Text mining for exploration of COVID-19 severity factors

Khang Duy LAI - Mariia KLIMINA

University Paris Cité

UFR des Sciences Fondamentales et Biomédicales

May 12, 2022

2022-05-12 Text mining for exploration of COVID-19 severity factors


Text mining for exploration of COVID-19 severity factors

Khang Duy LAI - Mariia KLIMINA

University Paris Cité

UFR des Sciences Fondamentales et Biomédicales

May 12, 2022

1. Introduction
2. State of the art
3. Data exploration
4. Data preprocessing
5. Data processing
6. Result

Introduction
State of the art
Data exploration
Data preprocessing
Data processing
Result

Introduction

COVID-19 is the disease caused by the Sar-COV-2 virus that originated in China at the end of the year 2019. Over the time, studies have shown that there is some form of background diseases and risk factors that can hugely affect the severity cases rate of COVID-19. This project will apply NLP and text mining methods in order to explore the CORD-19 dataset and extract background diseases and risk factors.

Introduction

COVID-19 is the disease caused by the Sar-COV-2 virus that originated in China at the end of the year 2019. Over the time, studies have shown that there is some form of background diseases and risk factors that can hugely affect the severity cases rate of COVID-19. This project will apply NLP and text mining methods in order to explore the CORD-19 dataset and extract background diseases and risk factors.

State of the art

In this project we used multiple state of the art NLP and Data Science libraries.

- Numpy,Pandas: Formatting the data and the calculations.
- Matplotlib: Library for drawing the charts and figures.
- Scikit-learn: LDA and T-SNE models.
- Spacy,Gensim, and NLTK: Important NLP libraries.
- Scispacy: NER,Spacy models for science papers.
- Bokeh: A library for visualising interacted charts.

State of the art

In this project we used multiple state of the art NLP and Data Science libraries.

- Numpy,Pandas: Formatting the data and the calculations.
- Matplotlib: Library for drawing the charts and figures.
- Scikit-learn: LDA and T-SNE models.
- Spacy,Gensim, and NLTK: Important NLP libraries.
- Scispacy: NER,Spacy models for science papers.
- Bokeh: A library for visualising interacted charts.

Data exploration

CORD-19 dataset

Figure 1: CORD-19 Structure

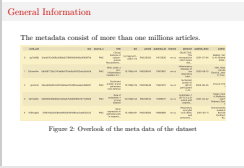
Data exploration
CORD-19 dataset
Figure 1: CORD-19 Structure

General Information

The metadata consist of more than one millions articles.

	cord_uid	sha	source_x	title	doi	pmcid	pubmed_id	license	abstract	publish_time	author
0	ug7v899j	d1aafb70c066a2068b02786f8929fd9c900897fb	PMC	Clinical features of culture-proven Mycoplasma...	10.1186/1471-2334-1-6	PMC35282	11472636	no-cc	OBJECTIVE: This retrospective chart review des...	2001-07-04	Madani, Tari- A; Al-Ghamd Aisha
1	02tnwd4m	6b0567729c2143a66d737eb0a2f63f2dce2e5a7d	PMC	Nitric oxide: a pro-inflammatory mediator in l...	10.1186/rr14	PMC59543	11667967	no-cc	Inflammatory diseases of the respiratory tract...	2000-08-15	Vliet, Alber van de Eiserich, Jaso P; Cros.
2	ejv2xln0	06ced00a5fc04215949aa72528f2eeaae1d58927	PMC	Surfactant protein-D and pulmonary host defense	10.1186/rr19	PMC59549	11667972	no-cc	Surfactant protein-D (SP-D) participates in th...	2000-08-25	Crouch, Erik
3	2b73a28n	3480556 49b6b8cf2b9a376498df9bf41f7123605	PMC	Role of endothelin-1 in lung disease	10.1186/rr44	PMC59574	11686871	no-cc	Endothelin-1 (ET-1) is a 21 amino acid peptide...	2001-02-22	Fagan, Kare A; McMurtry Ivan F Rodman, Davi-
4	9785vg6d	5f48792a5fa08bed9f56016f4981ae2ca6031b32	PMC	Gene expression in epithelial cells in respons...	10.1186/rr61	PMC59580	11686888	no-cc	Respiratory syncytial virus (RSV) and pneumoni...	2001-05-11	Domachowske Joseph E Bonville Cynthia A Ro.

Figure 2: Overlook of the meta data of the dataset



General Information

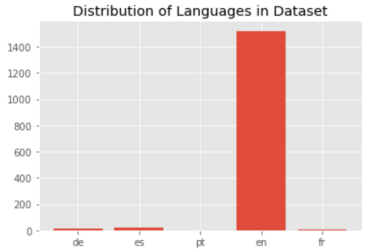
Columns in the metadata

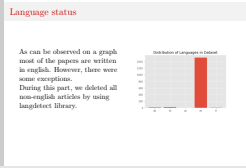
```
['cord_uid', 'sha', 'source_x', 'title', 'doi',
 'pmcid', 'pubmed_id', 'license', 'abstract',
 'publish_time', 'authors', 'journal',
 'mag_id', 'who_covidence_id', 'arxiv_id',
 'pdf_json_files', 'pmc_json_files', 'url', 's2_id']
```


General Information
Columns in the metadata
['cord_uid', 'sha', 'source_x', 'title', 'doi',
'pmcid', 'pubmed_id', 'license', 'abstract',
'publish_time', 'authors', 'journal',
'mag_id', 'who_covidence_id', 'arxiv_id',
'pdf_json_files', 'pmc_json_files', 'url', 's2_id']

Language status

As can be observed on a graph most of the papers are written in english. However, there were some exceptions.
During this part, we deleted all non-english articles by using langdetect library.





Data preprocessing

- Converting JSON format into DataFrame format.
- Removing all non-english paper.
- Removing special characters
- Removing numbers
- Tokenizing.
- Removing stopwords.
- Stemming.
- Lemmatisation.

Data preprocessing
- Converting JSON format into DataFrame format.
- Removing all non-english paper.
- Removing special characters
- Removing numbers
- Tokenizing.
- Removing stopwords.
- Stemming.
- Lemmatisation.

Data preprocessing

Data preprocessing

Data processing

- Data selection
  - Selecting articles with risk factors and severity key-words.
  - Clustering using Latent Dirichlet Allocation.
- NER (Named-entity recognition).

Data processing

- Data selection
  - Selecting articles with risk factors and severity key-words.
  - Clustering using Latent Dirichlet Allocation.
- NER (Named-entity recognition).

Risk factors and severity paper filtering

Dictionary of key words

Risk factors and severity paper filtering
Dictionary of key words

Risk factors and severity paper filtering

LDA

In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

The LDA algorithm structure:

- Providing to an algorithm a certain number of topics.
- The algorithm is assigning every word to a temporary topic.
- The algorithm is checking and updating topic assignments.

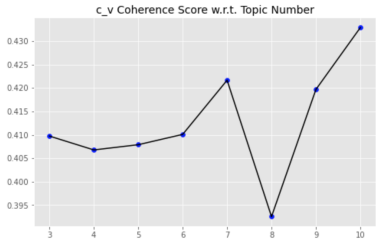
LDA

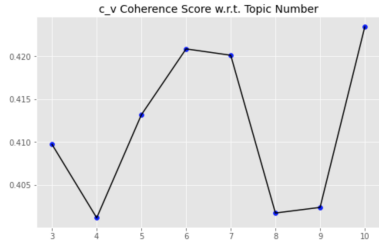
In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

The LDA algorithm structure:

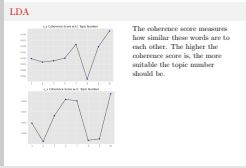
- Providing to an algorithm a certain number of topics.
- The algorithm is assigning every word to a temporary topic.
- The algorithm is checking and updating topic assignments.

LDA

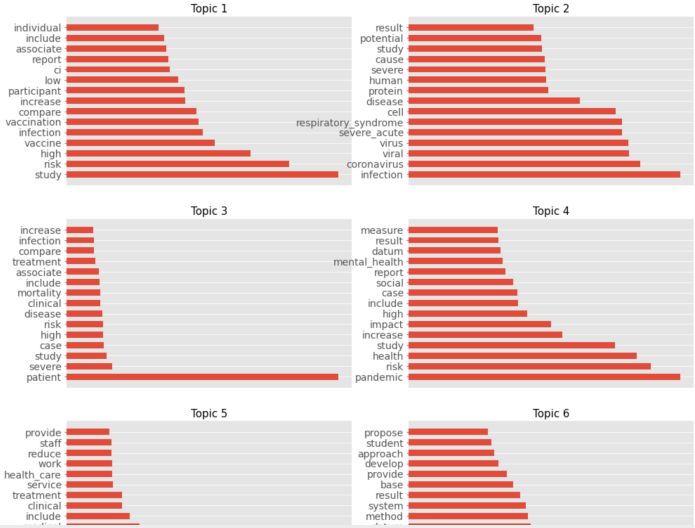


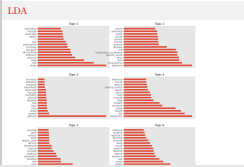


The coherence score measures how similar these words are to each other. The higher the coherence score is, the more suitable the topic number should be.



LDA


Topic 1
individual
include
associate
report
ci
low
participant
increase
compare
vaccination
infection
vaccine
high
risk
study
Topic 2
result
potential
study
cause
severe
human
protein
disease
cell
respiratory_syndrome
severe_acute
virus
viral
coronavirus
infection
Topic 3
increase
infection
compare
treatment
associate
include
mortality
clinical
disease
risk
high
case
study
severe
patient
Topic 4
measure
result
datum
mental_health
report
social
case
include
high
impact
increase
study
health
risk
pandemic
Topic 5
provide
staff
reduce
work
health_care
service
treatment
clinical
include
Topic 6
propose
student
approach
develop
provide
base
result
system
method


LDA

NER

NER

Result

Example of table of result

0	chronic obstructive pulmonary disease copd	DISEASE
1	death	DISEASE
3	copd	DISEASE
9	dyspnea	DISEASE
10	cough	DISEASE
11	copd pulmonary function	DISEASE
13	respiratory tract infection	DISEASE
14	chronic unstable disease system malignancy	DISEASE
19	obstructive pulmonary disease	DISEASE
21	copd airflow	DISEASE
25	hypertension	DISEASE
26	atherosclerotic heart disease	DISEASE

Result
Example of table of result
0 chronic obstructive pulmonary disease copd DISEASE
1 death DISEASE
3 copd DISEASE
9 dyspnea DISEASE
10 cough DISEASE
11 copd pulmonary function DISEASE
13 respiratory tract infection DISEASE
14 chronic unstable disease system malignancy DISEASE
19 obstructive pulmonary disease DISEASE
21 copd airflow DISEASE
25 hypertension DISEASE
26 atherosclerotic heart disease DISEASE
27 bronchiectasis DISEASE

Conclusion

- Basic diseases are filtered out
- The method have not sorted if the disease is a covid symtoms or related to the progress of severity cases.

Conclusion
- Basic diseases are filtered out
- The method have not sorted if the disease is a covid symtoms or related to the progress of severity cases.

Future improvement

- Create a knowledge graph
- Calculate the severity rate

Future improvement
- Create a knowledge graph
- Calculate the severity rate

Thank you

Thank you for your listening.

Thank you
Thank you for your listening.