

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

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.

2022-05-12

Text mining for exploration of COVID-19 severity factors

- └ State of the art
- └ 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

CORD-19 dataset











-  Kaggle
- ▼  cord_19_embeddings
 -  cord_19_embeddings_...
- ▼  document_parsers
 -  pdf_json
 -  pmc_json
 -  COVID.DATA.LIC.AGMT....
 -  json_schema.txt
 -  metadata.csv
 -  [metadata.readme](#)

Figure 1: CORD-19 Structure

2022-05-12

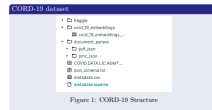
Text mining for exploration of COVID-19 severity

factors

└ Data exploration

└ Data exploration

Data exploration

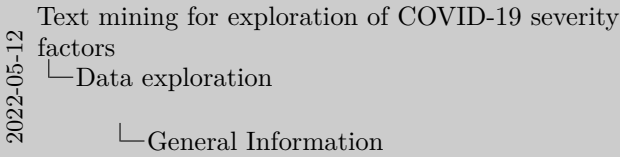


General Information

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, Tariq A; Al-Ghamdi 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 der Eiserich, Jasco 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 t
3	2b73a28n	348055649b6b8cf2b9a376498df9bf41f7123605	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, David h
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	Domachowski Joseph E Bonville Cynthia A Ro.

Figure 2: Overview 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']
```

2022-05-12

Text mining for exploration of COVID-19 severity factors

- Data exploration
- General Information

General Information

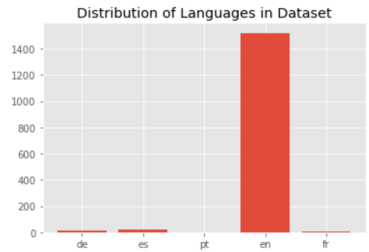
Columns in the metadata

```
['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.

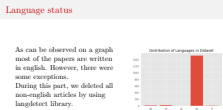


2022-05-12

Text mining for exploration of COVID-19 severity factors

Data exploration

Language status



Data preprocessing

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

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

Converting JSON format into DataFrame format.

- Merge body text into the same dataframe with metadata
- Add column to define language of paper
- Remove unnecessary columns.

Remove

- Filter out only paper with English
- Using regex to remove special characters, numbers.
- Remove stopwords

2022-05-12

Text mining for exploration of COVID-19 severity factors

└ Data preprocessing

└ Data preprocessing

Data preprocessing

Converting JSON format into DataFrame format.

- Merge body text into the same dataframe with metadata
- Add column to define language of paper
- Remove unnecessary columns.

Remove

- Filter out only paper with English
- Using regex to remove special characters, numbers.
- Remove stopwords

Data preprocessing

Stemming and Lemmatisation

- Lowers inflection in words to their root forms
- (connections, connected, connects, is connect)
- Using NLTK library

2022-05-12

Text mining for exploration of COVID-19 severity factors

└ Data preprocessing

└ Data preprocessing

Data preprocessing

Stemming and Lemmatisation

- Lowers inflection in words to their root forms
- (connections, connected, connects, is connect)
- Using NLTK library

Data preprocessing

[35]: 'quarantine isolation main containment strategy intended help protect public preventing spread contagious disease strategy primarily refer r
striction movement limitation personal contact quarantine definition person exposed disease isolation contagious person require separation
person infected finding previous research pointed increased risk negative psychological outcome depression anxiety isolation quarantined per
son equally heightened risk adverse mental health outcome rapid review brook reported increased negative psychological outcome including pos
ttraumatic stress symptom confusion anger person quarantine author concluded important stressor longer quarantine electronic supplementary m
aterial online version article doiorgs x contains supplementary material available authorized user duration infection fear frustration bored
om inadequate supply inadequate information financial loss stigma finding suggest containment strategy quarantine isolation negative impact
psychological outcome related broad spectrum psychosocial stressor need investigation mental health problem associated containment strategy
highlighted rising implementation quarantine isolation worldwide currently ongoing covid pandemic unprecedented number people worldwide affe
cted quarantine isolation identification individual elevated risk adverse mental health effect mandatory suggested vulnerable population ris
k negative psychological outcome implementation containment strategy eg person mental illness low income lack social network particular grea
ter risk quarantine isolation world health organization included covid list disease pathogen prioritized research development rd public heal
th emergency context pose greatest public health risk epidemic potential insufficient countermeasure established containment strategy main c
ountermeasure context systematic investigation evidence concerning psychological effect urgently need single study review suggest increased
risk negative psychological outcome person quarantine isolation presented partially contradicting result furthermore prevalence estimate poi
nt elevated level adverse outcome quarantined isolated population validity finding limited underlying uncontrolled study design conducted sy
stematic literature review metaanalysis mental health effect quarantine isolation based controlled primary study data best knowledge metaana
lysis including quarantine isolation exists date systematic literature review metaanalysis protocol project published prospero prospero regi
strationno crd method followed guideline cochrane collaboration conduction systematic review searched pubmed psycinfo embase database study
restriction beginning searched time period april assessing rate psychological effect quarantinedisolated person compared nonquarantinednonis
olated person search entry described online supplement supplement database search entry broad specific search term combined increase likelih
ood detecting eligible study research aim specific search term included list disease pathogen prioritized research development rd public hea
lth emergency context world health organization covid additional record identified manual search reference included study included language
restriction translation native speaker acquired test eligibility criterion article language english study author contacted case missing data
search carried endnote x clarivate analytics philadelphia usa trial considered appropriate test hypothesis included met following criterion
observation person quarantine isolation described second quantitative assessment psychological outcome parameter performed comparators perso
n quarantine isolation fourth data calculation effect size corresponding measure dispersion provided study observing psychological outcome p
arameter qualitative assessment excluded study excluded focused specific subpopulation primary infection controlassociation isolated person
prison study assessing correlation mental health outcome varying duration quarantine isolation excluded quantitative synthesis reported qual
itative synthesis determinant entire literature search study screening carried independently reviewer f jvb consensus unclear case reached d
iscussion additional member reviewing team lb jh testing eligibility criterion study selection classification coding data predefined excel s
preadsheet microsoft excel mac version microsoft corporation usa followed recommendation cochrane collaboration handbook performed indepen
dently reviewer lb jh reviewer jh lb independently extracted data characteristic study study sample quantitative data severity mean score freq
uency incidence prevalence mental health outcome group comparison group eg relative risk odds ratio result determinant testing reported reac
h statistical significance original study multiple measure outcome reported extracted data following hierarchy continuous measure mean score
categorical measure highest cutoff defined author original study ie severe manifestation disorder risk bias study classified independently r
eviewer lb jh according newcastleottawa scale no recommended cochrane handbook table summary assessment study classified holding low unknown
high risk bias taking account bias main domain selection comparability exposureoutcome disagreement resolved consensus additional review aut
hor calculated standardized mean difference smd confidence interval ci outcome measure primary study respective measure dispersion available
calculated ci p value recommended cochrane handbook stratified predefined mental health outcome effect size comparison quarantinedisolated n
onquarantinedisolated group summarized forest plot table quantitative synthesis result possible heterogeneity included study methodology pop
ulation outcome restricted quantitative synthesis predefined outcome primary study provided data categorical outcome based validated diagnos

2022-05-12

Text mining for exploration of COVID-19 severity factors

└─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).

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

- Create a dictionary of key words related to risk factors and severity.
- Filter out only the paper that contain words in the dictionary.

2022-05-12

Text mining for exploration of COVID-19 severity factors

- └ Data processing
 - └ Risk factors and severity paper filtering

Risk factors and severity paper filtering

- Create a dictionary of key words related to risk factors and severity.
- Filter out only the paper that contain words in the dictionary.

Risk factors and severity paper filtering



2022-05-12

Text mining for exploration of COVID-19 severity factors

- └ Data processing
- └ Risk factors and severity paper filtering

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.

2022-05-12

Text mining for exploration of COVID-19 severity factors

- └ Data processing
 - └ LDA

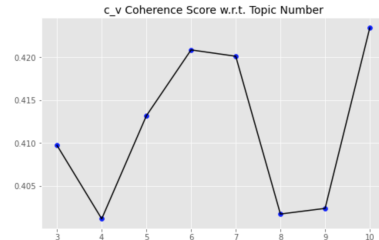
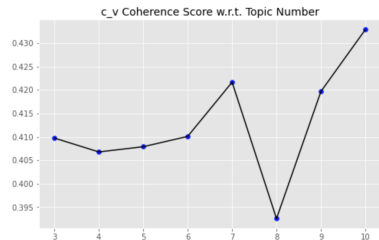
LDA

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.

2022-05-12

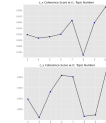
Text mining for exploration of COVID-19 severity

factors

└ Data processing

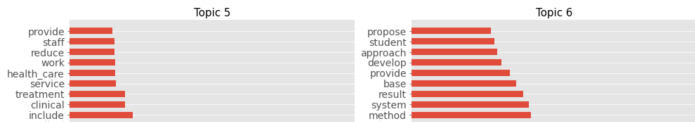
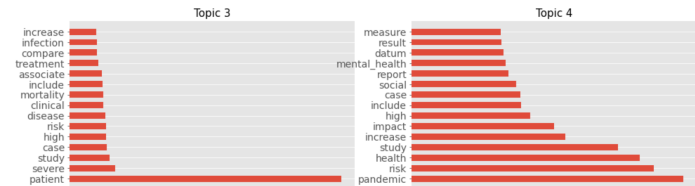
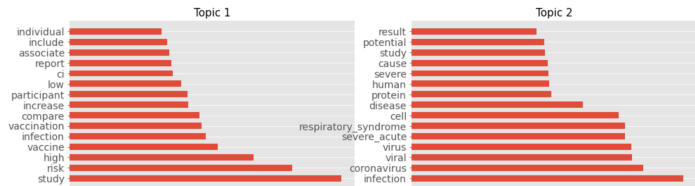
└ LDA

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



2022-05-12

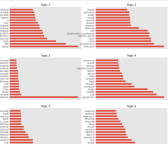
Text mining for exploration of COVID-19 severity

factors

└ Data processing

└ LDA

LDA



Named entity recognition (NER) is probably the extraction-method that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages. In our case it’s biomedical entities: diseases,chemicals etc.

We used Scispacy library with different SpaCy models for biomedical text processing.

2022-05-12

Text mining for exploration of COVID-19 severity factors

- Data processing
 - NER

NER

Named entity recognition (NER) is probably the extraction-method that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages. In our case it's biomedical entities: diseases,chemicals etc.

We used Scispacy library with different SpaCy models for biomedical text processing.

Khang Duy LAI - Mariia KLIMINA

Université Paris Cité

Text mining for exploration of COVID-19 severity factors19 of 24

NER

Scispacy

- Science pretrain model
- en_ner_bc5cdr_md for biology, mark diseases and chemicals.

2022-05-12

Text mining for exploration of COVID-19 severity factors
└ Data processing
└ NER

NER

Scispacy

• Science pretrain model

• en_ner_bc5cdr_md for biology, mark diseases and chemicals.

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

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.

2022-05-12

Text mining for exploration of COVID-19 severity factors

└ Result

└ Conclusion

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

2022-05-12

Text mining for exploration of COVID-19 severity factors
└ Result
└ Future improvement

Future improvement

- Create a knowledge graph
- Calculate the severity rate

Introduction	State of the art	Data exploration	Data preprocessing	Data processing	Result
○	○	○○○○	○○○○	○○○○○○○○	○○○●
Thank you					
Thank you for your listening.					
Khang Duy LAI - Mariia KLIMINA					
Text mining for exploration of COVID-19 severity factors					

2022-05-12

Text mining for exploration of COVID-19 severity factors

└ Result

└ Thank you

Thank you

Thank you for your listening.