Data cleaning

First data cleaning

- Select five columns: ‘title’, ‘abstract’, ‘authors’, and ‘publish time’
- Lower words in ‘title’ and ‘abstract’ and connect these two column as ‘text’

Identify coronavirus type

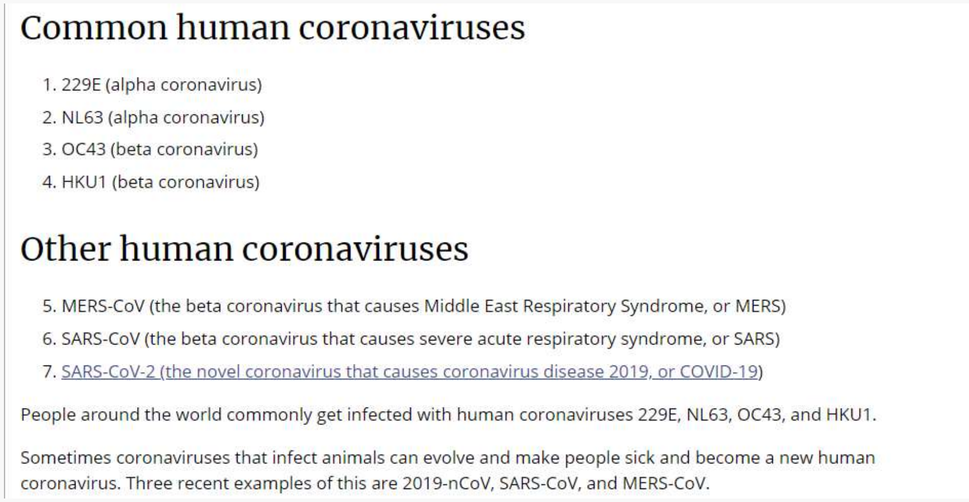
- Type_list = ['229e','nl63','oc43','hku1','mers-cov','mers','sars-cov','sars','sars-cov-2','covid-19’]
- Set other types of coronavirus or no specific type papers call ‘other’
- Set more than one coronavirus types call ‘multi’

Paper publish time

- For the pubilish_time column only keep the year value

COVID19 paper authors analyzation

- Split author name with “;”, and build author name list



Common human coronaviruses

1. 229E (alpha coronavirus)
2. NL63 (alpha coronavirus)
3. OC43 (beta coronavirus)
4. HKU1 (beta coronavirus)

Other human coronaviruses

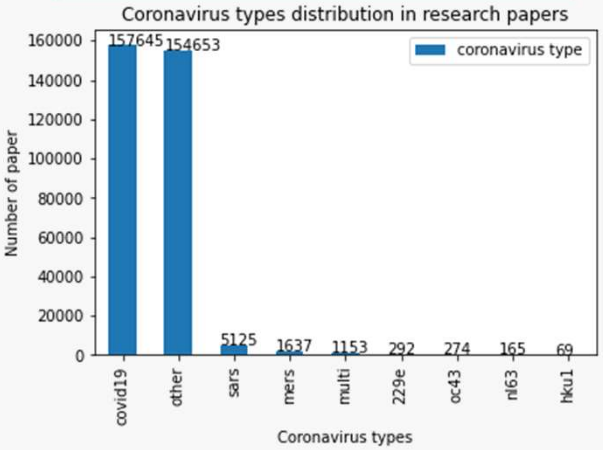
5. MERS-CoV (the beta coronavirus that causes Middle East Respiratory Syndrome, or MERS)
6. SARS-CoV (the beta coronavirus that causes severe acute respiratory syndrome, or SARS)
7. SARS-CoV-2 (the novel coronavirus that causes coronavirus disease 2019, or COVID-19)

People around the world commonly get infected with human coronaviruses 229E, NL63, OC43, and HKU1.

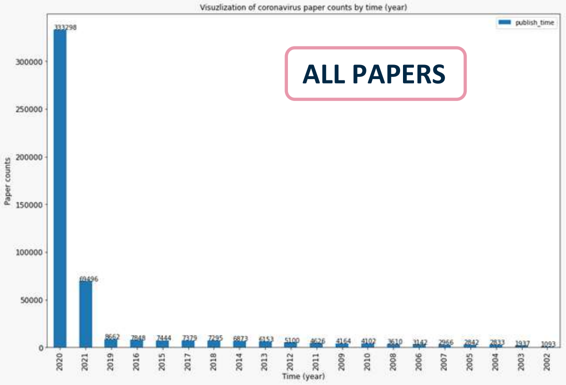
Sometimes coronaviruses that infect animals can evolve and make people sick and become a new human coronavirus. Three recent examples of this are 2019-nCoV, SARS-CoV, and MERS-CoV.

Data Visualization and Exploratory Data Analysis

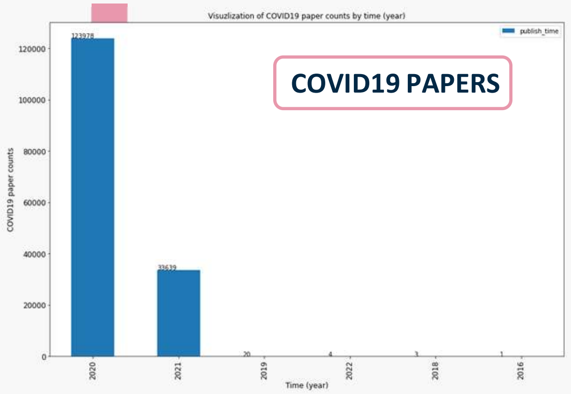
Visualization of paper coronavirus types



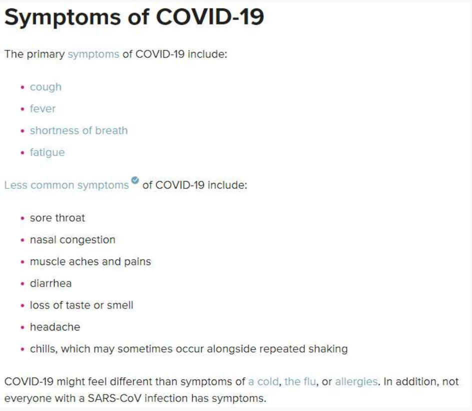
Paper publish time analyzation



ALL PAPERS



COVID19 PAPERS



Symptoms of COVID-19

The primary symptoms of COVID-19 include:

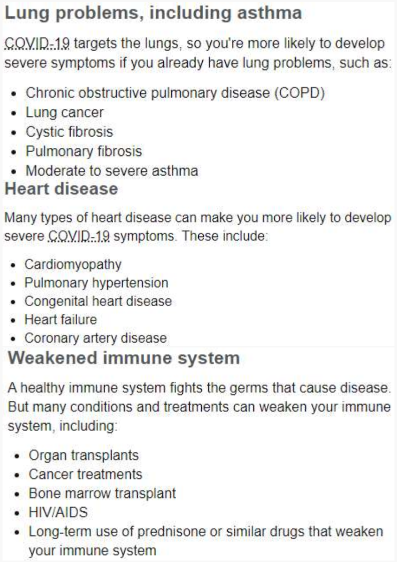
- cough
- fever
- shortness of breath
- fatigue

Less common symptoms of COVID-19 include:

- sore throat
- nasal congestion
- muscle aches and pains
- diarrhea
- loss of taste or smell
- headache
- chills, which may sometimes occur alongside repeated shaking

COVID-19 might feel different than symptoms of a cold, the flu, or allergies. In addition, not everyone with a SARS-CoV infection has symptoms.

```
symptoms = ['cough','fever','fatigue','sore throat','headache','diarrhea','chills','nasal congestion','pneumonia','bronchitis','cold','flu','allergies','sneezing','runny nose','stuffy','weakness','pains','aches']
```



Lung problems, including asthma

COVID-19 targets the lungs, so you're more likely to develop severe symptoms if you already have lung problems, such as:

- Chronic obstructive pulmonary disease (COPD)
- Lung cancer
- Cystic fibrosis
- Pulmonary fibrosis
- Moderate to severe asthma

Heart disease

Many types of heart disease can make you more likely to develop severe COVID-19 symptoms. These include:

- Cardiomyopathy
- Pulmonary hypertension
- Congenital heart disease
- Heart failure
- Coronary artery disease

Weakened immune system

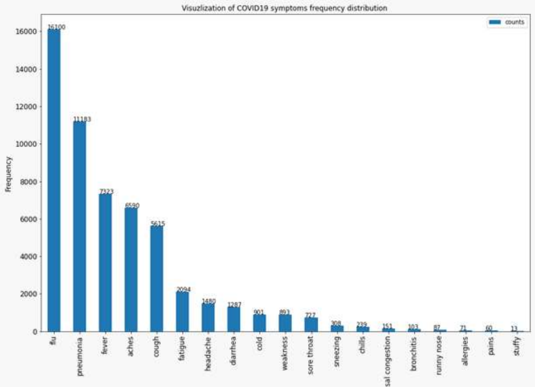
A healthy immune system fights the germs that cause disease. But many conditions and treatments can weaken your immune system, including:

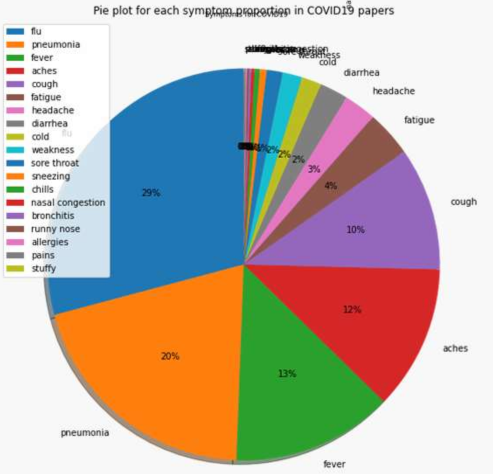
- Organ transplants
- Cancer treatments
- Bone marrow transplant
- HIV/AIDS
- Long-term use of prednisone or similar drugs that weaken your immune system

```
risk_factors = ['male', 'female', 'age', 'asthma', 'copd',
            'lung cancer','cystic fibrosis','pulmonary fibrosis',
            'heart disease','cardiomyopathy','heart failure','hypertension',
            'diabetes','obesity','cancer','hiv','aids','smoking', 'alcohol']
```

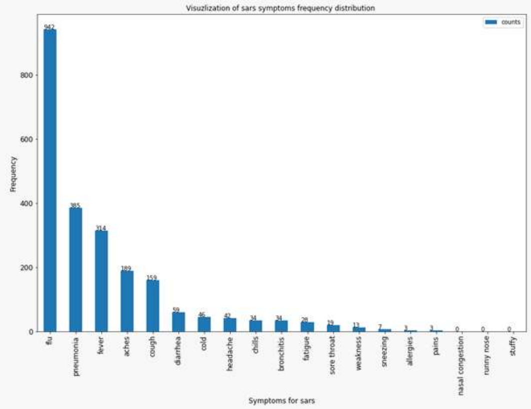
Data Visualization and Exploratory Data Analysis

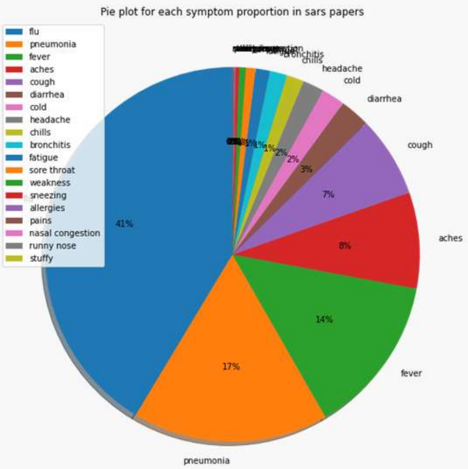
Symptoms frequency distribution in COVID19 papers



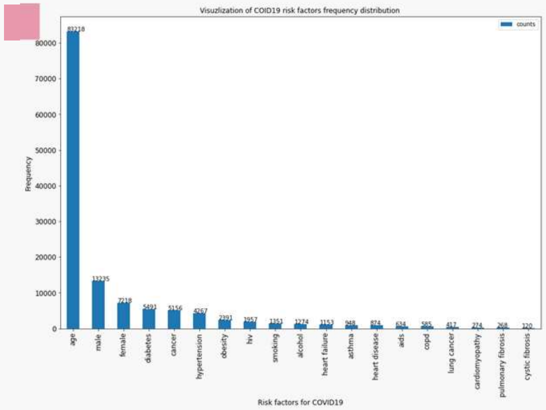


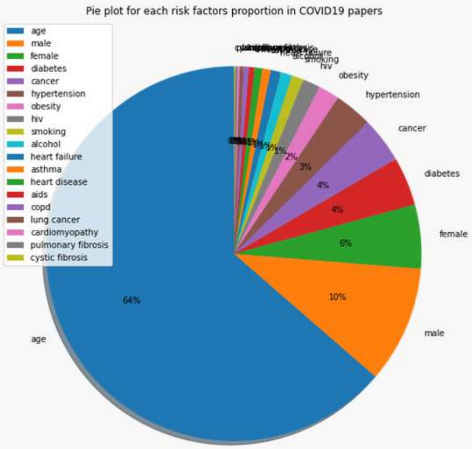
Symptoms frequency distribution in sars papers





Symptoms frequency distribution in COVID19 papers

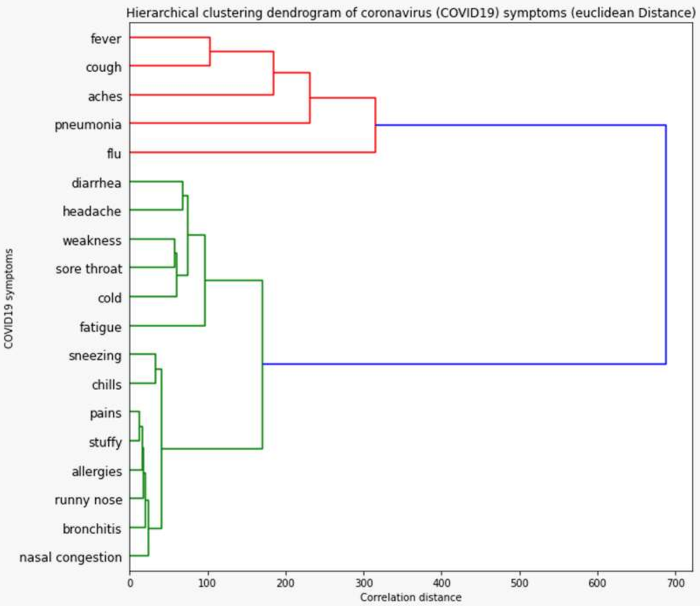




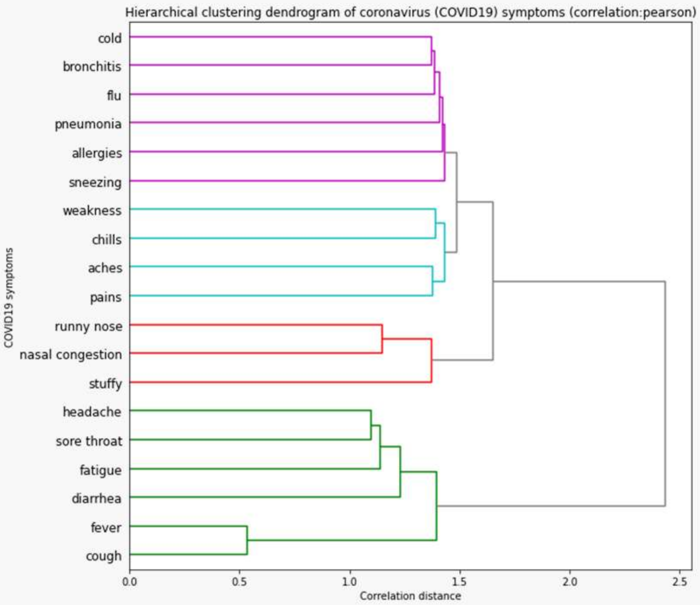
Model selection and fitting to data

Association of coronavirus symptoms:

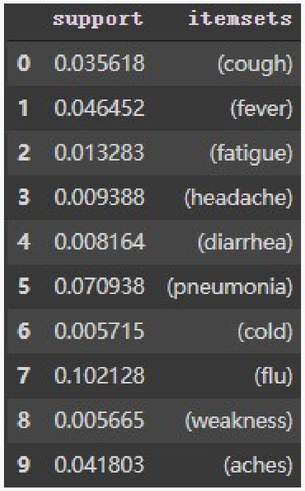
Analyze association of symptoms using Euclidean Distance



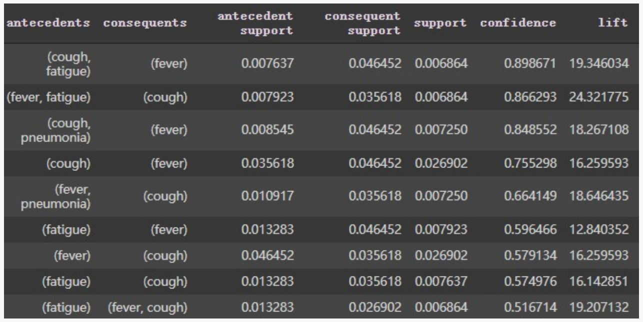
Analyze association of symptoms using correlation method



Symptoms analyzation by apriori algorithm



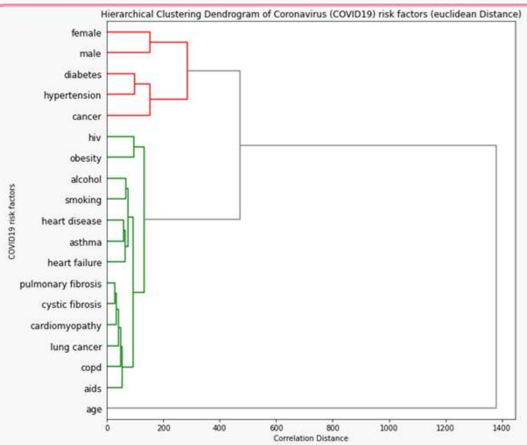
	support	itemsets
0	0.035618	(cough)
1	0.046452	(fever)
2	0.013283	(fatigue)
3	0.009388	(headache)
4	0.008164	(diarrhea)
5	0.070938	(pneumonia)
6	0.005715	(cold)
7	0.102128	(flu)
8	0.005665	(weakness)
9	0.041803	(aches)



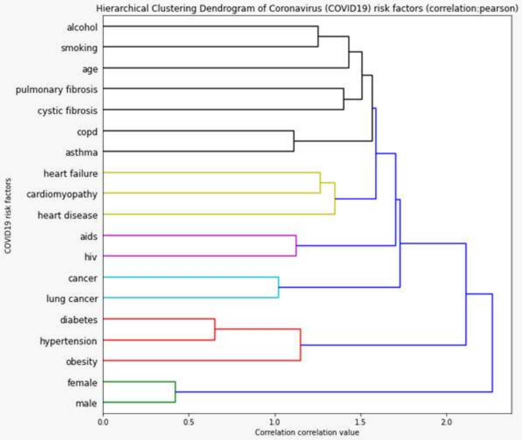
antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(cough, fatigue)	(fever)	0.007637	0.046452	0.006864	0.898671	19.346034
(fever, fatigue)	(cough)	0.007923	0.035618	0.006864	0.866293	24.321775
(cough, pneumonia)	(fever)	0.008545	0.046452	0.007250	0.848552	18.267108
(cough)	(fever)	0.035618	0.046452	0.026902	0.755298	16.259593
(fever, pneumonia)	(cough)	0.010917	0.035618	0.007250	0.664149	18.646435
(fatigue)	(fever)	0.013283	0.046452	0.007923	0.596466	12.840352
(fever)	(cough)	0.046452	0.035618	0.026902	0.579134	16.259593
(fatigue)	(cough)	0.013283	0.035618	0.007637	0.574976	16.142851
(fatigue)	(fever, cough)	0.013283	0.026902	0.006864	0.516714	19.207132

Association of coronavirus risk factors:

Association of risk factors using Euclidean Distance



Association of risk factors using correlation method



Deriving insights and guidance

- Determine the association of coronavirus symptoms
- Determine the association of coronavirus risk factors

Governments

- Governments should increase awareness and understanding of the disease
- Targeted outreach to at-risk populations

  For example, older adults, male, and people with other diseases.
- Quarantine and keep social distance

Hospital

- Quick screening and evaluation of patients which based on the symptoms' analyzation
- Risk prevention

  People with diabetes, cancer, hypertension, obesity, and hiv
  will get higher risk to get covid19.

Scientists

- Quick paper researching by author and paper classification