Covid-19 Research Papers Text Mining

-N R SRIRANJANI (19BCT0165)

-GRAVIT ARORA (19BCE0422)

-TALHAR JANHAVI AJAY (19BCE0736)

Team Id 4

Professor : Rajeshkannan R

G1 slot

**1.ABSTRACT**

We present a text mining system on a corpus of scientific articles related to COVID-19. We build a similarity network on the articles where similarity is determined via shared citations and biological domain-specific sentence embedding’s

We uses tf-idf algorithm to find the similarity between the query and the research paper and get the relevant documents related to the query

**2.INTRODUCTION**

Since the discovery of the novel coronavirus SARS-CoV-2 toward the tail end of 2019, the disease caused by the virus, COVID-19, has swept through the globe and drastically altered all aspects of our lives. Governments and researchers, academic and industry alike, have coalesced around the common goals of healthcare resource management, social policy determination, prevention and treatment and vaccine development.

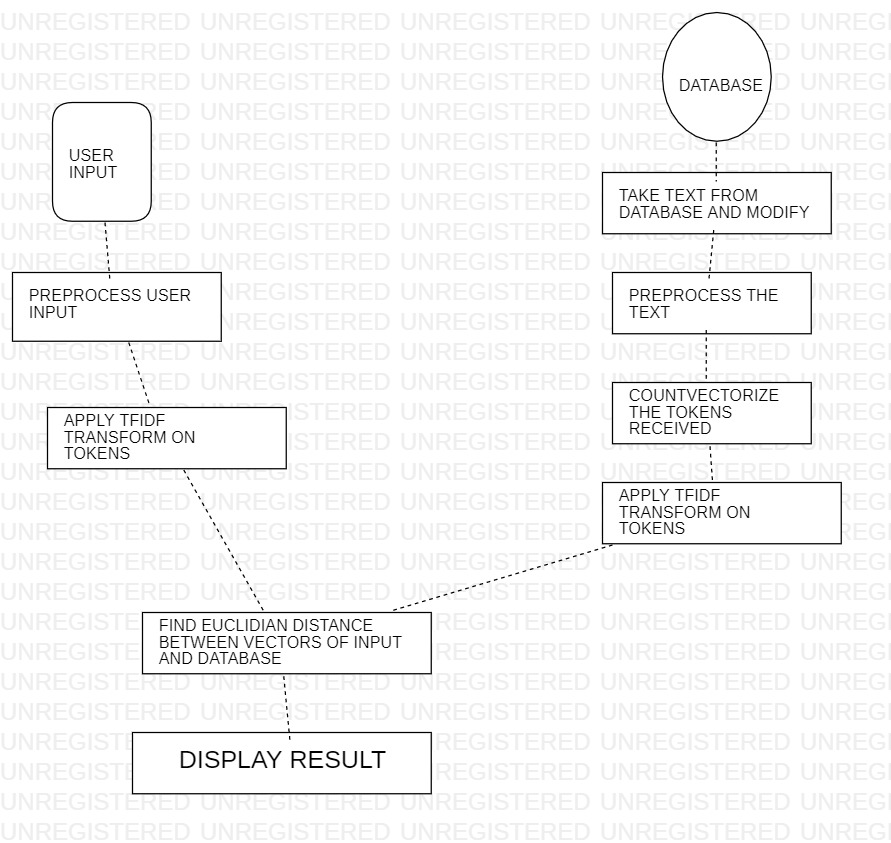
Text mining is the process of examining large collections of documents to discover new information or help answer specific research questions. The structured data created by text mining can be integrated into databases, data warehouses or business intelligence dashboards and used for descriptive, prescriptive or predictive analytics.

One of the major application areas of biomedical text mining is managing information overload. As per text mining focuses on solving specific problems such as retrieving relevant documents or extracting nuggets of information from those documents. In the process of addressing these problems, text mining systems may use techniques for information retrieval, information extraction, text classification, etc. and leverage methods from related fields such as natural language processing

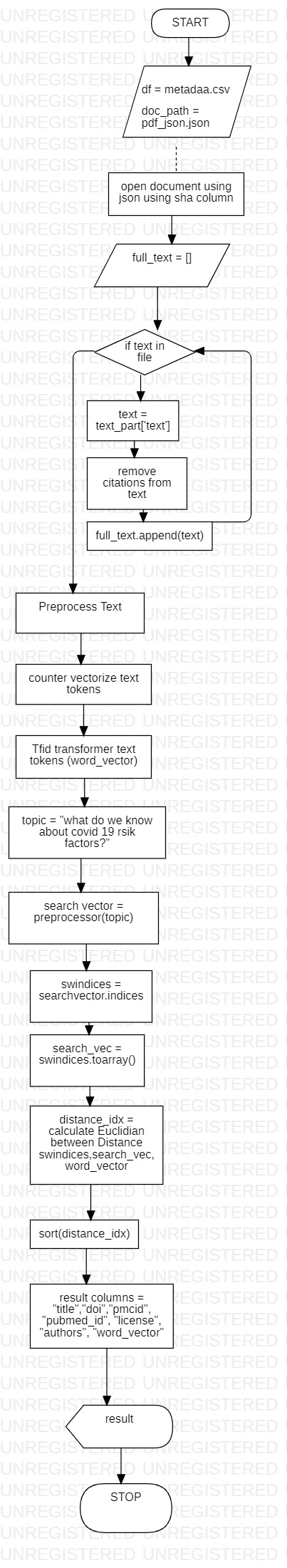
**3.PROBLEM STATEMENT**

Novel coronavirus (COVID-19) has resulted in a pandemic in a short span of time owing to its quick transmission. The incredible rate of scientific research on COVID 19 lead to information overload, making it difficult for researchers, clinicians and public health officials to keep up with the latest findings

**->ARCHITECTURE DIAGRAM**



**FLOW CHART:**



**->PSEUDO CODE**

1. def get\_text(sha): // to get the abstract

if sha == "":

return ""

document\_path = [x for x in doc\_paths if sha in x]

if not document\_path:

return ""

with open(document\_path[0]) as f:

file = json.load(f)

full\_text = []

#iterate over abstract and body part

for part in ['abstract', 'body\_text']

# iterate over each paragraph

for text\_part in file[part]

text = text\_part['text']

# remove citations from each paragraph

for citation in text\_part['cite\_spans']

text = text.replace(citation['text'], "")

full\_text.append(text)

return str.join(' ', full\_text)

2. stemmer = SnowballStemmer("english") //calling the snowballstemmer

analyzer = CountVectorizer().build\_analyzer() // handles preprocessing ,tokenization and n-gram genaration

3.def preprocess(doc): This function preprocesses the text document

doc=doc.lower()

return str.join(" ", [stemmer.stem(w) for w in analyzer(doc)])

def preprocess\_row(row)

text = str.join(' ', [str(row.title), str(row.abstract]) //to join the title abstract

return preprocess(text)

df.apply(preprocess\_row) // applies for whole dataset

5. cv = CountVectorizer(max\_df=0.95, stop\_words='english') // it calculates the tf-idf for each research paper

word\_count = cv.fit\_transform(df.preprocessed) Convert a collection of text documents to a matrix of token counts

tfidf\_tr = TfidfTransformer(smooth\_idf=True, use\_idf=True) Transform a count matrix to a normalized tf or tf-idf representation

tfidf\_tr.fit(word\_count)

6.def get\_word\_vector(document)

w\_vector = tfidf\_tr.transform(cv.transform([document]))

return w\_vector

df.apply(preprocessed text) // applies for whole dataset

7.def calculate\_distance\_between\_words\_vectors(search\_words\_indices, search\_vec, document\_vector) // calculates the distance between the query and papers using tf idf

document\_vec = document\_vector [0, search\_words\_indices].toarray()

returns distance.euclidean([search\_vec], document\_vec)

8. def get\_rel\_doc

search\_vector = get\_word\_vector(preprocess(topic))

search\_words\_indices = search\_vector.indices

search\_vec = search\_vector.data

distance\_idx = df.apply(lambda x: calculate\_distance\_between\_words\_vectors(search\_words\_indices, search\_vec, x.word\_vector), axis=1)

relevant\_indexes = distance\_idx.sort\_values().head(10).index // we sort the distances and get the titles of the papers

result\_columns = ["title", "doi", "pmcid", "license", "authors"]

result = df[result\_columns].iloc[relevant\_indexes].fillna("")

return result

9. def display\_friendly\_results(df\_result): Results is displayed

display\_columns = ["title", "doi", "pmcid", "authors"]

display(df\_result[display\_columns].reset\_index(drop=True))

Sample Input question: "What do we know about COVID-19 ?"

4. EXPERIMENT AND RESULT

Datasets:

[CORD-19: The Covid-19 Open Research Dataset - NCBI - NIH](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955/)

CORD-19 integrates papers from several sources ([Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955/figure/F1/)). Sources make openly accessible paper metadata, and in most cases, documents associated with each paper.

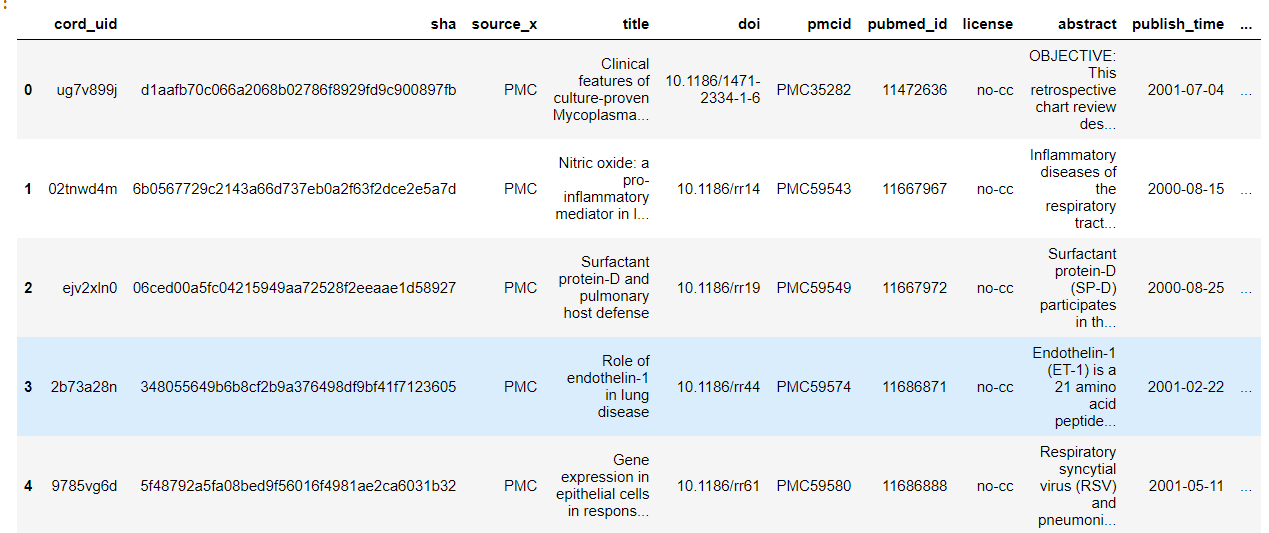
The CORD-19 effort combines paper metadata and documents from different sources, and generates harmonized and deduplicated metadata as well as structured full text parses of paper documents as output. We provide full text parses of all papers for which we have access to a paper document, and for which the documents are available under open access copyright licenses

CORD-19 has grown rapidly, now consisting of over 52K papers with over 41K full texts. The increase can be attributed to major publishers offering favorable terms on text/data mining uses that make the inclusion of their publications possible.

The resulting collection of sourced papers suffers from duplication and incomplete or conflicting metadata. We perform the following operations to harmonize and deduplication all metadata entries:

1. Cluster duplicate papers using identifiers
2. Select canonical metadata for each cluster
3. Filter clusters to remove non-papers

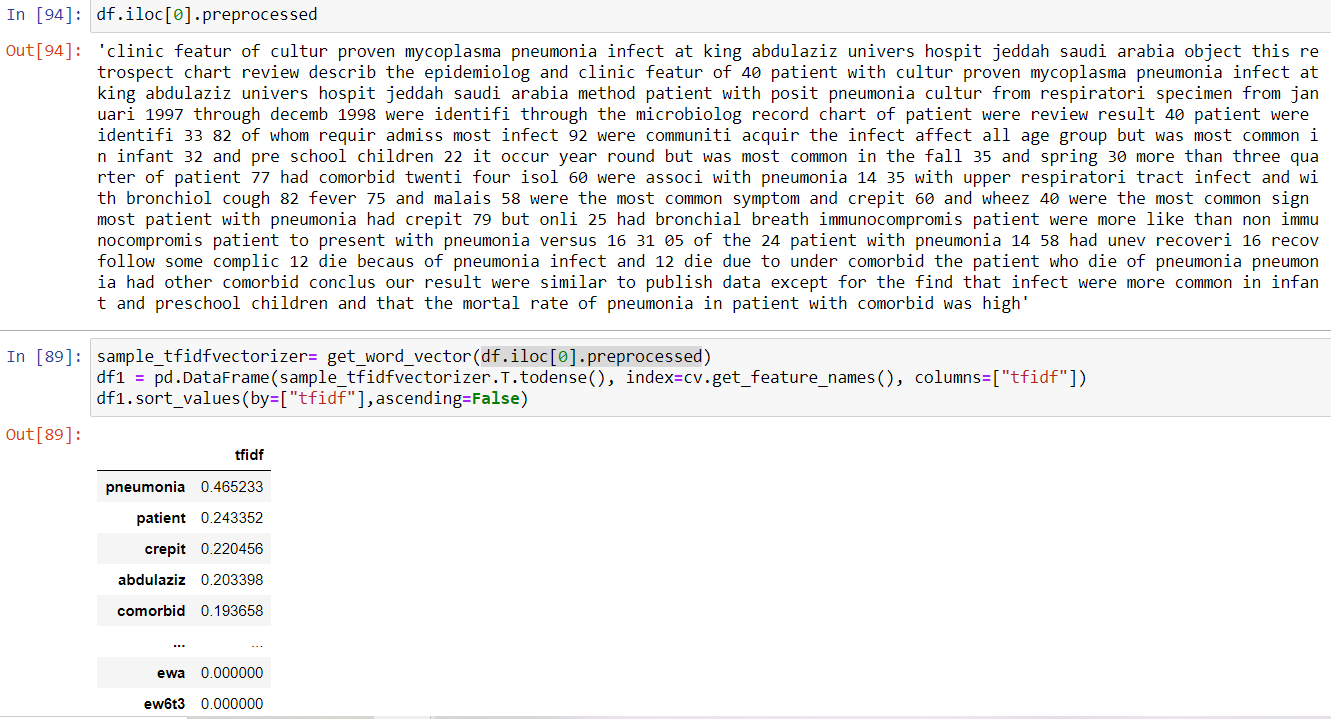
We start with approximately 73K metadata entries. After processing, the metadata consists of papers from PMC (28.6K), medRxiv (1.1K), and bioRxiv (0.8K), with another 1.1K from the WHO paper list and 19.5K contributed directly by publishers.

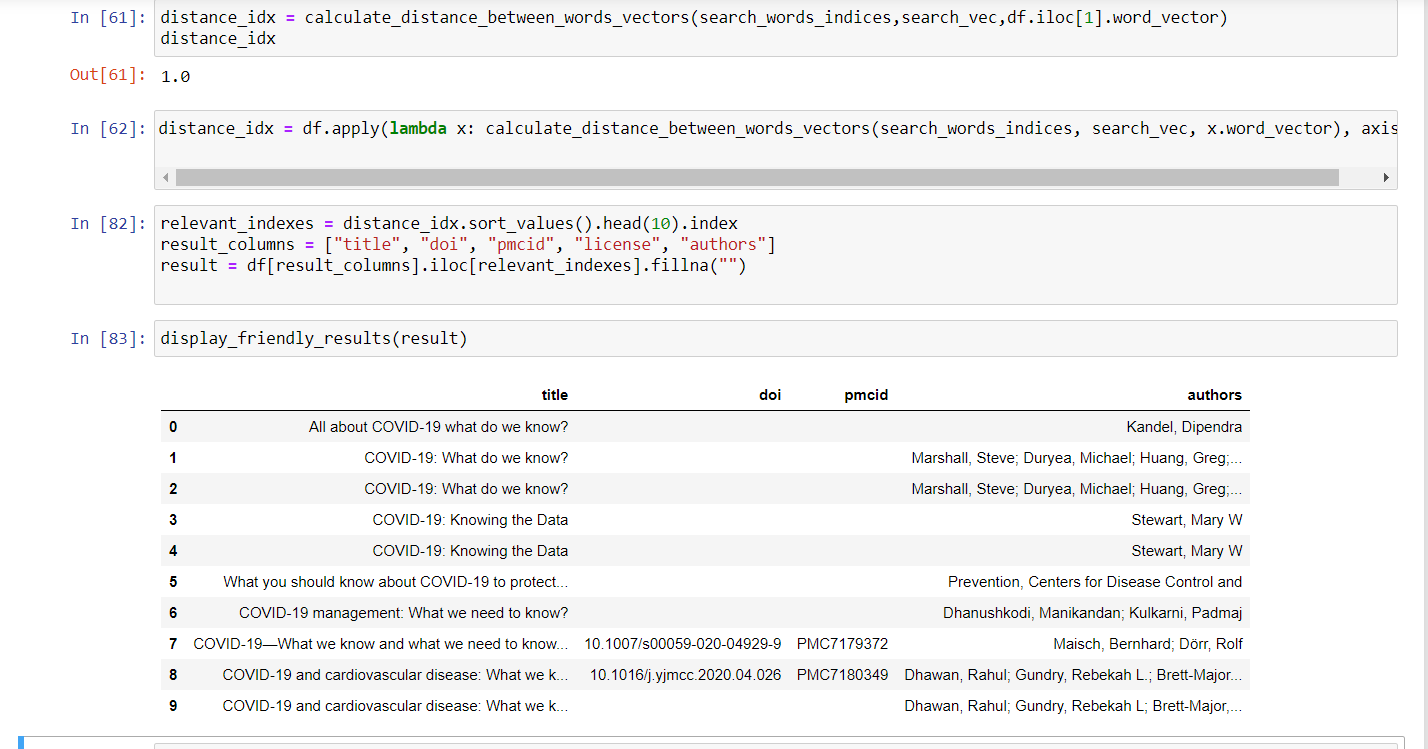


df shape =(522159, 22)

Sample Output:

It gets the relevant documents for query from the datasets.





5. CONCLUSION

Toward enhancing rigor and integrity of biomedical research, we proposed text mining as complementary to efforts focusing on standardization and guideline development.

Using tf-idf algorithm, we get a successful information retrieval system. That is the relevant research papers regarding the query is given as the output. This model will limit the time to search for relevant documents and useful to further creating a question answering bot.

6. REFERENCES

1. Debasmita Das, Yatin Katyal, Janu Verma Shashank Dubey, Aakash Deep Singh, Kushagra Agarwal, Sourojit Bhaduri, Rajesh Kumar Ranjan ,Information Retrieval and Extraction on COVID-19 Clinical Articles Using Graph Community Detection and Bio-BERT Embeddings
2. Ramzan Talib , Muhammad Kashif Hanif , Shaeela Ayesha, and Fakeeha Fatima ,Text Mining: Techniques, Applications and Issues
3. Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.
4. Kim, SW., Gil, JM. Research paper classification systems based on TF-IDF and LDA schemes. *Hum. Cent. Comput. Inf. Sci.* **9,**30 (2019).
5. Simon, C., Davidsen, K., Hansen, C. *et al.* BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* **19,**57 (2019).
6. Biomedical text mining for research rigor and integrity: tasks, challenges, directions, [Halil Kilicoglu](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kilicoglu%20H%5BAuthor%5D&cauthor=true&cauthor_uid=28633401)
7. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, Volume 36, Issue 4, 15 February 2020
8. Lucy Lu Wang, Kyle Lo, Text mining approaches for dealing with the rapidly expanding literature on COVID-19, Briefings in Bioinformatics, Volume 22, Issue 2, March 2021
9. Tracking and Mining the COVID-19 Research Literature
10. <https://kavita-ganesan.com/tfidftransformer-tfidfvectorizer-usage-differences/#.YHRQGugzZPZ>
11. <https://kavita-ganesan.com/how-to-use-countvectorizer/#.YHRQEegzZPa>
12. <https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089>
13. <https://scipy-lectures.org/advanced/scipy_sparse/csc_matrix.html>
14. <https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing>
15. John M Giorgi, Gary D Bader, Towards reliable named entity recognition in the biomedical domain, Bioinformatics, Volume 36, Issue 1, 1 January 2020