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.

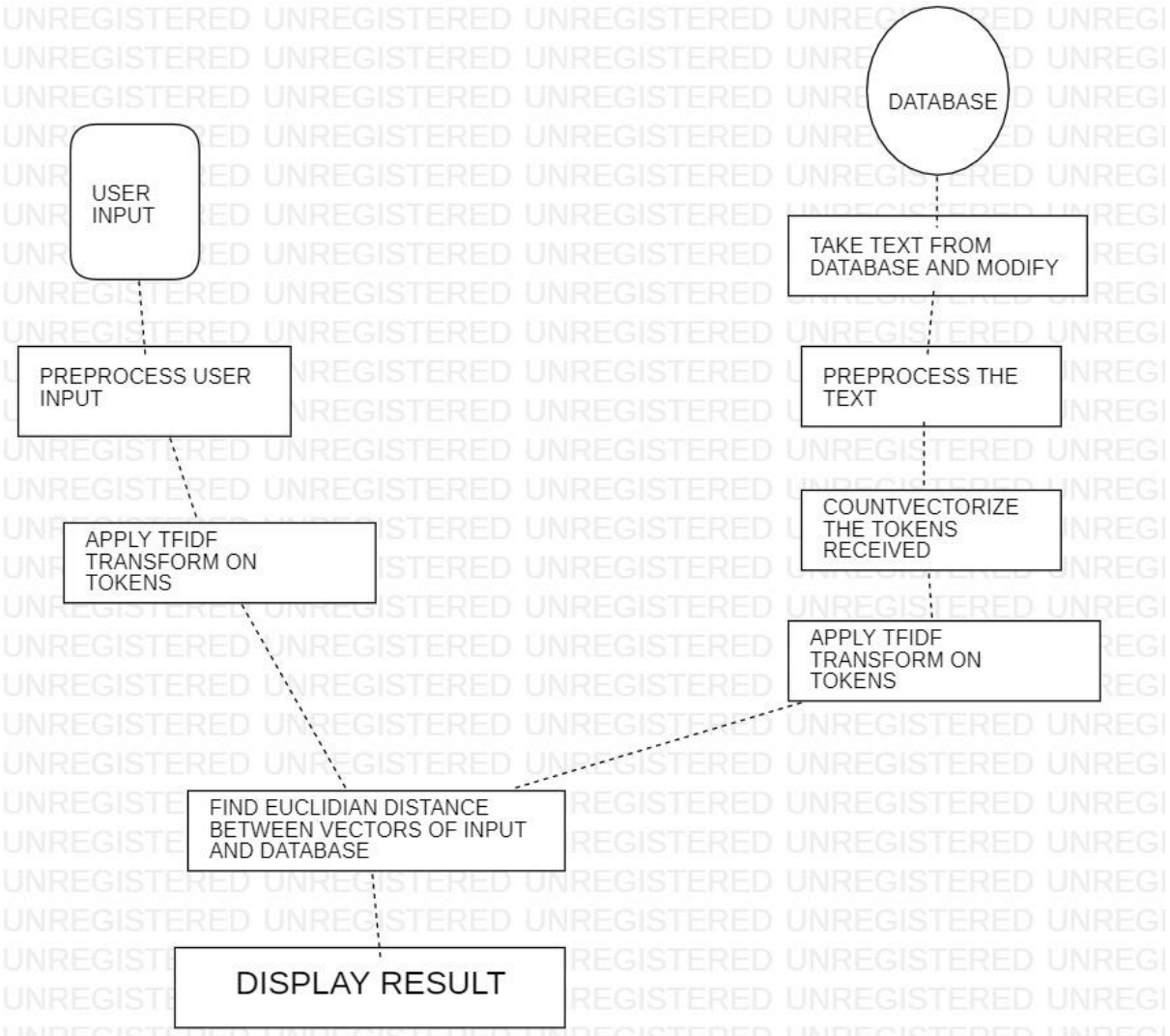
Text mining is the process of examining large collections of documents to discover new information or help answer specific research questions. The structured data created by text mining can be integrated into databases, data warehouses or business intelligence dashboards and used for descriptive, prescriptive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

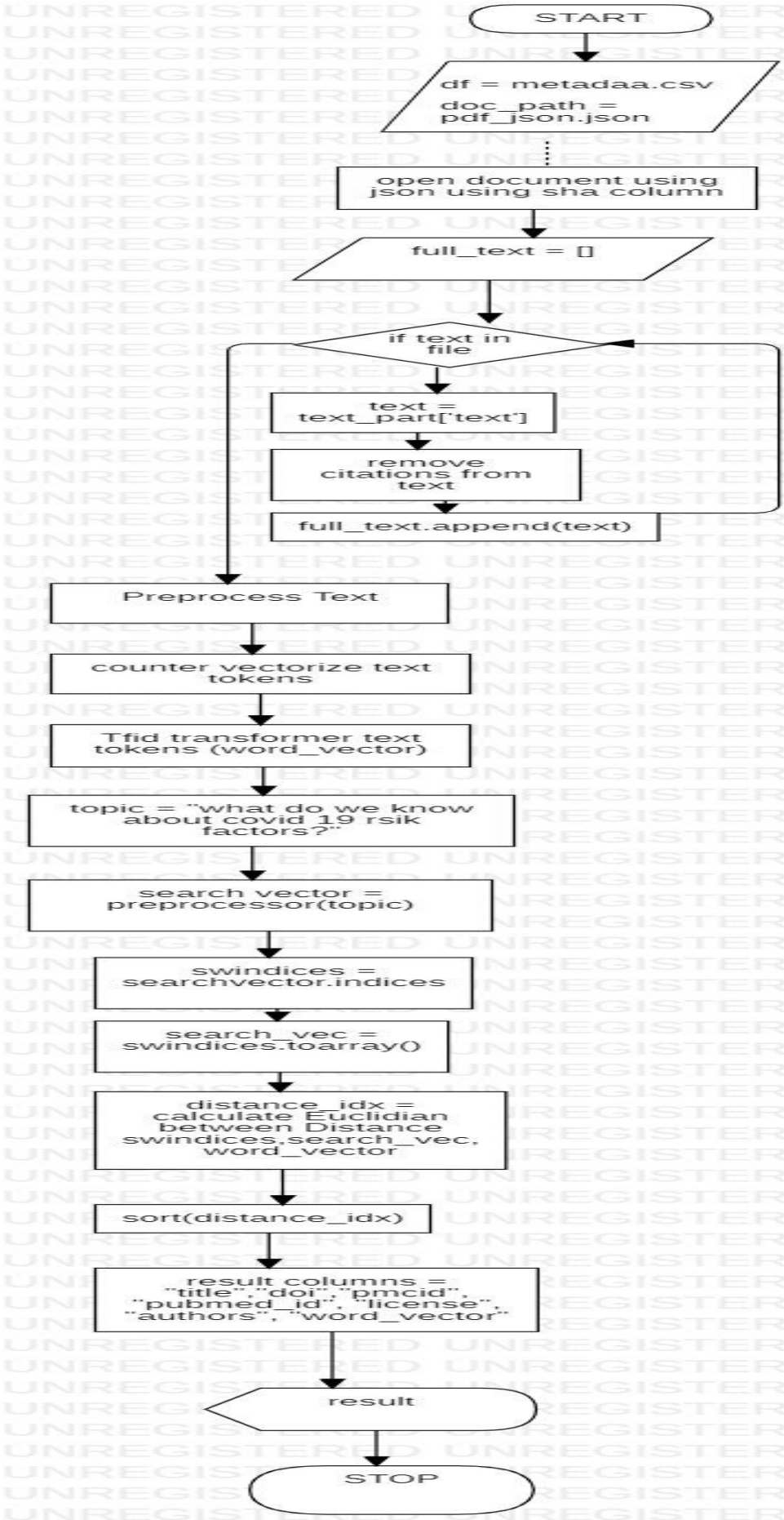
3.PROBLEM STATEMENT

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:


START
df = metadaa.csv
doc_path = pdf_json.json
open document using json using sha column
full_text = []
if text in file
text = text_part['text']
remove citations from text
full_text.append(text)
Preprocess Text
counter vectorize text tokens
Tfid transformer text tokens (word_vector)
topic = "what do we know about covid 19 rsik factors?"
search vector = preprocessor(topic)
swindices = searchvector.indices
search_vec = swindices.toarray()
distance_idx = calculate Euclidian between Distance swindices,search_vec, word_vector
sort(distance_idx)
result columns = "title","doi","pmcid", "pubmed_id", "license", "authors", "word_vector"
result
STOP

->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

CORD-19 integrates papers from several sources (Figure 1). 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.

:

	cord_uid	sha	source_x	title	doi	pmcid	pubmed_id	license	abstract	publish_time	...
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	...
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	...
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	...
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	...
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	...

df shape =(522159, 22)

Sample Output:

It gets the relevant documents for query from the datasets.

In [94]:
```
df.iloc[0].preprocessed
```

Out[94]:
```
'clinic featur of cultur proven mycoplasma pneumonia infect at king abdulaziz univers hospit jeddah saudi arabia object this re
trospect chart review describ the epidemiolog and clinic featur of 40 patient with cultur proven mycoplasma pneumonia infect at
king abdulaziz univers hospit jeddah saudi arabia method patient with posit pneumonia cultur from respiratori specimen from jan
uari 1997 through decemb 1998 were identifi through the microbiolog record chart of patient were review result 40 patient were
identifi 33 82 of whom requir admiss most infect 92 were communiti acquir the infect affect all age group but was most common i
n infant 32 and pre school children 22 it occur year round but was most common in the fall 35 and spring 30 more than three qua
rter of patient 77 had comorbid twenti four isol 60 were associ with pneumonia 14 35 with upper respiratori tract infect and wi
th bronchiol cough 82 fever 75 and malais 58 were the most common symptom and crepit 60 and wheez 40 were the most common sign
most patient with pneumonia had crepit 79 but onli 25 had bronchial breath immunocompromis patient were more like than non immu
nocompromis patient to present with pneumonia versus 16 31 05 of the 24 patient with pneumonia 14 58 had unev recoveri 16 recov
follow some complic 12 die becaus of pneumonia infect and 12 die due to under comorbid the patient who die of pneumonia pneumon
ia had other comorbid conclus our result were similar to publish data except for the find that infect were more common in infan
t and preschool children and that the mortal rate of pneumonia in patient with comorbid was high'
```

In [89]:
```
sample_tfidfvectorizer= get_word_vector(df.iloc[0].preprocessed)
df1 = pd.DataFrame(sample_tfidfvectorizer.T.todense(), index=cv.get_feature_names(), columns=["tfidf"])
df1.sort_values(by=["tfidf"],ascending=False)
```

Out[89]:

	tfidf
pneumonia	0.465233
patient	0.243352
crepit	0.220456
abdulaziz	0.203398
comorbid	0.193658
...	...
ewa	0.000000
ew6t3	0.000000

In [61]:
```
distance_idx = calculate_distance_between_words_vectors(search_words_indices,search_vec,df.iloc[1].word_vector)
distance_idx
```

Out[61]:
```
1.0
```

In [62]:
```
distance_idx = df.apply(lambda x: calculate_distance_between_words_vectors(search_words_indices, search_vec, x.word_vector), axis
```

In [82]:
```
relevant_indexes = distance_idx.sort_values().head(10).index
result_columns = ["title", "doi", "pmcid", "license", "authors"]
result = df[result_columns].iloc[relevant_indexes].fillna("")
```

In [83]:
```
display_friendly_results(result)
```

	title	doi	pmcid	authors
0	All about COVID-19 what do we know?			Kandel, Dipendra
1	COVID-19: What do we know?			Marshall, Steve; Duryea, Michael; Huang, Greg;...
2	COVID-19: What do we know?			Marshall, Steve; Duryea, Michael; Huang, Greg;...
3	COVID-19: Knowing the Data			Stewart, Mary W
4	COVID-19: Knowing the Data			Stewart, Mary W
5	What you should know about COVID-19 to protect...			Prevention, Centers for Disease Control and
6	COVID-19 management: What we need to know?			Dhanushkodi, Manikandan; Kulkarni, Padmaj
7	COVID-19—What we know and what we need to know...	10.1007/s00059-020-04929-9	PMC7179372	Maisch, Bernhard; Dörr, Rolf
8	COVID-19 and cardiovascular disease: What we k...	10.1016/j.yjmcc.2020.04.026	PMC7180349	Dhawan, Rahul; Gundry, Rebekah L.; Brett-Major...
9	COVID-19 and cardiovascular disease: What we k...			Dhawan, Rahul; Gundry, Rebekah L; Brett-Major,...

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
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