Text mining for exploration of COVID-19 severity factors

Khang Duy LAI - Mariia KLIMINA

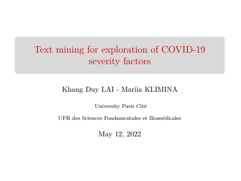
University Paris Cité

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

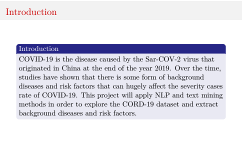
UFR des Sciences Fondamentales et Biomédicales

May 12, 2022

Introduction

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

Introduction

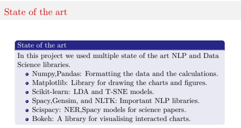
COVID-19 is the disease caused by the Sar-COV-2 virus that originated in China at the end of the year 2019. Over the time, studies have shown that there is some form of background diseases and risk factors that can hugely affect the severity cases rate of COVID-19. This project will apply NLP and text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

State of the art

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

State of the art

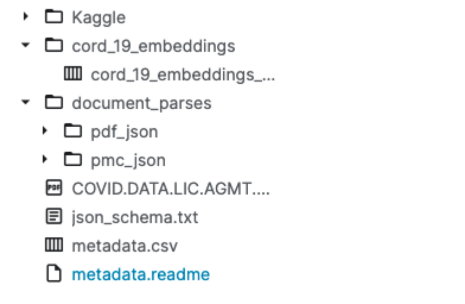
In this project we used multiple state of the art NLP and Data Science libraries.

- 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

Data exploration

CORD-19 dataset



- Kaggle
- cord_19_embeddings
  - cord_19_embeddings_...
- document_parses
  - pdf_json
  - pmc_json
- COVID.DATA.LIC.AGMT....
- json_schema.txt
- metadata.csv
- metadata.readme

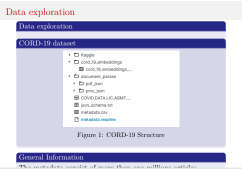
Figure 1: CORD-19 Structure

General Information

2022-05-12

Text mining for exploration of COVID-19 severity factors

└Data exploration



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

Data preprocessing

Data preprocessing

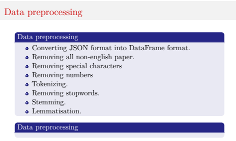
- Converting JSON format into DataFrame format.
- Removing all non-english paper.
- Removing special characters
- Removing numbers
- Tokenizing.
- Removing stopwords.
- Stemming.
- Lemmatisation.

Data preprocessing

2022-05-12

Text mining for exploration of COVID-19 severity factors

└Data preprocessing



Data preprocessing

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 processing

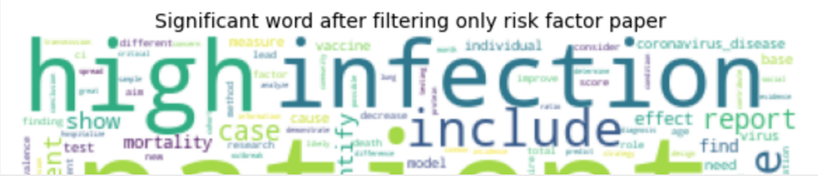
Data processing

- 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

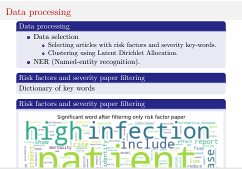
Risk factors and severity paper filtering


Significant word after filtering only risk factor paper

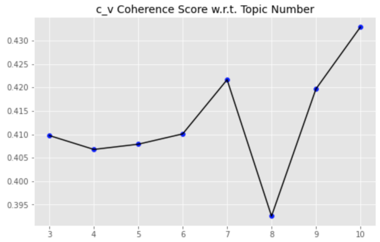
2022-05-12

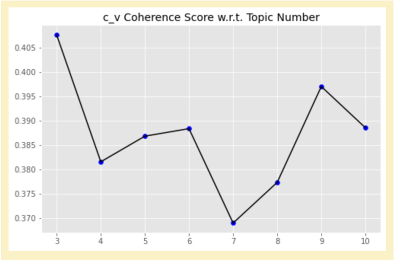
Text mining for exploration of COVID-19 severity factors

└─Data processing


Data processing
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
Significant word after filtering only risk factor paper

LDA



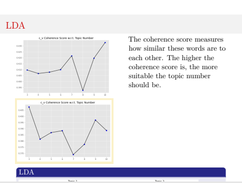


The coherence score measures how similar these words are to each other. The higher the coherence score is, the more suitable the topic number should be.

2022-05-12

Text mining for exploration of COVID-19 severity factors

└LDA



Result

Result

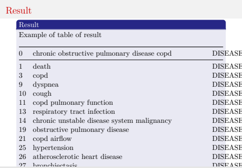
Example of table of result

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

2022-05-12

Text mining for exploration of COVID-19 severity factors

└─Result



Result

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