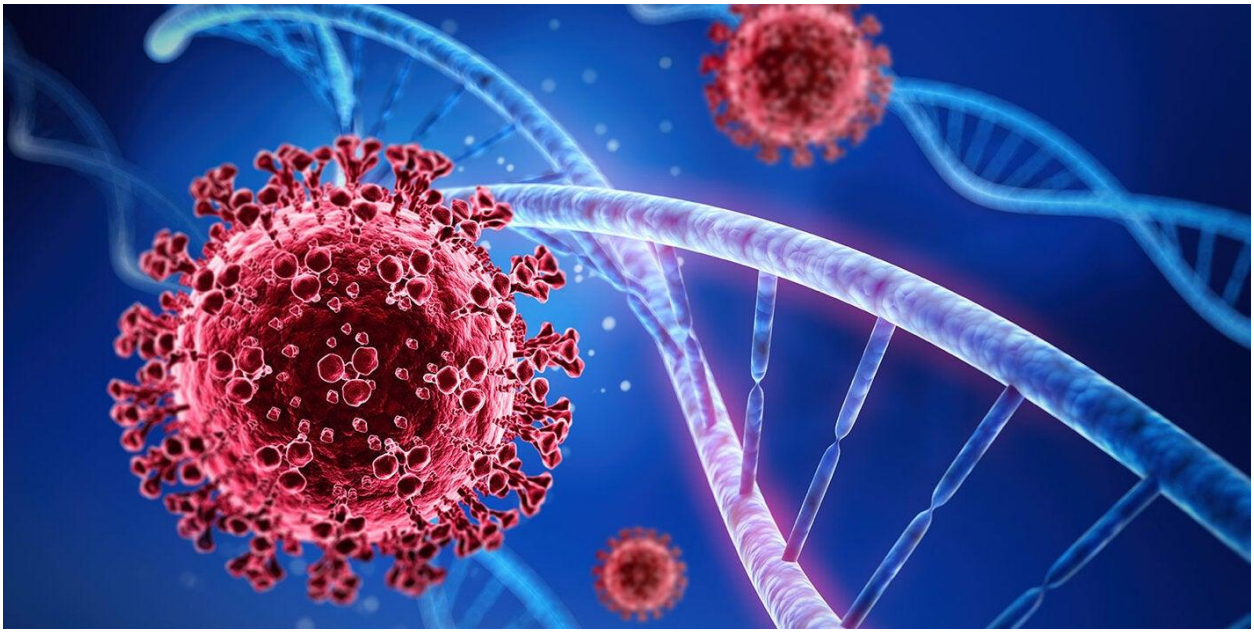
Clustering Analysis on Open Research Dataset CORD 19



Project Description:

- In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19).
- CORD-19 is a resource of over 57,000 scholarly articles, including over 45,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses.
- This freely available dataset is provided to the global research community. As a big data community, how can we help researchers to easily find the related research papers easily?

Project Objective:

- Given the large number of literature and the rapid spread of COVID-19, It is difficult for health professionals to keep up with the new information on the virus.
- The objective is to find out the best way to cluster the research papers then build a recommender system to receive the title of the research paper and recommend the most N similar papers to it based on its cluster.

Project Documentation:

1- Reading and Exploring the dataset

1.1- comm_use_subset, non_comm_use_subset and biorxiv_medrxiv data were loaded from databricks and merged into a single spark data farame.

1.2- bib_entries & ref_entries had to be dropped as data frames had them structured differently and they don't hold relevant info

2- Optimizing the Performance

2.1- Since parquet is much more optimized for spark operations, the dataframe was repartitioned and written as parquet.

3- Exploratory Data Analysis

3.1- Explore the Language:

3.1.1- Procedure:

- A language detection library was used to tag each record.
- Title and abstract are used for language detection. Access to full text may only provide marginal improvement.
- Some quality issues on the content of title and abstract fields were discovered.

3.1.2- Conclusions:

- Most articles are in English, with small proportions in French, Spanish, German or Italian
- A visual inspection of results shows that results are quite accurate for English even if the text are not long.
- Nevertheless, for those records that have no abstract and short text, English articles get tagged as other languages.
- There are only three articles in Chinese (abstract, title is in English)
- As English records are the vast majority, probably we should ignore other languages after cleaning empty records.

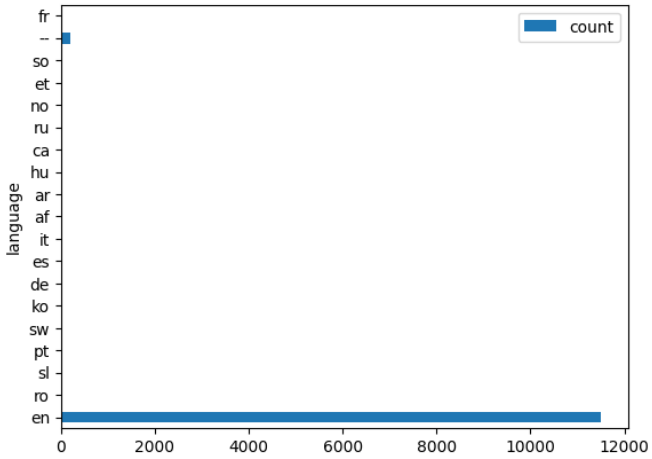


Fig.1 The bar chart shows the total number of the papers per language.

3.2- Explore empty values:

3.2.1- Essential data as title and abstract are missing in some rows, we need to handle that later.

3.2.2- Title, abstract and body are not plain text columns, they need specific handling to access them.

3.2.3- Non-essential data is missing in some rows as well, but that won't affect clustering.

3.3- Explore data size:

3.3.1- As expected, body is much larger than abstract.

3.4- Explore most frequent words:

3.4.1- Generated a WordCloud to visualize the most frequent words

Fig.2 The generated WordCloud.

3.4.2- Most frequent words seem legit (related to COVID-19) and nothing out of the ordinary.

3.4.3- Custom stop words should cover irrelevant and frequent words.

4- Preparation and Cleaning the data

4.1- Explore metadata:

4.1.1- Metadata has been first compressed using gzip format to be ready by spark.

4.1.2- Handling null ids and duplicates in metadata, ids with nulls were removed and duplicates were reduced.

4.1.3- Metadata was joined with the main data.

4.2- Handling Nulls:

4.2.1- Missing Titles and Abstracts were filled with metadata counterparts.

4.2.2- Abstract Nulls are handled in later steps by concatenating titles on them.

4.3- Keep only English Documents:

4.3.1- Only a tiny proportion of the documents are non-English.

5-Preprocessing

5.1- Merge Text columns in one column

5.2- Pre-stop words removal:

5.2.1- HTML Tags removal.

5.2.2- Convert Accented characters using unidecode

5.2.3- Expanding Contradictions

5.2.4- Removing Punctuations using this Regex '!()-[]{};:'",<>./?@#$%^&*_~'

5.2.5- Convert Text to Lower Case.

5.2.6- Treatment for Numbers, converting words to numeric forms.

5.2.7- Lemmatization, converting a word to its base form.

5.3- Stop words:

5.3.1- Removing default and custom stop words using spaCy's stopwords feature.

6-Vectorization

6.1- TF-IDF

6.1.1- TF-IDF will convert our string formatted data into a measure of how important each word is to the instance out of the literature as a whole.

6.1.2- CountVectorizer is used to convert the text to numerical indices

7-Principal Component Analysis

7.1- PCA is used to reduce the dimensions while still keeping 95% variance for better performance and hopefully remove some noise/outliers

7.2- Ran PCA initially with big k to cover almost whole variance range. This needs trial and error. Then, picked the k that would just cover 95% variance.

7.3- By trial and error we found that PCA can take no more than 250 entries.

8-Clustering

8.1- Applied K-means.

8.1.1- k defines the number of clusters and the seed defines the value used to set the cluster centers. A different value of seed for the same k will result in clusters being defined differently. In order to reproduce similar clusters when re-running the clustering algorithm use the same values of k and seed.

9-Evaluation

9.1- Used Elbow Method and Silhouette Method to pick the best k for training.

9.2- Re-trained the model with the value of the best k (k=5)

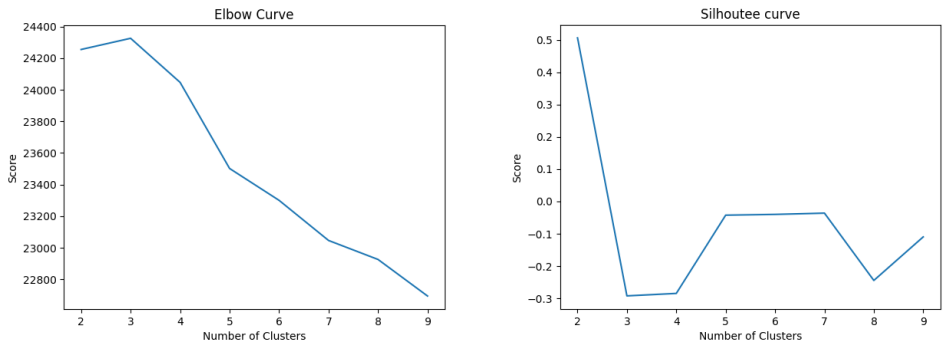


Fig.3 The output of the elbow and silhouette methods.

10-Recommender System

10.1- Create a function with the signature RecommendPaper (paper_title,N) where N is the number of recommended papers in the list and it returns the recommendation list.

10.1.1- Append test paper_title as a new row on the existing dataframe

10.1.2- Run all transformers and model on the appended dataframe

10.1.3- Filter results by predicted cluster, sort them with cosine distance and limit to N recommendations required

10.1.4- Return papers with same paper_id as filtered results

```
display(recommendPaper('Outlook of therapeutic and diagnostic competency of nanobodies against SARS-CoV-2', 5))
```

	metadata_title	doi	pmcid	pubmed_id	license	metadata_abstract
	Potential impact of seasonal forcing on a SARS-CoV-2 pandemic	10.1101/2020.02.13.20022806	null	null	medrxiv	A novel coronavirus (SARS-CoV-2) first detected in Wuhan, China, has spread rapidly since December 2019 80,000 confirmed infections and 2,700 fatalities (as of Feb 27, 2020). Imported cases and transmission clust have been reported globally suggesting a pandemic is likely. Here, we explore how seasonal variation in tran modulate a SARS-CoV-2 pandemic. Data from routine diagnostics show a strong and consistent seasonal v

```
display(recommendPaper('Covid tests and PCR', 3))
```

	metadata_title	doi	pmcid	pubmed_id	license	metadata_abstract
	Selection of a set of reliable reference genes for quantitative real-time PCR in normal equine skin and in equine sarcoids	10.1186/1472-6750-6-24	PMC1484482	16643647	cc-by	"BACKGROUND: Real-time quantitative PCR can be a very powerful and accurate technique to patterns in different biological conditions. One of the critical steps in comparing transcription profi most of the studies published on real-time PCR in horses, normalisation occurred against only or GAPDH or ACTB, without validation of its expression stability. This might result in unreliable conc

Fig.4 The output of the recommendation function for two different inputs for the paper titles.