Visualization and search tool for COVID-19 data

Zhengjie He A1808830

Contents

- □ Data preprocessing
- □ Visualization
- □ Sentence embedding
- □ Search tool

Data preprocessing

- First step: Set the path and combine the names of all PDF JSON files.
- Second step: Read the JSON file in turn to obtain the Paper ID, Title, Author, Abstract, Text, etc.
- Third step: Write the resulting dataframe to CSV so that it can be easily read next time.

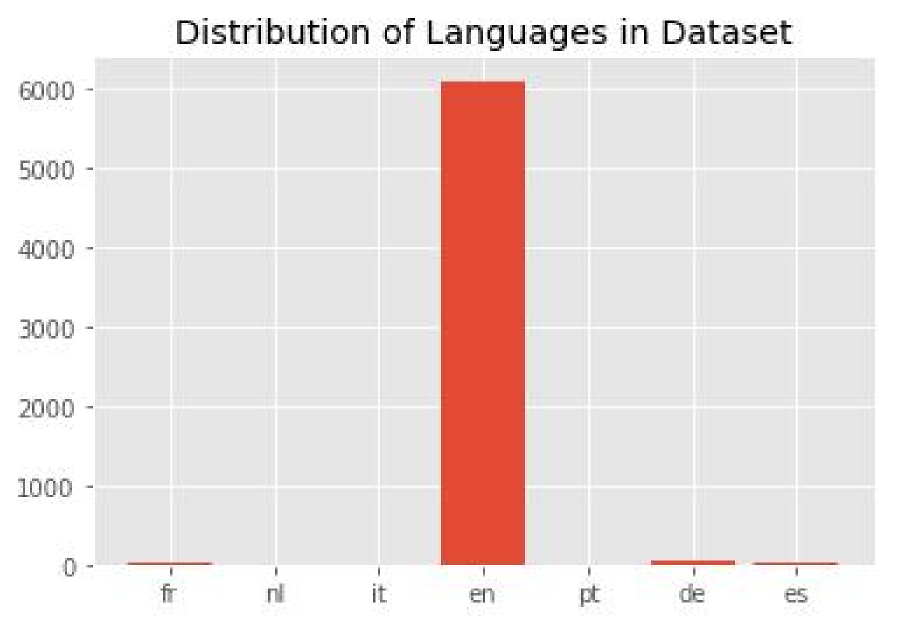
Data preprocessing

- Fourth step: Filter 10000 data from 1 million data as the data base.
- Fifth step: Remove null value.
- Sixth step: Determine the language of each document and keep the English document only.
- Seventh step: Reorder

Visualization

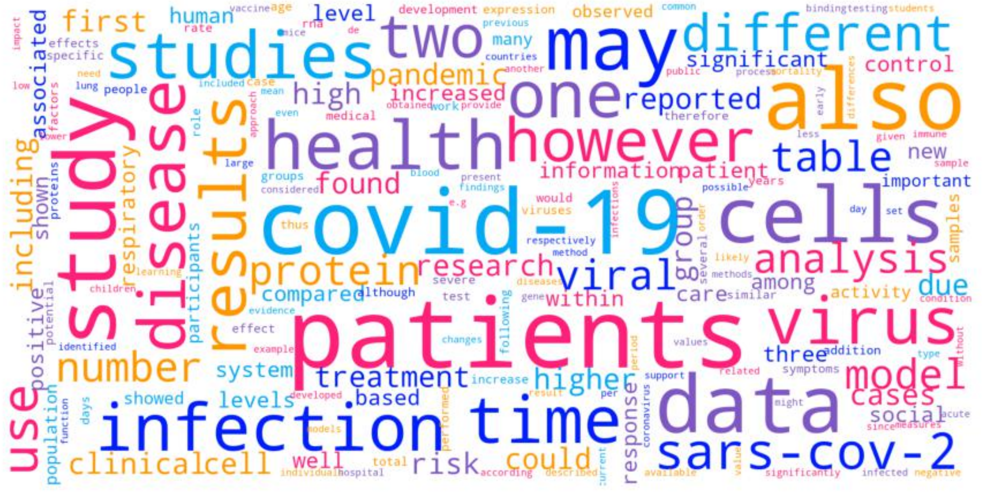
Tag articles by language

Languages include: English, French, Italian, Spanish, Portuguese, German, etc. Among them, English articles are far more than articles in other languages.



Tag articles by word

From the result we can see that most literatures focus on the situation of patients. Just as the key words "patient", "health" and "cell" have always appeared.

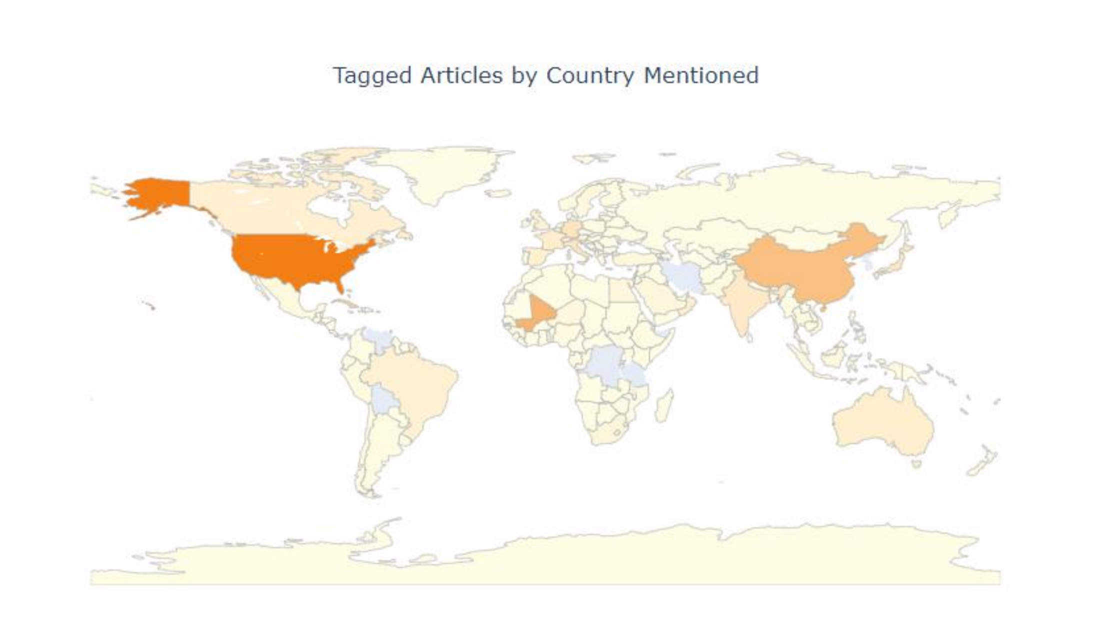


Most frequent words in article titles tagged as COVID-19

Visualization

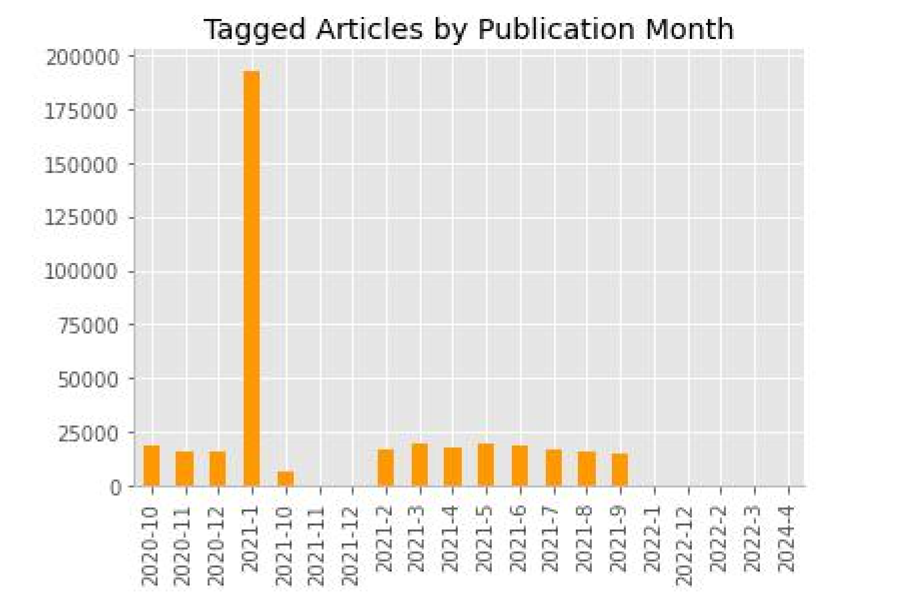
Tag articles by countries

From the key words about countries selected in the text, it can be seen that the United States is mentioned far more frequently than other countries. On the other hand, China and some countries are often concerned.

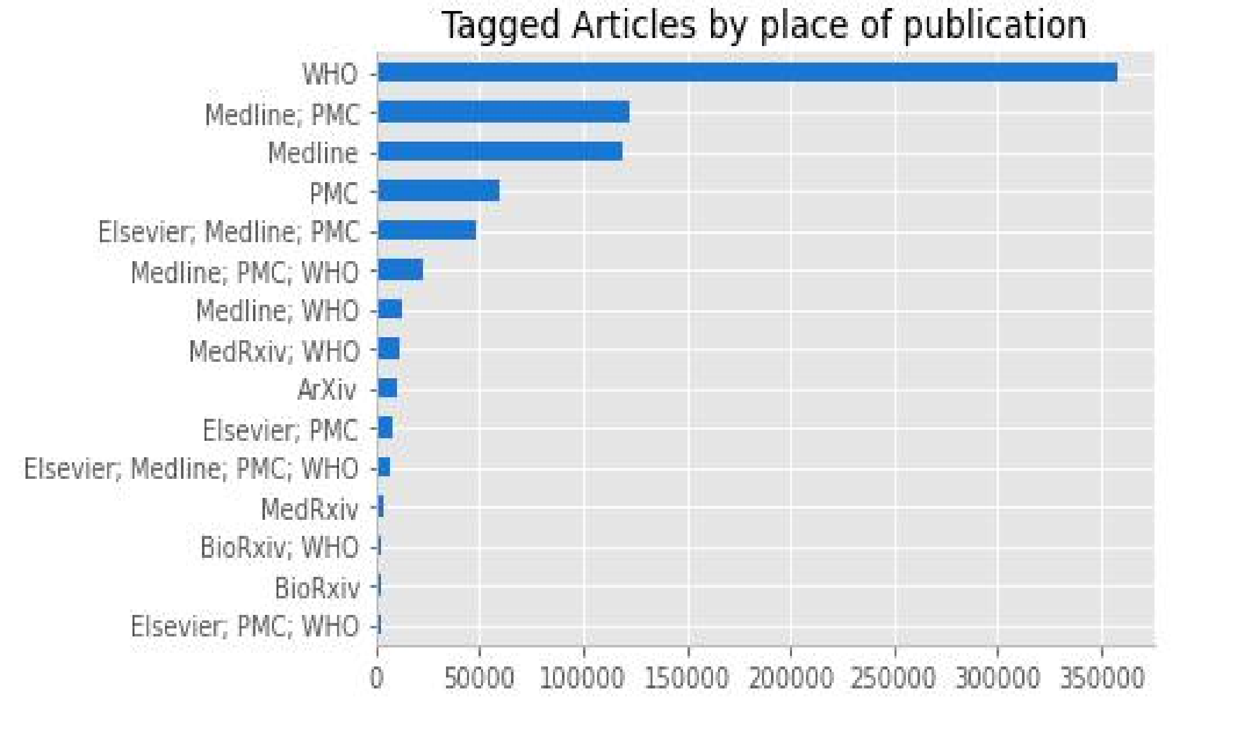

Tagged Articles by Country Mentioned

Tag articles by publish time

Surprisingly, after screening the literature after 2020 (because covid-19 has not erupted before 2021), I counted the publication date of the literature and found that a large number of relevant papers were published in January 2021, even exceeding the total number of other papers published so far.


Tagged Articles by Publication Month

Visualization



Tag articles by publishing platform

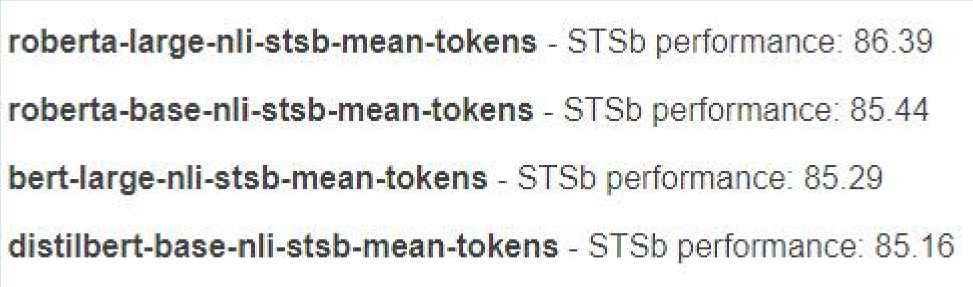
I also counted the publishing platforms in the data literature. As can be seen from the picture, who has attracted the most researchers to publish articles, followed by Medline and PMC

Sentence embedding

Sentence embedding

SentenceTransformer

Sense transformers is an unsupervised sentence embedding model. The model is based on Transformer Models (such as Bert, Roberta, distilbert, Albert, xlnet, etc.) and fine tuned for semantic similarity. Therefore, it can be used for the following tasks: semantic text similarity, clustering and semantic search.



roberta-large-nli-stsb-mean-tokens - STSb performance: 86.39

roberta-base-nli-stsb-mean-tokens - STSb performance: 85.44

bert-large-nli-stsb-mean-tokens - STSb performance: 85.29

distilbert-base-nli-stsb-mean-tokens - STSb performance: 85.16

roberta-large-nli-stsb-mean-tokens

It can be seen that the Roberta large NLI stsb mean tokens model achieves the highest score in the semantic text similarity test of similar models. Therefore, it is suitable for sentence learning and constructing search tools.

Search tool

Cut sentence

Using the send tokenize tool in nltk library can easily cut sentences in English documents.

Calculate mathematical vector

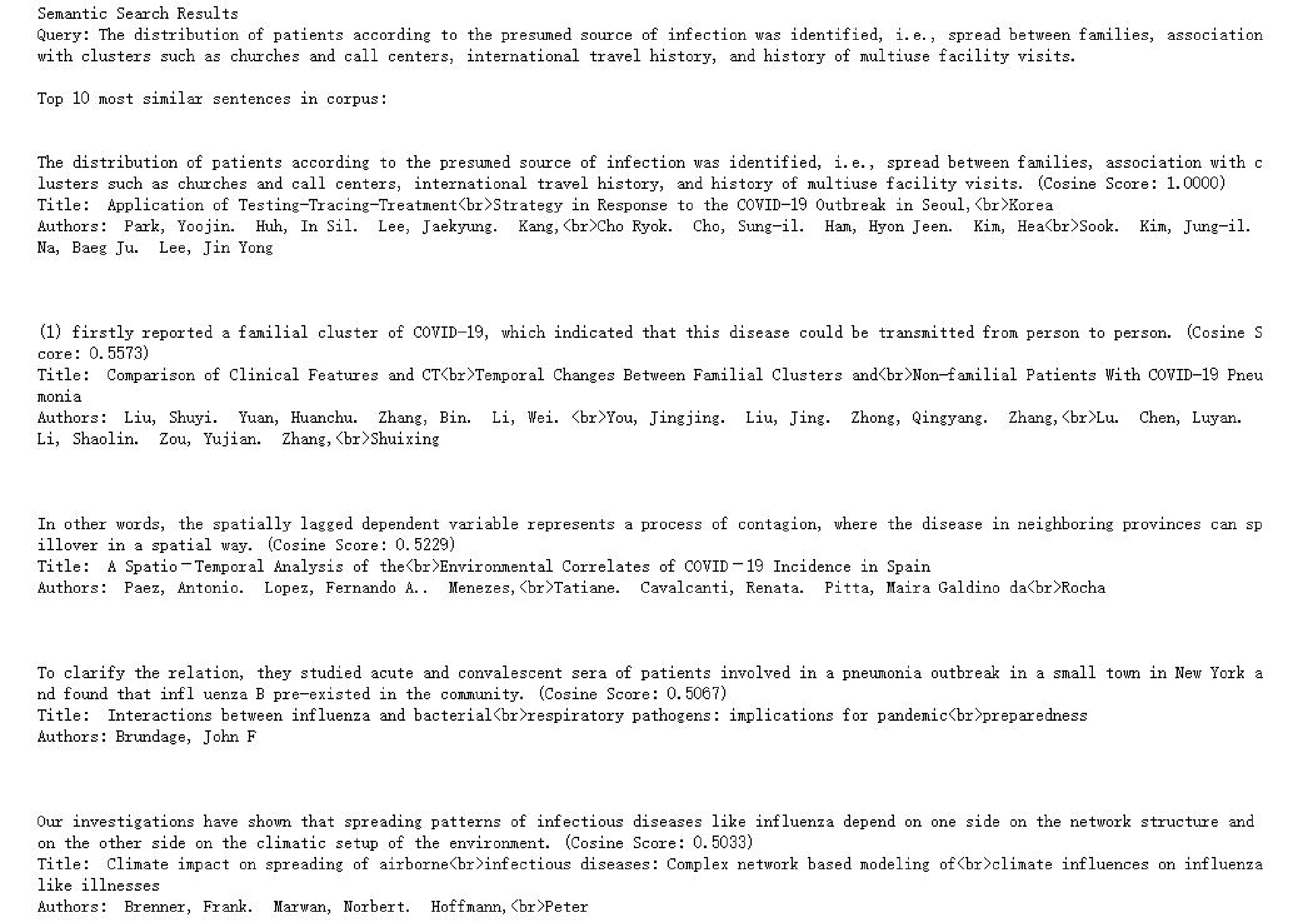
Using the 'Roberta large NLI stsb mean tokens' model in the sensetransformer library, it is convenient to calculate the sentences from the sentence resources. It takes about an hour to get a sentence vector of 10000 documents.

Compare cosine value and generate a sort

Using SciPy library, calculate the cosine value of the vector of literature sentence and query sentence vector. The sentence closest to the query sentence is obtained by comparing the query with the cosine value of each sentence, and the sentence and its source are returned.

Search tool

Search results



```
Semantic Search Results
Query: The distribution of patients according to the presumed source of infection was identified, i.e., spread between families, association
with clusters such as churches and call centers, international travel history, and history of multiuse facility visits.

Top 10 most similar sentences in corpus:


The distribution of patients according to the presumed source of infection was identified, i.e., spread between families, association with c
lusters such as churches and call centers, international travel history, and history of multiuse facility visits. (Cosine Score: 1.0000)
Title:  Application of Testing-Tracing-Treatment<br>Strategy in Response to the COVID-19 Outbreak in Seoul,<br>Korea
Authors:  Park, Yoojin.  Huh, In Sil.  Lee, Jaekyung.  Kang,<br>Cho Ryok.  Cho, Sung-il.  Ham, Hyon Jeen.  Kim, Hea<br>Sook.  Kim, Jung-il.
Na, Baeg Ju.  Lee, Jin Yong



(1) firstly reported a familial cluster of COVID-19, which indicated that this disease could be transmitted from person to person. (Cosine S
core: 0.5573)
Title:  Comparison of Clinical Features and CT<br>Temporal Changes Between Familial Clusters and<br>Non-familial Patients With COVID-19 Pneu
monia
Authors:  Liu, Shuyi.  Yuan, Huanchu.  Zhang, Bin.  Li, Wei. <br>You, Jingjing.  Liu, Jing.  Zhong, Qingyang.  Zhang,<br>Lu.  Chen, Luyan.
Li, Shaolin.  Zou, Yujian.  Zhang,<br>Shuixing



In other words, the spatially lagged dependent variable represents a process of contagion, where the disease in neighboring provinces can sp
illover in a spatial way. (Cosine Score: 0.5229)
Title:  A Spatio－Temporal Analysis of the<br>Environmental Correlates of COVID－19 Incidence in Spain
Authors:  Paez, Antonio.  Lopez, Fernando A..  Menezes,<br>Tatiane.  Cavalcanti, Renata.  Pitta, Maira Galdino da<br>Rocha



To clarify the relation, they studied acute and convalescent sera of patients involved in a pneumonia outbreak in a small town in New York a
nd found that infl uenza B pre-existed in the community. (Cosine Score: 0.5067)
Title:  Interactions between influenza and bacterial<br>respiratory pathogens: implications for pandemic<br>preparedness
Authors: Brundage, John F



Our investigations have shown that spreading patterns of infectious diseases like influenza depend on one side on the network structure and
on the other side on the climatic setup of the environment. (Cosine Score: 0.5033)
Title:  Climate impact on spreading of airborne<br>infectious diseases: Complex network based modeling of<br>climate influences on influenza
like illnesses
Authors:  Brenner, Frank.  Marwan, Norbert.  Hoffmann,<br>Peter
```

Thank You For Watching !