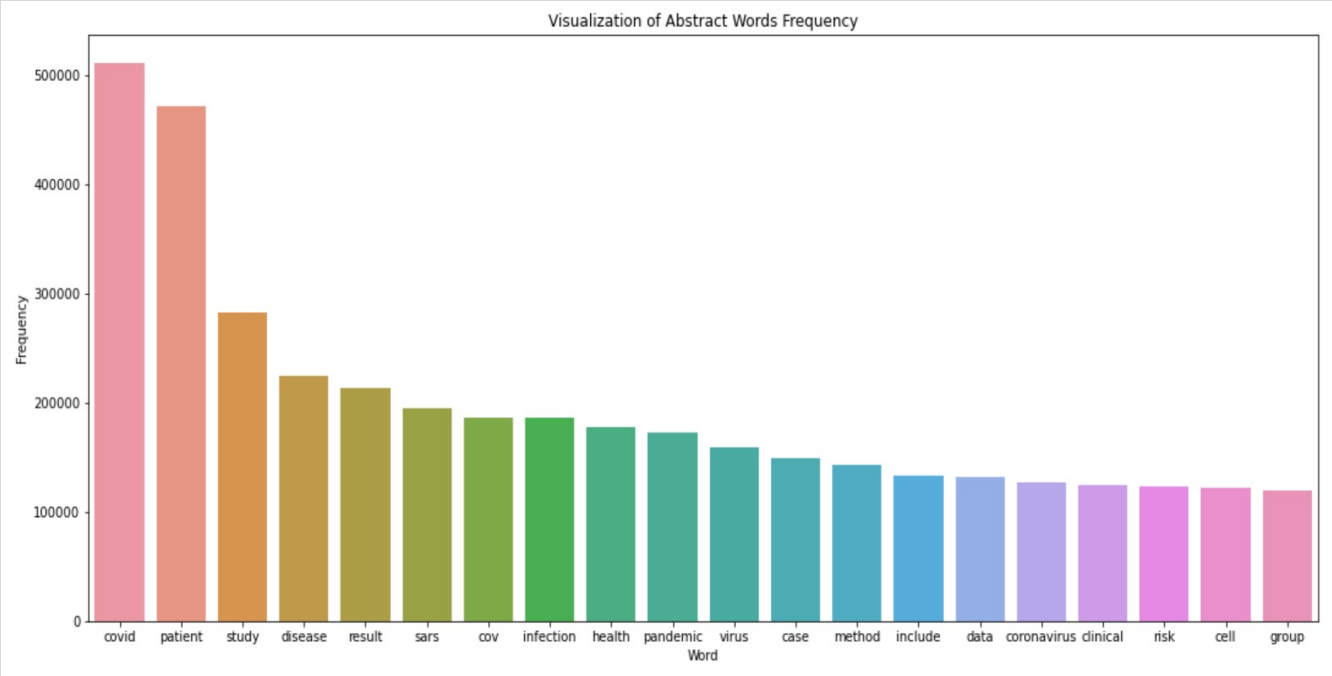
There are potential virus sources from animals, we can pay attention to ‘birds’, 'bats’, ‘camels’, and ‘pigs’. As the histogram below shown, pig has the highest count which is even more than the count of bats. On the contrary, the count of camels is the lowest and is about 400, thus can be ignored.

Visualize Ways of Transmissions

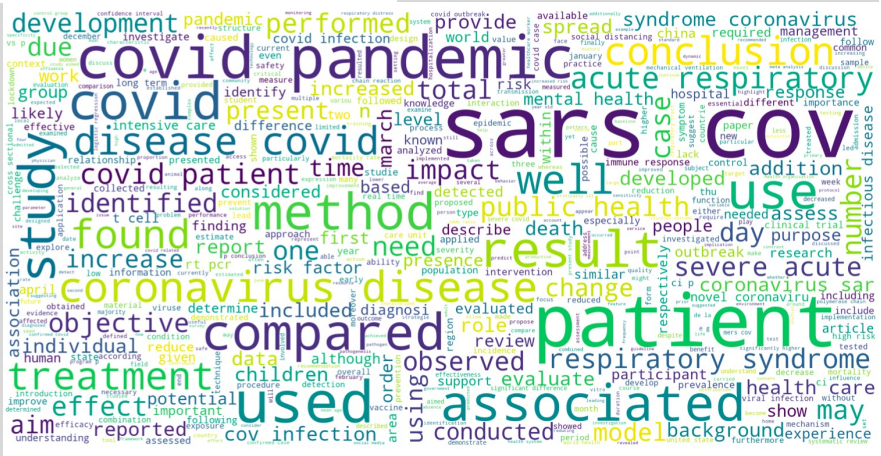


Virus can be transmitted through ‘mouth’, ‘nose’, ‘eye’, and ‘hand’. It can be shown that the transmission of virus from nose is the most prevalent one and the transmission of virus from hand is also highly likely. On the contrary, 'mouth' is the least mentioned.

Model Feature Importance



The 20 most frequent words from the abstract are shown in the graph, which are “covid, patient, study, disease, result, sars, cov....”.



	Word	Frequency
0	covid	511780
1	patient	471427
2	study	282750
3	disease	224585
4	result	214090
5	sars	194739
6	cov	186766
7	infection	185761
8	health	177699
9	pandemic	173097
10	virus	159562
11	case	149736
12	method	142850
13	include	133804
14	data	132132
15	coronavirus	127099
16	clinical	124747
17	risk	123401
18	cell	121736
19	group	119371

Supervised Learning Models x 3

Training set

```
Model:  Logistic Regression
Accuracy Score:  97.771 %
     Precision:  97.823 %
        Recall:  97.555 %
      F1 score:  97.688 %
```

Test set

```
Model:  Logistic Regression
Accuracy Score:  96.284 %
     Precision:  96.875 %
        Recall:  95.723 %
      F1 score:  96.278 %
```

Training set

```
Model:  Random Forest
Accuracy Score:  91.602 %
     Precision:  92.817 %
        Recall:  91.294 %
      F1 score:  91.920 %
```

Test set

```
Model:  Random Forest
Accuracy Score:  88.223 %
     Precision:  90.109 %
        Recall:  87.379 %
      F1 score:  88.357 %
```

Training set

```
Model:  Stochastic Gradient Descent
Accuracy Score:  93.299 %
     Precision:  93.724 %
        Recall:  93.211 %
      F1 score:  93.429 %
```

Test set

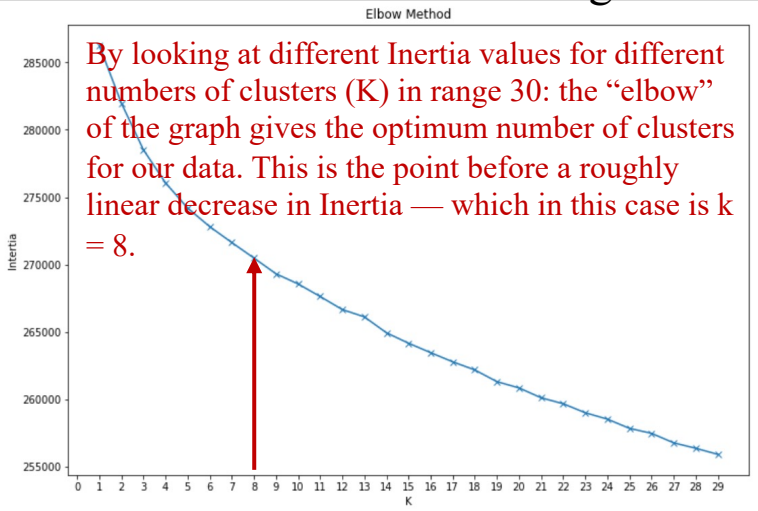
```
Model:  Stochastic Gradient Descent
Accuracy Score:  92.633 %
     Precision:  93.142 %
        Recall:  92.572 %
      F1 score:  92.805 %
```

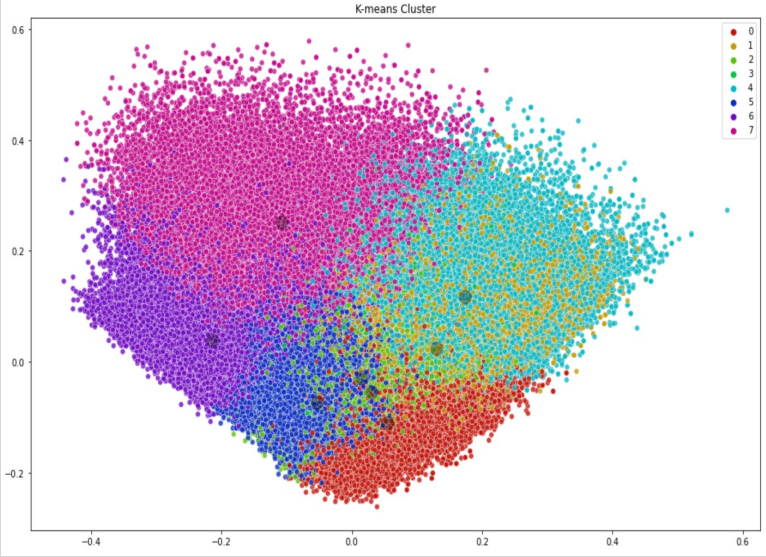
Among logistic regression, random forest, and stochastic gradient descent models, logistic regression model fits the best with highest accuracy and F1-score while random forest model fits the worst on both training and test set.

Also, the difference of accuracy score between training and test set is about 1% to 3%, suggesting no overfitting. High precision score relates to low false positive rates and high recall score relates to high rates of correctly predicted positive observations. F1-score is the weighted average of precision and recall.

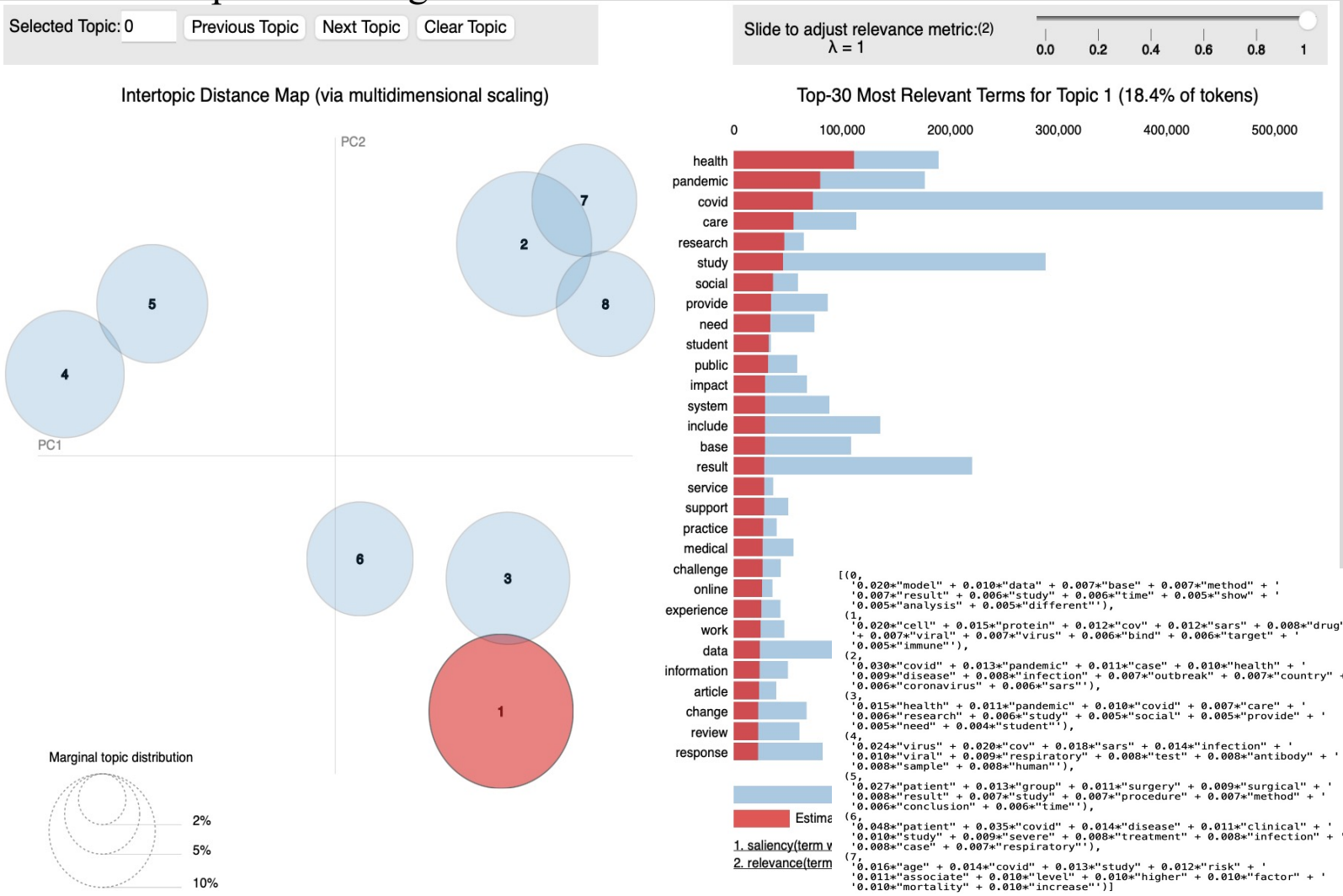
Unsupervised Learning Model LDA

K-Means Clustering

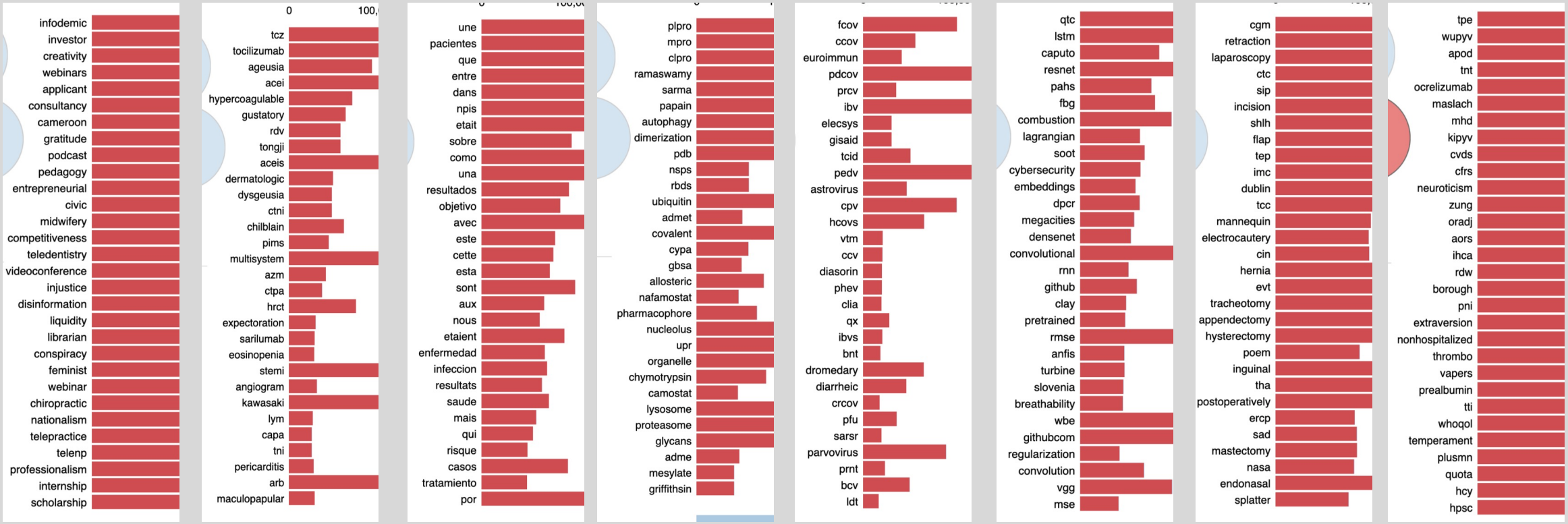




Topic Modeling LDA



Visualize 8 Topics-Keywords From LDA



Guidance: Scientists should learn more on the conoravirus variants and mutations, and develop more deep learning models on those. Doctors should strictly comply with the protective measures and technical recommendations for high-risk sugery such as "laparoscopy" and "tracheotomy". Nurse should pay more attention to elders since they are easily infected. Nurse should also be aware of the prevalent symptoms related to Covid 19 such as fever, cough, and pneumonia. Industry should pay more attention to pigs and because there are emerging conoravirus such as PEDV, PDCoV, and SADS-CoV in pigs. Government should host webinars or videoconference or podcast to give an pedagogy on Covid 19. Also, government could encourage psychological consultancy so that people could maintain a grateful and positive state, which may be helpful. Government should also inform the public to wear the masks all the time and wash hands more frequently.