The Pennsylvania State University
The Graduate School

S_COVID: A FRUGAL YET EFFECTIVE ENGINE TO EXPLORE COVID-19 SCIENTIFIC LITERATURE

A Thesis in
Computer Science and Engineering
by
Raj Ratn Pranesh

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2023

The thesis of Raj Ratn Pranesh was reviewed and approved by the following:

Rui Zhang
Assistant Professor of Computer Science and Engineering
Thesis Advisor

Abhinav Verma
Assistant Professor of Computer Science and Engineering

Chitaranjan Das
Professor of Computer Science and Engineering
Head of the Department of Computer Science and Engineering

Abstract

This paper presents the S_Covid, a cutting-edge end-to-end unsupervised learning-based question-answering engine. This novel system has been designed to aid in the exploration of vast COVID-19 scientific literature collections, which can often be overwhelming and difficult to navigate. Utilizing advanced unsupervised machine learning techniques, S_Covid is an innovative, yet user-friendly system that empowers researchers and scientists to search for specific research literature containing valuable information to answer complex user queries.

The engine functions by strategically isolating and extracting relevant sentences from research papers, which may provide viable answers to complex COVID-19-related user queries. The system accomplishes this feat without being overly reliant on exhaustive computation resources, making it both practical and accessible for researchers worldwide. We conducted a series of rigorous experiments on a vast collection of 80,000 COVID-19-related papers to demonstrate the system's performance. These experiments yielded promising results, which were corroborated by statistical analyses and feedback from real users. We also conducted a comparative analysis of S_Covid with existing search engines designed for information retrieval of COVID-19 scientific literature.

Overall, S_Covid is a powerful and effective information retrieval system that offers significant value to researchers and scientists who are exploring COVID-19 scientific literature. Its advanced unsupervised machine learning techniques and user-friendly design make it a valuable addition to the scientific community's arsenal of tools for exploring COVID-19 research literature.

Table of Contents

List of Figures vi

List of Tables vii

Acknowledgments viii

Chapter 1
Introduction 1

Chapter 2
Related Work 3

Chapter 3
S_Covid Methodology 5
3.1 Exploring COVID-19 Scientific Literature with S_Covid 5
3.1.1 Phase-A: Building the S_Covid corpus 5
3.1.1.1 **Step-1 Generating a vocabulary** 6
3.1.1.2 **Step-2 Topic modeling** 6
3.1.2 Phase-B: Exploring COVID-19 scientific articles 7
3.1.2.1 **Step-3 Extracting top relevant papers for a query** 7
3.1.3 Phase-C: Extracting candidate answers and ranking relevant papers 8
3.1.3.1 **Step 1: Extracting candidate answer sentences** . . 9
3.1.3.2 **Step 2: Ranking relevant papers** 9
3.2 S_Covid Algorithm 11

Chapter 4
Experiments 13
4.1 Experiment settings 13
4.2 **Dataset** 14
4.2.1 **Data pre-processing** 14

Chapter 5
Performance Analysis of S_Covid: A Comparative Study 18
5.1 Existing Tools and Baseline Models Analysis 18

5.1.1 Search Engines . . . 18
5.1.1.0.1 **COVIDEX**: . . . 18
5.1.1.0.2 **KDCOVID**: . . . 18
5.1.1.0.3 **COVID-19 Research Explorer (Google)**: . . 19
5.1.2 Baseline Models . . . 19
5.1.2.0.1 **LDA**: . . . 19
5.1.2.0.2 **LDA-BM25**: . . . 20
5.1.2.0.3 **LDA-WHOOSH:** . . . 20
5.1.3 Experimental comparison . . . 21
5.1.4 Result and Discussion . . . 21
5.2 An additional error analysis method . . . 22
5.2.1 **Comparative survey of S_Covid with baseline models**: . . 24
5.2.1.1 **Experimental comparison** . . . 24
5.2.2 Results and Discussion . . . 25
5.3 S_COVID features . . . 27

Chapter 6
Conclusions **31**
6.1 Future Work . . . 31

Bibliography **33**

List of Figures

3.1 Dominating set reduction with k = 3 and one isolated vertex.. 5

5.1 Screenshot of our S_COVID application, which builds on LDA and our rankings algorithm. 20

5.2 Question wise comparative evaluation of S_Covid against baseline models based on *Evaluation_score* (refer 5.2). 24

5.3 Overall performance score obtained by baseline models and S_Covid . . 26

List of Tables

4.1 COVID-19 Tasks 17

4.2 COVID-19 Questions 17

5.1 Evaluation result 19

5.2 S_Covid answers for COVID-19 related questions 29

5.3 Related sentences extracted by S_Covid for the questions in Table 4.2 . . 30

Acknowledgments

I extend my most heartfelt appreciation to my esteemed advisor, Dr. Rui Zhang, for his sagacious guidance in illuminating my interests in this particular realm of research and for his unwavering support and nurturance throughout my academic pursuit. His invaluable encouragement and sage advice were instrumental to the successful culmination of this research endeavor and thesis. Our every encounter served as a wellspring of inspiration and motivation, spurring me to persist in my scholarly pursuits. Engaging in dialogues about new areas of research and novel ideas with him was an intellectually stimulating experience that significantly enriched the quality of my thesis. His continued stewardship and counsel facilitated my advancement in my research endeavors.

I am fervently thankful to my esteemed committee member, Dr. Abhinav Verma, who is a professor of Electrical Engineering and Computer Science at Pennsylvania State University. His seminal course on machine learning kindled within me the spark for applying artificial intelligence and natural language processing concepts in my thesis. The valuable discussions we held regarding the progression and future prospects of my research study proved to be pivotal.

Finally, yet importantly, I must extend my sincere gratitude to my dear friends, my parents, and my sibling, who provided me with steadfast support and encouragement throughout the course of my studies and research. I am indelibly grateful to my parents and sibling for instilling in me the courage to take calculated risks and pursue my aspirations. Your unconditional love and affection are truly appreciated. I would also like to thank my co-authors for working with me and giving me consent to include our paper, "S_COVID: Frugal yet Effective Question-Answering Engine for COVID-19 Queries," as a part of my Master's thesis. Thank you from the bottom of my heart.

Chapter 1
Introduction

The COVID-19 pandemic has necessitated significant research and development efforts. To develop methods and innovative ways to study the virus and its resulting illness, scientists are constantly experimenting and consequently publishing their work. With the increasing volume of research documents regarding COVID-19 daily, it has become a challenge for scientists to keep updated on relevant new scientific literature. Search engines like Google are a powerful medium for an all-encompassing query; however, they have limitations when it comes to understanding the meaning behind research-driven questions that aim to explore document collections as they simply obtain one-shot batch answers. As a result, researchers have developed various deep learning-based information retrieval systems for answering COVID-19-related queries by exploring and filtering through thousands of COVID-19 scientific literature. Despite the success of the deep-learning model, it typically requires a perturbing use of computational resources to build the system along with the back-end infrastructure needed to successfully deploy for commercial usage. A recent study [1] has shown that simple systems without any complex machine learning components have outperformed deep-learning-based commercial COVID-19 information extraction systems on various evaluation matrices.

Therefore, in this paper, we propose S_COVID, a frugal yet effective end-to-end unsupervised learning-based interactive framework for COVID-19 scientific literature search and information retrieval. S_COVID understands a user query and adopts an exploration approach for finding relevant research literature that contains information that can most closely answer the query by a successive refinement process.

Our model utilizes the COVID-19 Open Research Dataset (CORD-19 [2]), consisting of over 50,000 scholarly articles related to COVID-19 and research related to coronavirus.

S_COVID refines and reranks the documents in the CORD-19 dataset successively to extract candidate results set by understanding the semantic meaning of the user query before presenting the final results. The S_COVID pipeline is divided into two major components: (a) data exploration and relevant paper extraction given a query; and (b) retrieving and ranking candidate sentences from papers that may answer the user query, thereby giving a final ranking of the papers.

With S_COVID, when the user asks an initial query, the tool not only returns a set of papers (like in a traditional search engine) but also returns sentences from the paper that are possible answers to the question. The user can review the sentences and decide whether or not the paper is pertinent or worthwhile for further reading. The overarching goal of S_COVID is to aid scientists with fast and efficient information retrieval and exploration of documents regarding COVID-19. Scientists can thereby conveniently find relevant information and answers for complex COVID-19-related queries, hence saving significant time that was previously spent on exploring and filtering scientific papers.

To evaluate S_COVID and ascertain that a simple framework can provide effective results, we worked with biologists and physicians from the Golestan University of Medical Sciences. They tested our model on a collection of Kaggle tasks based on scientific questions developed with the World Health Organization (WHO) and the National Academies of Sciences, Engineering, and Medicine (NASEM).

S_COVID has a simple and interactive web application interface with a front-end that was built using Javascript, HTML and CSS. Python Flask was used for building customized APIs that send user requests to the model and send the model's output back to the web application. The model is deployed on the Google Cloud Platform with required dependencies for supporting back-end functions and storing data.

Chapter 2
Related Work

This section summarises research work in information retrieval and question-answering proposal devoted to searching and exploring COVID-19 scientific literature.

Information retrieval is the most researched area when it comes to using NLP techniques to build a model using CORD-19 dataset. In [3], the author proposes a neural search and ranking engine for COVID-19 literature exploration based on T5 [4] language model fine-tuned on the medical text. It leverages the BM25 algorithm for ranking the documents, followed by reranking using T5 model. The model also highlights the most relevant sentences of each research paper using the pre-trained BioBERT [5] unsupervised model.

A number of research organisations have launched online web applications to provide a platform to the scientist and research to explore and understand COVID-19 research literature. The COVIDSCHOLAR[1] is a web application that adapts the MATSCHOLAR [6] system for exploring and searching relevant COVID-19 research papers based on entity-centric queries.

The KDCOVID[2] also presents a similar document searching framework which uses BioSentVec [7] to encode the query sentences and scientific literature followed by KNN search to find the relevant papers. Papers ranking was done based on the similarity score of the query vector and the document sentences vector. The key sentences were also highlighted in the paper. This also uses DisGeNET [8] to represent the relation between genes and disease in the form of a knowledge graph.

[1] https://covidscholar.orghttps://covidscholar.org
[2] http://kdcovid.nl/about.htmlhttp://kdcovid.nl/about.html

Plenty of work has adopted question-answering approaches based on the CORD-19 dataset. In [9], the authors presented a COVID-19 specific question-answering dataset built using CORD-19 [2]. The paper also presents a performance analysis of various language models for question-answering the same dataset.

The Google AI team also developed an NLU-powered tool to explore COVID-19 research papers- COVID-19 Research Explorer [3]. It is a framework for searching and ranking relevant papers. It also helps users to find the portion of each research paper related to the query. COVIDASK[4] and AUEB system[5] also present a question-answering framework that present user answer snippets for their query.

Existing work has some limitations. For example, in [9] the dataset was small in size. The performance of the language model shows that a larger question-answer dataset is necessary for training a supervised model. The paper [3] reports the use of BioBERT to generate a vector of words present in sentences. Since BioBERT's vocabulary might not have those words used within the COVID-19 papers, the contextual understanding and the word/sentence vector representation quality can decrease. In our proposal, we addressed this problem by training a COVID-19 specific `word2vec` model using CORD-19 dataset. The `word2vec` model generates a more representative contextual and semantic vector representation of all the COVID-19 terms and words present in the dataset. Also, the large transformer based information retrieval models such as [3] and Google's COVID-19 Research Explorer, tends to have a higher computational resource requirement as compared to S_covid which have comparatively simpler architecture design.

[3]https://covid19-research-explorer.appspot.comhttps://covid19-research-explorer.appspot.com
[4]https://covidask.korea.ac.krhttps://covidask.korea.ac.kr
[5]http://cslab241.cs.aueb.gr:5000http://cslab241.cs.aueb.gr:5000

Chapter 3
S_Covid Methodology

3.1 Exploring COVID-19 Scientific Literature with S_Covid

This section describes the approach for exploring scientific literature proposed by S_Covid. The main contribution is an end-to-end unsupervised COVID-19 scientific literature exploration process with a question-answering protocol to assist scientists going through query results about COVID-19. **Algorithm** 1 presents the pseudo-code of the S_Covid algorithm.

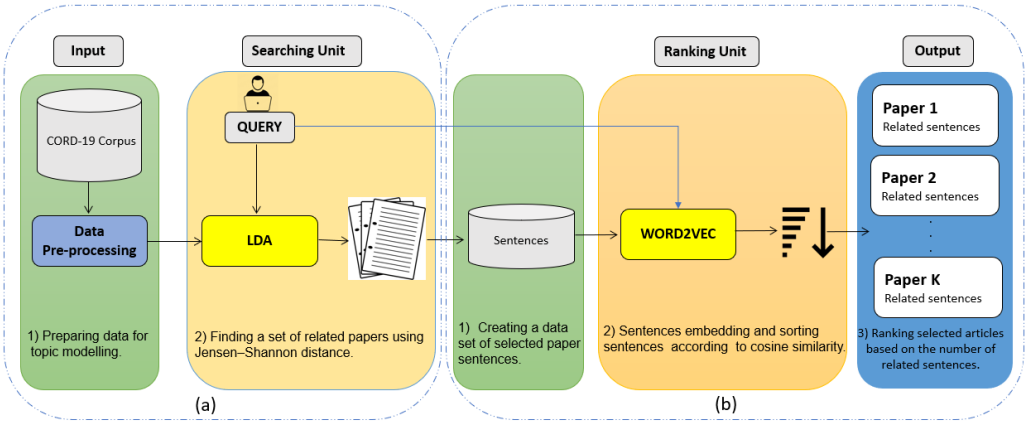


Figure 3.1: Dominating set reduction with k = 3 and one isolated vertex..

3.1.1 Phase-A: Building the S_Covid corpus

S_Covid first processes and indexes the paper collection to discover a set of topics (step-1) and then assigns relevant topics to the paper according to their content (step-2).

3.1.1.1 Step-1 Generating a vocabulary

The COVID-19 dataset consists of a large number of scientific papers containing text in the form of words, sentences, and paragraphs. To explore the documents (i.e., papers), it is necessary to generate a model of their content. In our approach, this model is a vocabulary. Vocabulary is particularly useful in the case of COVID-19 because there is no official vocabulary about this topic at the time of writing.
We first pre-processed documents to generate vector representations for words (i.e., word embedding). The word embedding representation most effectively captures words' contextual, lexical, and sentimental characteristics.
Once texts were pre-processed, we used the `Gensim Phrases` Python library [10] to automatically detect common phrases (bigrams) from each paper in the corpus. For example, sentences like 'infectious disease' or 'public health' must occur together. We applied the `skip-gram` method to predict the context of words. The principle of the method `skip-gram` is using the word to predict its target context.

We trained a `Word2vec` model, [11] which is a two-layer neural net that processes text by "vectorizing" the words in the document to build a vocabulary using the Gensim [10] python library and applying the `skip-gram` method. Then, we set up parameters and built the vocabulary to train the Word2vec model. The underlying assumption of `Word2Vec` is that two words sharing similar contexts also share a similar meaning and vector representation in the model. We set the dimensionality of the feature vectors to 300 and trained the model for 15 epochs. Finally, the resulting `Word2Vec` model was stored for future use in the document exploration pipeline.

3.1.1.2 Step-2 Topic modeling

The COVID-19 dataset used in our model is an unstructured text corpus. This characteristic made it difficult to extract relevant and desired information from the dataset. To tackle this challenge, we assume that large text documents can be understood by their overarching topics. The statistical process of learning and extracting topics from document collections is called topic modeling [12] [1].
In the S_Covid pipeline, we used the topic modeling method called Latent Dirichlet allocation (LDA) [13]. It is a Bayesian model for classifying discrete data that contains

[1]Through topic modeling, one can identify the topics that best describe a set of documents. Knowing the topics representing the content of documents can be useful for search engines and customer service automation.

uncorrelated topics. Through topic modeling, we represented each scientific document in the corpus as a distribution of topics described as a distribution of words. As an unsupervised machine learning method, LDA automatically analyzes a text for clustering the words from a given set of documents.
The prime objective of topic modeling is to extract latent semantic topics from large volumes of textual documents (i.e., corpora). To understand the topics of articles, we used Latent Dirichlet Allocation (LDA) [13], which is a well-known topic modeling technique, to generate topics from large amounts of textual data. Topic modeling helps understand and summarize large collections of textual information We used a trained a LDA model for topic modeling on the given dataset.
The number of topics has a crucial role in determining the performance of the model computed with LDA. After several iterations, we discovered that 50 is the optimum number of latent semantic topics that can be extracted from the COVID-19 literature corpus. For producing fine-grained results, the discovered topics were specified and refined with minimum overlapping. Then, with a set of topics assigned to each paper (i.e., a distribution over words), we organized the papers in an LDA space, namely a simplex. The dimensionality of the space depends on the number of topics i.e. 50 with each topic assigned a corner in the LDA space. Depending upon a topic's word distribution in a paper, each paper in the corpus is closer to the topics that represent it more strongly.

3.1.2 Phase-B: Exploring COVID-19 scientific articles

The objective of this phase is to extract the relevant papers that possibly contain the answer to a user query about COVID-19 from a large COVID-19 scientific literature corpus.
A query sentence (consisting of words/terms) is assigned with relevant topics and represented as a query vector. The semantic of a query is: *which papers are closest to these topics?* A ranked result set is computed by determining which papers of the corpus contain the topics that are closest to those of the query and ordering them.

3.1.2.1 Step-3 Extracting top relevant papers for a query

Given a user query regarding COVID-19, the engine S_Covid retrieves the set of ranked most relevant candidate papers that would most possibly answer the query.
For this, we first calculate the probability distributions of 50 topics over a given scientific literature corpus "P_p". The probability "P_p" represents how well each topic describes the

content of the paper. We also calculate the probability distribution of topics over the user query "P_q".
Then, we calculate the similarity score "S" to measure the similarity between two probability distributions i.e., "P_p" and "P_q" representing the paper and query probability vectors, respectively. The similarity score is given as **S = 1 - Jensen–Shannon distance**. Jensen–Shannon distance is the square root of the Jensen-Shannon divergence, which is a method of measuring the similarity between two probability distributions.

$$S = 1 - \sqrt{\frac{D(P_p \parallel m) + D(P_q \parallel m)))}{2}} \tag{3.1}$$

In the equation, "m" represents the mean of "P_p" and "P_q", D represents Kullback-Leibler divergence, and "S" represents the similarity score between a user query and the research paper in the corpus. The range of "S" lies between 0 to 1. The value of the similarity score determines the similarity between the topic distribution of a paper and a query. Consequently, the probability of a paper represents its relevance concerning the query.
Using the similarity score, we select the top relevant candidate papers to build a `related articles set`. This set contains papers that are the closest to the user query. We experimented with various values of similarity scores ranging from 0.2 to 1. Domain experts from biology and medicine manually reviewed the quality of extracted papers using a different similarity score. They agreed to consider articles having a similarity score more than **0.5** on a scale of 0 to 1 as a candidate for being in the related articles set.

3.1.3 Phase-C: Extracting candidate answers and ranking relevant papers

The objective of this phase is to extract the top-k papers possibly containing the answer to a user query out of a relevant papers list computed in Phase-B about COVID-19 and that are most relevant. Therefore, with the relevant papers' list, S_Covid extracts candidate answer sentences from each paper that can be possible answers to the user query. This phase consists of two steps. (i) Step-1: Out of all the selected top relevant papers extract the candidate answer sentences based on the user query and, (ii) Step-2: Ranking the relevant papers based on the value associated with candidate sentences.

3.1.3.1 Step 1: Extracting candidate answer sentences

This step converts each relevant paper in the `top related papers set` computed in Stage-B into a set of "representative" sentences. We create a list of tuples $Top\_relevant\_paper\_sentences :< sentence, paperID >$. Each sentence in the list is converted into a vector representation using our previously trained `word2vec` model.
Given the query vector representing the user query, we use a Cosine similarity, to calculate a similarity score between the query vector and all the elements of the $Top\_relevant\_paper\_sentences$ list. Based on the similarity score, we choose sentences with a similarity score above a fixed threshold and then sort the sentences in a manner such that the sentences with a higher similarity score are placed at the top of the list of candidate answers. Our experiments showed that sentences having a Cosine similarity score equal to or above 0.5 on a scale of 0 to 1, can be considered candidates for an answer to the query.

3.1.3.2 Step 2: Ranking relevant papers

Once the top candidate papers have been determined, the sentences are grouped with their respective research paper using $< paperID >$. So, now for each paper having a set of answer candidates, our model finally re-ranks the papers based on the number of answer candidates of each paper. For example, if papers 'A' and 'B' have respectively ten and seven candidate sentence-answers, then 'A' is ranked above 'B' in the ranking of relevant papers list. The idea behind this method is that if a paper has a high number of candidate sentences, it means that its content is contextually more aligned with the user query. Consequently, the paper should be ranked higher in the relevant paper ranking.

S_Covid is an innovative tool designed to help users efficiently navigate through the vast and ever-growing amount of COVID-19 research papers available. With a dataset consisting of over 80,000 scientific papers, S_Covid utilizes state-of-the-art natural language processing techniques to quickly and accurately extract the most relevant research papers in response to a user's query.
Whether you're a researcher, healthcare professional, or simply someone interested in keeping up-to-date with the latest COVID-19 findings, S_Covid is a powerful resource that can save you time and energy in your search for information. Simply input your question, or "query", and specify the number of relevant papers you want to extract, or "k". S_Covid will then scan through its extensive database and present you with the

top-k related papers that best match your query.

With S_Covid, you can feel confident that you're getting the most accurate and up-to-date information on COVID-19 research, allowing you to stay informed and make informed decisions.

In **Stage-A**, our approach involves several steps to discover and rank relevant papers for a given user query. Specifically, we:

- Create an *LDA* topic model with 50 topics to help identify key themes in the CORD-19 dataset [2].
- Train the *LDA* model using the CORD-19 dataset, which consists of over 200,000 scholarly articles related to COVID-19.
- Calculate the probability distribution of topics over each paper in the CORD-19 dataset, as well as for the user's query. This is done using the *doc_topic_dist* and *query_topic_dist* measures, respectively.
- For each paper in the dataset, calculate the similarity score between the paper's probability distribution and the user's query probability distribution using the Jensen-Shannon distance. This gives us a measure of how closely the paper aligns with the user's interests and needs.
- Filter the papers using a threshold on the similarity score, such as *similarity_score (article)* > 0.5. This step helps us identify the papers that are most likely to be relevant to the user's query.
- As a result of these steps, we create a set of the most relevant papers, which we call *top_nearest_articles*. These papers are sorted in descending order of their similarity score so that the most relevant papers are presented first.

In **Stage-B**, for papers in *top_nearest_articles* we extract a list of sentences- sentences: We take the most relevant articles from Stage-A and extract all of the sentences from each article to create a list of candidate answer *sentences*.

- Vectorize sentences using our trained Word2vec model - *sentences-vectorized*: We use the Word2vec algorithm to convert each sentence in our list of candidate answer sentences into a vector representation.

- Vectorize user question - *query_vectorized*: We also convert the user's query into a vector representation using the same Word2vec algorithm.

- For each sentence, calculate the cosine similarity between the *sentences-vectorized* and *query_vectorized* - *similarity*: We calculate the cosine similarity score between each sentence in our list of candidate answer sentences and the user's query vector to measure how similar they are.

- Using a condition on cosine similarity score - $similarity > 0.5$, selecting top sentences as candidate answer sentences - *top_similar_sentence*: We select the sentences with a cosine similarity score above a threshold of 0.5 as candidate answer sentences.

- Followed by sorting the candidate sentences based on the cosine similarity score - *sorted_sentences*: We sort the candidate answer sentences based on their cosine similarity score to the user's query in descending order.

- Then grouping the sentences with their respective research paper using paper ID - *Group_by_paperid* to have candidate-relevant papers - *candidated_results*: We group the candidate answer sentences by the research paper they come from, so we can identify the papers that have the most relevant information for the user's query.

Finally, **Stage-C** ranks the papers in *candidated_results* such that the papers with a greater number of answer candidate sentences are at the top i.e. more relevant, and presenting *Top_k_related_paper*: We rank the papers in *candidated_results* by the number of candidate answer sentences they have, and present the top k papers as the most relevant research papers for the user's query.

3.2 S_Covid Algorithm

Below we have presented our proposed algorithm in detail given the dataset consisting of 80,000 COVID-19 papers. For a given user question input:- Query, the number of relevant papers the user wants to extract:- k; S_Covid outputs:- Top-k related paper to the Query.

Algorithm 1: S_Covid algorithm

80000 COVID-19 articles
Input: (Query , k , Data)
Output: Top-k related paper to the Query

[1] STAGE-A;

$lda \longleftarrow LatentDirichletAllocation(n\_components = 50)$;
lda.fit ($data\_vectorized$);
$doc\_topic\_dist \longleftarrow lda.transform(data\_vectorized)$;
$query\_topic\_dist \longleftarrow lda.transform(query\_vectorized)$;
foreach *article in the Data* **do**
    $similarity\_score(article) \longleftarrow 1 - Jensenshannon(query\_topic\_dist, doc\_topic\_dist[article])$; **if** $similarity\_score(article) > 0.5$ **then**
        $top\_nearest\_articles \longleftarrow article$;

STAGE-B;

$sentences \longleftarrow extract\_sentence(top\_nearest\_articles)$;
$sentences\_vectorized \longleftarrow word2vec(sentences)$;
$query\_vectorized \longleftarrow word2vec(Query)$;
foreach *sentence in the sentences_vectorized* **do**
    $similarity \longleftarrow cosine\_similarity(query\_vectorized, sentence)$;
    if $similarity > 0.5$ **then**
        $top\_similar\_sentence \longleftarrow sentence$;

$sorted\_sentences \longleftarrow sorted(top\_similar\_sentence)$;
$candidated\_result \longleftarrow Group\_by\_paperid(sorted\_sentences)$;
$Top\_k\_related\_paper \longleftarrow$ Select top k articles which have most sentences in $candidated\_result$;

Chapter 4
Experiments

In this section, we present the experimental setting used to assess the performance of our approach, as well as a comparative study of various baseline models along with our proposed S_Covid model. We also provide a detailed explanation of our method for evaluating the performance of various data exploration algorithms.

To evaluate the accuracy and efficiency of our search engine, we conducted a series of experiments using feedback from real users. This involved presenting users with a set of search queries related to COVID-19 research, and assessing the performance of our approach compared to various baseline models. We collected and analyzed user feedback to identify areas for improvement and refine our approach over time.

We systematically compared the performance of our approach against several state-of-the-art data exploration algorithms, including keyword-based search and content-based filtering. Our results demonstrate that our approach outperforms these baseline models, providing more accurate and efficient results for users.

To ensure the robustness of our approach, we conducted extensive testing on a large dataset of over 80,000 scholarly articles related to COVID-19 and coronavirus, including articles with full text. This allowed us to assess the scalability and effectiveness of our approach, as well as identify any potential limitations or areas for improvement.

4.1 Experiment settings

Our study involved conducting experiments on the Google Colab platform, utilizing a powerful computational setup that includes 12GB NVIDIA GPU RAM and Xeon processors running at 2.3Ghz, with support for CUDA-10.0. This state-of-the-art hardware allowed us to efficiently process and analyze large datasets, making it an ideal platform for our research. To conduct our study, we used the COVID-19 Open Research Dataset

(CORD-19) [2], which is a comprehensive and constantly-updated collection of over 100,000 scholarly articles related to COVID-19 and coronavirus. Developed and released by the Allen Institute for AI, this dataset includes over 80,000 articles with full text, making it a valuable resource for researchers in the field.

4.2 Dataset

In our experiment, we used CORD-19 [2] (COVID-19 Open Research Dataset) dataset. The corpus currently includes over 100,000 scholarly articles and updates weakly. The Allen Institute for AI published this dataset for the global research community. The objective was to enable the community to apply techniques in natural language processing/natural language understanding and Artificial Intelligence techniques to generate new insights about this infectious disease. CORD-19 is a collection of research papers published by various international publishing bodies.
The sources of papers come from the databases PMC, bioRxiv, medRxiv, Elsevier, Springer Nature, and WHO. The metadata involves papers published by bioRxiv, medRxiv, PMC, WHO, and individual publishers. The data was provided in JSON format which we converted into CSV files with the schema columns: paper_id title, authors affiliations, abstract, text, bibliography, raw_authors, and raw_bibliography. We further preprocessed the data to produce a clean and structured dataset before using it in S_Covid.

4.2.1 Data pre-processing

We used the version-52 of original CORD-19 [2] dataset which contained around 116,005 COVID-19-related research papers. The dataset schema consisted of the following attributes:

- *paper_id*
- *body_text*
- *methods*
- *results*
- *cord_uid*

- *source*
- *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*
- *is_covid19*
- *publish_year*

We initiated a filtering process to select 8 attributes from the dataset schema that best represented the content of the papers, based on the following criteria:

- *paper_id*: unique paper id for each article
- *title*: title of the paper

- *abstract*: abstract of the article
- *body_text*: full-length text of the article
- *doi*: DOI id of the article
- *url*: hyperlink to the article's publication website
- *publish_year*: publication year of the article
- *is_covid19*: contains either true (COVID-19 only article) or false value.

We went further into an attribute engineering process to add an attribute named '*complete text*' to the schema. This new attribute concatenates three attributes of the dataset schema - Title, Abstract, and Body text. In this manner, we effectively simplified the schema. We removed research papers that were not written in English. We also removed the abstract-only papers and considered only those papers which have all the three body text [1]. The final result is a collection of 80,000 COVID-19 research papers. We then performed the text preprocessing on the '*complete text*'. We used `ScispaCy` [15], a Python package containing spaCy models for processing biomedical, scientific or clinical text. We removed the stop words and performed word lemmatization on the text data. We also removed some common unnecessary words such as author, figure, copyrights, license, and fig from the '*complete text*'. But, we kept important information such as citation numbers intact so that the user would get a complete answer without any missing values. We used '*complete text*' to train the `word2vec` model and to extract sentences from each paper.

[1]Research [14] has proven that using full text for information retrieval is more effective than just using the abstract section of the research paper.

Table 4.1: COVID-19 Tasks

Task Id	Task details
$Task1$	What is known about transmission, incubation, and environmental stability?
$Task2$	What do we know about COVID-19 risk factors?
$Task3$	What do we know about vaccines and therapeutics?
$Task4$	What do we know about virus genetics, origin, and evolution?
$Task5$	What has been published about medical care?
$Task6$	What do we know about non-pharmaceutical interventions?
$Task7$	What has been published about ethical and social science considerations?
$Task8$	What do we know about diagnostics and surveillance?
$Task9$	What has been published about information sharing and inter-sectoral collaboration?

Id	Question
$Q1$	The incubation period of coronavirus disease
Q_2	The effect of seasons on transmission of COVID-19
Q_3	Risk factors for severe disease and death
Q_4	Efforts to develop a SARS-CoV vaccine
Q_5	Risk-reduction strategies
Q_6	Misinformation relate to COVID-19
Q_7	The economic impact of COVID-19
Q_8	Use of diagnostics markers to detect early COVID-19 disease
Q_9	Protocols for screening and testing of COVID-19
Q_{10}	Outcomes data for COVID-19 after mechanical ventilation

Table 4.2: COVID-19 Questions

Chapter 5
Performance Analysis of S_Covid: A Comparative Study

5.1 Existing Tools and Baseline Models Analysis

We performed an in-depth quality analysis of S_Covid model by comparing its performance with the existing tools and baseline models. We used a set of COVID-19 related questions shown in Table 4.2 and Table 5.2 for the comparative study.

5.1.1 Search Engines

For assessing the results of our experiments, we used the following baseline models to compare the performance of our proposed model- S_Covid. The following are the models:

5.1.1.0.1 COVIDEX: It is a COVID-19 research papers' search engine [3]. It leverages the BM25 [16] algorithm and T5 [4] language model fine-tuned on medical text data for relevant paper retrieval and ranking. It also uses a pre-trained BioBERT [5] model to extract and highlights the interesting sentences in each paper based on the user query.

5.1.1.0.2 KDCOVID: It is a tool for exploring COVID-19 research dataset [1]. It uses the similarity between the user query and sentences in the full text of papers in CORD19 corpus using a similarity metric derived from BioSentVec [7] to find relevant

[1]http://kdcovid.nl/about.htmlhttp://kdcovid.nl/about.html

System	P@3	P@5	NDCG@5	bpref
COVIDEX	0.6333	0.6133	0.5398	0.1187
KDCOVID	0.6733	0.6333	0.5390	0.1049
Google	0.7200	**0.7133**	0.5583	0.1063
LDA	0.5733	0.5700	0.4972	0.1069
LDA-BM25	0.6467	0.6133	0.5225	0.1303
LDA-WHOOSH	0.6467	0.6367	0.5466	0.1279
S_Covid	**0.7300**	0.6400	**0.6109**	**0.1352**

Table 5.1: Evaluation result

papers and highlight key points.

5.1.1.0.3 COVID-19 Research Explorer (Google): It is based on neural networks for natural language understanding [2]. This model is trained on 100,000+ scholarly articles on COVID-19. It is a supervised model powered by semantic search to help the model understand the context of the query. It encodes the query and documents as vectors and then performs retrieval by looking for the document vectors that are most similar to the query vector using k-nearest neighbour retrieval.

5.1.2 Baseline Models

We implemented different LDA-based information retrieval models. We coupled LDA with various searching algorithms (i.e. BM25 and WHOOSH) to design a COVID-19 literature exploration model such as (LDA, LDA-BM25 and LDA-WHOOSH). In our study, we compared and analysed the performances of these models against our proposed S_Covid model.

5.1.2.0.1 LDA: For this baseline model, we used LDA topic modeling to first calculate the probability distribution of topics over the papers and the given query. Then, it uses a similarity score **(S = 1 - Jensen–Shannon distance)** to select and rank the papers based on their relevance for the query.

[2]https://covid19-research-explorer.appspot.comhttps://covid19-research-explorer.appspot.com

Figure 5.1: Screenshot of our S_COVID application, which builds on LDA and our rankings algorithm.

5.1.2.0.2 LDA-BM25: For this baseline model, we used the same LDA topic modeling for finding relevant papers followed by ranking the relevant papers using the scoring function- Okapi BM25 [16]. BM25 is a bag-of-words retrieval function that ranks COVID-19 research papers based on the user query terms appearing in each document, regardless of their proximity within the papers.

5.1.2.0.3 LDA-WHOOSH: In this baseline model, we used the LDA topic modeling and Whoosh[3] to find and rank relevant papers. First, it selects the papers using LDA and the similarity score **(S = 1 - Jensen–Shannon distance)**. Then it uses Whoosh to create an index for selected papers and, based on the query, it searches the index to find the relevant papers. Finally, the relevant papers are sorted using BM25, the Whoosh default ranking algorithm.

[3]Whoosh is a python library for indexing text and then searching indexes https://pypi.org/project/Whooshhttps://pypi.org/project/Whoosh

5.1.3 Experimental comparison

For a systematic assessment of the model's performance, extensive manual evaluation is done by biomedical field experts. We worked with three field experts of medicine and biology from the National Institute of Genetic Engineering and Biotechnology in Tehran, Iran, and Golestan University of Medical Sciences to assess the S_Covid algorithm and the baseline models used to compare S_Covid.

To have a standard set of diverse queries, we chose a subset consisting of 30 questions (see Table 4.2 and 5.2) out of 100 questions of the Kaggle CORD-19-research-challenge[4] (CORD-19). The challenge consists of 17 sub-tasks out of which 9 sub-tasks (refer to Table 4.1) were related to information retrieval. Each of the 9 sub-tasks consists of 5-10 questions. We selected 3-4 questions for each of the 9 tasks. The selection pattern of our choice was according to two criteria. First, we chose topics with enough variety to cover all aspects of the pandemic. Second, we selected questions that engaged scientific minds about COVID-19. Since, out of 100 questions, various questions were overlapping and repetitive. To overcome this issue, our biology and medical experts identified the 30 most interesting questions which were most distinctive.

For a given query, each of the extracted documents was assigned to one out of three categories: *relevant*, *partially relevant*, or *not relevant*. In the experimental comparison. we utilized the following standard evaluation matrices: NDCG@5 (normalized discounted cumulative gain with top 5 documents), bpref (binary preference), and P@5 (precision at 5 documents). For the bpref calculation, we only used judged documents whereas, for the other two measures, we assumed the non-judged documents as *not relevant*. It is important to note that we extracted the top 5 documents for each system which were manually judged by the experts.

For a better understanding of the systems, we performed an additional in-depth error analysis of each system using a custom evaluation method (please refer 5.2).

5.1.4 Result and Discussion

This section presents a comprehensive analysis of the performance of various models used in the experiments, with a focus on the results obtained from the evaluation metrics. The evaluation outcomes are summarized in Table 5.1. Our study revealed that S_Covid outperformed the baseline models across several evaluation metrics. While Google's system achieved a higher precision at 5 (P@5) value, its precision at 3 (P@3) was

[4]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

comparatively lower. This suggests that although Google was able to retrieve a larger number of relevant papers, S_Covid's recommendations were more focused and accurate for the user's specific query. It is worth noting that this difference in performance could be due to Google's access to a larger database, enabling it to retrieve more relevant papers. Conversely, in terms of NDCG@5 and bpref metrics, S_Covid achieved higher scores than the other baseline models, indicating that it was able to provide more relevant and diverse results for the given user query. Overall, the results demonstrate the superiority of S_Covid in comparison to other existing tools and models.

5.2 An additional error analysis method

Search engines are designed to cater to the specific needs of different groups of people, and it is important to continuously evaluate them based on user feedback. In light of this, we conducted a thorough evaluation of our algorithm, soliciting feedback at every stage of development to enable continuous improvement. However, given the lack of a standardized question-answering dataset for COVID-related topics, we developed an additional manual assessment method to gain more detailed insights and compare the performance of S_Covid with other baseline models. To quantitatively evaluate our approach, we assigned scores to each result based on its relevance to the user queries. Our aim was to provide a comprehensive and rigorous evaluation of the performance of S_Covid, taking into account both existing evaluation metrics and additional assessment criteria that we developed to ensure a more nuanced understanding of the system's strengths and limitations. In this regard, we allocate three scores for our findings, as mentioned in the following formula:

$$\text{Evaluation (article)} = \left\{ \begin{array}{cc} 1 & Relevant \\ 0 & PartiallyRelevant \\ -0.5 & NotRelevant \end{array} \right\} \tag{5.1}$$

In the formula, the score of -0.5 is a penalty for completely irrelevant articles, the score of zero is for papers that may be a candidate for parts of our findings and are partially relevant, and the score of one is for documents that are completely related to our questions. We calculated the evaluation score of all top-k articles extracted by the

models for each given query as:

$$\text{Evaluation\_score } (Q_j^m) = \frac{\sum_{i=1}^{k} Evaluation(topkarticle_j^m(i))}{k} \tag{5.2}$$

where m represents a model, i.e.:
m ∈ { LDA , LDA+BM25, LDA+Whoosh,Google, kdcovid.nl,covidex.ai, S_Covid } and j is a question id.

To have a standard set of diverse queries, we chose a subset consisting of 30 questions (see Table 4.2 and 5.2) out of 100 questions of Kaggle CORD-19-research-challenge[5] (CORD-19). The challenge consists of 17 sub-tasks out of which 9 sub-tasks (refer table 4.1) were related to information retrieval. Each of the 9 sub-tasks consists of 5-10 questions. We selected 3-4 questions for each of the 9 tasks. The selection pattern of our choice was according to two criteria. First, we chose topics with enough variety to cover all aspects of the pandemic. Second, we selected questions that engaged scientific minds about the COVID-19. Since, out of 100 questions, various questions were overlapping and repetitive. To overcome this issue, we identified the 30 most interesting questions which were most distinctive. We utilized 10 questions (refer to Table 4.2) out of 30 interesting questions selected by the experts (refer 5.1.4). For each task (refer to Table 4.1), we chose at least 1 question (refer to Table 4.2) to keep the error analysis consistent and diverse.

To illustrate our scoring method, we show three results in response to a COVID-19 question. As given in Table 4.2, "the effect of seasons on the transmission of COVID-19" is one of the searched query items. The following papers are three of our first five related papers extracted by the algorithm.
As shown in figure 5.1, we refined our results by selecting COVID-19 related articles, which have terms like COVID-19, Coronavirus, and 2019-nCoV. To clarify our scoring method, we show three results in response to a specific question. Take "the effect of seasons on the transmission of COVID-19" as one of our ten searched items, the following papers are three of our first five.

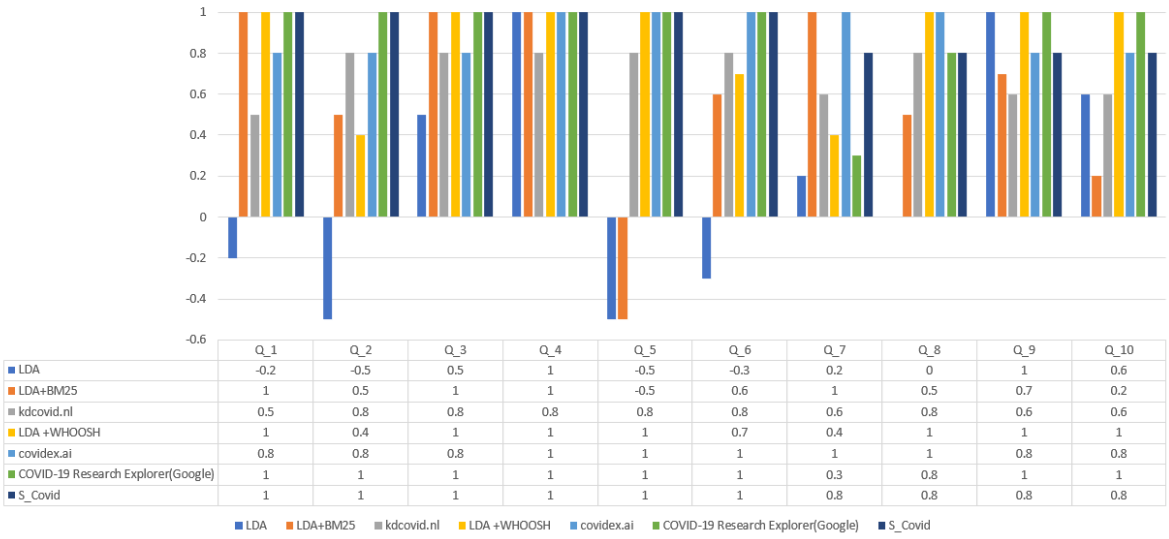
1. "Climate effect on COVID-19 spread rate: an online surveillance tool [17]."

2. "Projecting the transmission dynamics of SARS-CoV-2 through the post pandemic period [18]."

[5]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challengehttps://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

3. "Excess cases of influenza suggest an earlier start to the coronavirus epidemic in Spain than official figures tell us: an analysis of primary care electronic medical records from over 6 million people from Catalonia [19]."

The first one gained a score of one because it has relevant data related to our query. The second article gained zero because it "seems" to be related to our questions and the third one gained a -0.5 score as a penalty because of its irrelevant result.



	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_10
LDA	-0.2	-0.5	0.5	1	-0.5	-0.3	0.2	0	1	0.6
LDA+BM25	1	0.5	1	1	-0.5	0.6	1	0.5	0.7	0.2
kdcovid.nl	0.5	0.8	0.8	0.8	0.8	0.8	0.6	0.8	0.6	0.6
LDA +WHOOSH	1	0.4	1	1	1	0.7	0.4	1	1	1
covidex.ai	0.8	0.8	0.8	1	1	1	1	1	0.8	0.8
COVID-19 Research Explorer(Google)	1	1	1	1	1	1	0.3	0.8	1	1
S_Covid	1	1	1	1	1	1	0.8	0.8	0.8	0.8

Figure 5.2: Question wise comparative evaluation of S_Covid against baseline models based on *Evaluation_score* (refer 5.2).

5.2.1 Comparative survey of S_Covid with baseline models:

We performed an in-depth quality analysis of the S_Covid model by comparing its performance with the existing tools and baseline models. We used the same set of COVID-19 related questions shown in Table 4.2 and Table 5.2 for the comparative study.

5.2.1.1 Experimental comparison

The experimental comparison consists of the following three steps:

1. **Step-1:** Each model outputs a set of papers ranked according to the relevance of a paper for a given query. We selected top 5 (k=5) papers from the set for each model.

2. **Step-2:** Using our scoring method, experts assign a score to each of the paper extracted by that model based on the given query. The expert thoroughly reads the paper's title and abstract to determine the quality of the paper and its relevance for a given query.

3. **Step-3:** For each paper, we received three scores, each one corresponding to an expert. We finally assigned the most frequent score (0, 1 or -0.5) to the paper. Then, for comparing different models performance, we calculated the average performance score of each model based on the score obtained by the papers by using the *Evaluation_score* formula 5.2.

For evaluating the performance of the seven models (engines, baselines and S_Covid), experts manually reviewed 1050 papers (35 papers for each of the 30 questions). In some cases, experts also considered reading the extracted research paper's full text for a better understanding of why they were relevant for a query and making their final decision. A complete comparative evaluation of S_Covid against baseline models for 10 randomly selected questions (out of 30 questions), where each question belongs to one of the 9 Kaggle's challenge task (refer to Table 4.1) is shown in Figure 5.2. A detailed analysis of the results is given next.

5.2.2 Results and Discussion

Figure 5.3 shows the overall performance of S_Covid and the baseline models over 30 selected COVID-19 questions. We can see that S_Covid outperformed other models in terms of average paper quality. A key observation is that the performance score of 'COVID-19 Research Explorer (Google)' and 'coveidex.ai' is very close to S_Covid. To understand this, we can refer to Figure 5.2, which shows question wise performance of each model.

But the quality of related output papers is not the only advantage of S_COVID. By finding the most relevant sentences (Table 5.3) of selected documents, it helps researchers to save time. Our sentence finder can also be customized based on the number of convictions and arranged according to their importance (figure 5.1). It is crystal clear that finding and reading-related papers are one of the most critical but time-consuming prosses of scientific work. It is mainly about the researchers who work on systematic reviews and meta-analyses. Therefore such a "helper" like S_COVID can facilitate research prosses and drive the force of researchers toward scientific innovation rather than just wandering and being lost in an immense amount of scientific papers.

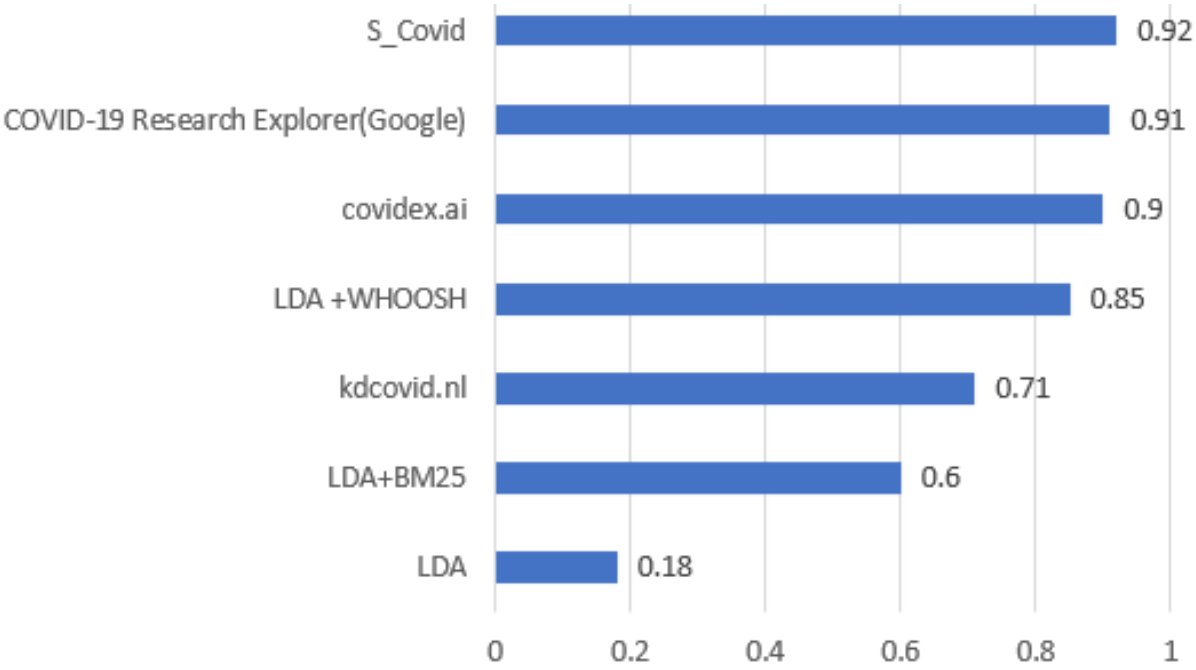


Figure 5.3: Overall performance score obtained by baseline models and S_Covid Average score

In our survey, we found that almost all algorithms faced issues while extracting an answer for the question Q_2. One possible reason might be that due to the prevalence of COVID-19 in all countries, there is no noticeable seasonal change, hence there is not any published work which can answer the query. For question Q_7, there is a similar reason: because the COVID-19 outbreak is a relatively new phenomenon in the world, there are not enough publications in our database related to its economic impacts on people and countries as of now. In the later update, we can evaluate these search algorithms concerning noted questions.

In order to work concerning, the user needs. We also developed an application, and the results in figure 5 are a shot of the app to question one in table 1. This app brings related papers and the most significant sentence related to the queries.it is just the simplest version of our app as soon as possible brings a user-friendly version of our search engine. As shown in table 5.3, S_COVID shows essential sentences of related papers given a query. It can also be customised according to the number of sentences that users need (figure 5.1). It is clear that reading the essential sentences of papers facilities research, especially in the systematic review and meta-analysis research area, where scientists need to understand and evaluate hundreds of documents.

In our survey, we noticed that almost all algorithms have a problem concerning question

two (figure 5.2). One possible reason may be that since the prevalence of COVID-19 in all countries, there is not noticeable seasonal change, and there is not enough data about this query. About question seven, also a similar reason because this outbreak is a relatively new phenomenon in the world. There is not enough publication about the economic impacts of this epidemic on peoples and countries.

As shown in Figure 5.2, S_Covid achieves much more accurate results than other search methods. Why did COVID-19 Research Explorer (Google) perform better with respect to S_Covid in questions Q_9,Q_{10}? Because they have a larger dataset (100,000+ papers). As a result, they can retrieve more papers than S_Covid. But having a larger dataset might be disadvantageous, as seen in question Q_7. In this case, COVID-19 Research Explorer (Google) performance dropped because of the fourth retrieved article "A national prospective cohort study of SARS/COV2 pandemic outcomes in the U.S.: The CHASING COVID Cohort [16]" which is not related to the query. Therefore, by acquiring a -0.5 penalty, its overall result falls to 0.3. For Q_7, covidex.ai gave better results than S_COVID and COVID-19 Research Explorer (Google). S_COVID scored a bit lower because the fourth paper in its result ("COVID-19 outbreak: Migration, effects on society, global environment and prevention [20]") is partially related to our query, so it does not gain the perfect score of 1. On the other hand, for questionQ_2 covidex.ai had a score of zero because of the retrieved article "Modeling strict age-targeted mitigation strategies for COVID-19 [21]," which is only partially related to our query. S_Covid and COVID-19 Research Explorer (Google) performed relatively better for this query.

5.3 S_COVID features

The core of the powerful search engine S_COVID is user-provided keywords. The user provides the desired keywords in S_COVID's search box and, after clicking the "Search" button, all relevant articles appear at the bottom of the page (see Figure 5.1). The distinct strength of this engine is that it returns sentences from the paper that are possible answers to the question. The search engine also provides a traditional ranking option based on the publication year of the articles.

We have also showcased S_Covid's capability of finding the most relevant candidate answer sentences (see Table 5.3 and 5.2) out of selected papers. This shows that the quality of related output papers is not the only advantage of S_COVID. Our application response to question Q_1 in Table 4.2 is shown in Figure 5.1. In terms of filtering options, users can select (a) the number of papers and (b) the number of sentences which wants

to review them, by clicking on the slider buttons below the search box.
We have also provided filtering options to enhance the user experience. Users can choose the number of papers and sentences they want to review by clicking on the slider buttons below the search box. Moreover, users can exclude unrelated papers and only consider COVID-19 associated articles with terms like COVID-19, 2019 Novel Coronavirus, SARS Coronavirus 2, SARS-CoV-2, and 2019-nCoV.

Table 5.2: S_Covid answers for COVID-19 related questions

Question	S_Covid Answer
11. Persistence of virus on surfaces of different materials	A report by van Doremalen and colleagues found survival of both SARS-CoV and SARS-CoV-2 of up to 2 days (on surfaces) and 3 days (in aerosols generated in the laboratory), but again with a large inoculum [22].
12. Role of the environment in transmission	The Respiratory diseases are often simply assumed to be transmitted via "close contact"; however, the complex transmission mechanisms often involve with more than one transmission route including direct or indirect contact, large droplet, and airborne routes [23].
13. Approaches to evaluate risk for enhanced disease after vaccination	NHPs could be utilized to evaluate COVID-19 vaccine candidates without adjuvants and guide in the selection of vaccines that elicit desired attributes that could reduce the risk of vaccine-mediated enhanced disease [24].
14. Effectiveness of drugs being developed and tried to treat COVID-19 patients	Remdesivir (GS-5734™) is an antiviral drug developed by Gilead Sciences initially developed to treat Ebola, but experimental tests are also being carried out to treat diseases such as MERS and COVID-19 [25].
15. Outcomes data for COVID-19 after mechanical ventilation adjusted for age	As age increased, so did the proportion of patients who required ICU admission, invasive mechanical ventilation, and vasopressors. Median age was 76 years (IQR, 66-85); 58% (n=244) were male; 71% (n=299) were admitted to the ICU; and 59% (n=246) received invasive mechanical ventilation [26].
16. Guidance on the simple things people can do at home to take care of sick people and manage disease.	The best health advice for older, frail patients was to stay home and health care providers offered televisits and telephonic symptom management to avoid unnecessary emergency department visits [27].
17. Policies and protocols for screening and testing.	In this study we propose a novel group testing protocol using a commercially available RT-dPCR assay and compare empirically the sensitivity of individual identification through RT-PCR with group testing by RT-dPCR for three groups sizes of 8, 16 and 32 samples [28].
18. Efforts to track the evolution of the virus	Mutation and adaptation have driven the co-evolution of coronaviruses (CoVs) and their hosts, including human beings, for thousands of years [29].
19. Effectiveness of case quarantine of exposed individuals	If 90% of cases are asymptomatic or undetected, as could happen in a location making no effort to follow people during their isolation, the efficacy of quarantine would be about 70% [30] .
20. Effectiveness of inter/inner travel restriction	During the peak of the COVID-19 outbreak in Europe, about 3-6% of air passengers were SARS-CoV-2 positive on repatriation flights [31].
21. Effectiveness of school distancing	We fit our model with the data until August 29th, and then simulate what would happen in the event that schools open in mid-September [32].
22. How does temperature and humidity affect the transmission of 2019-nCoV?	Analyzed meteorological data of 30 cities in China and suggested that low temperature, mild diurnal temperatures, and low humidity likely aid the transmission of novel coronavirus disease 2019 (COVID-19) [33].
23. What is the efficacy of novel therapeutics being tested currently?	Remdesivir and favipiravir are the most promising antiviral drugs that have been tested in clinical trials so far [34].
24. Risk factor studies related to impact of diabetes	Conclusion: Older people above 65 years old and diabetic patients are significant risk factors for COVID-19 [35].
25. Risk factor studies related to impact of male gender	The multivariate analyses, age over 50 years, male gender and low-medium socioeconomic status were also positively associated with the risk of COVID-19 infection [36].
26. Risk factor studies related to impact of kidney disease	Kaplan-Meier analysis demonstrated that patients with kidney disease had a significantly higher risk for in-hospital death [37].
27. Risk factor studies related to impact of cancer	Combined with previously published results , we can conclude that patients with cancer have an increased risk of COVID-19 [38].
28. Risk factor studies related to impact of overweight	Our findings are consistent with other reports from New York that did not find obesity to be an independent risk factor for mortality, though reports from reviews suggest obesity does play a role in mortality [39].
29. What do we know about viral shedding in stool?	In addition, we identified six studies presenting indirect evidence on the potential for SARS-CoV-2 transmission by children, three of which found prolonged virus shedding in stools [40].
30. What is the longest duration of viral shedding?	This happens to coincide with the fact that the viral RNA shedding of children is much longer than that of adults(13, 14) [41].

Id	S_Covid Answer
Q_1	In our analysis of 44 patients with clear contact history, we found that the mean incubation period of COVID-19 was 8 [42].
Q_2	Our findings of decreased replication and spread rates of COVID-19 in warm climates may suggest that the inevitable seasonal variance will alter the dynamic of the disease spread in both hemispheres in the coming months [17].
Q_3	Evidence from China, Italy and the USA indicates that older individuals, males and those with underlying conditions, such as CVD, diabetes and CRD, are at greater risk of severe COVID-19 illness and death [43].
Q_4	Here, we discuss therapeutic and prophylactic interventions for SARS-CoV-2 with a focus on vaccine development and its challenges [44].
Q_5	In addition, we provide suggestions to aid further development of epidemic prevention and control strategies, and scientific decision-making [45].
Q_6	The World Health Organization have emphasized that misinformation - spreading rapidly through social media - poses a serious threat to the COVID-19 response [46].
Q_7	We identify a total of 35 potential determinants that describe a diverse ensemble of social and economic factors,including healthcare infrastructure, societal characteristics,economic performance, demographic structure etc [47].
Q_8	Those findings indicate that our N antigen assay is an accurate, rapid, early and simple diagnosis method of COVID-19 [48].
Q_9	We established routines for SARS-CoV-2 RNA extraction-free single-reaction RT-qPCR testing [49].
Q_{10}	All patients demonstrated stable physiology and ventilation for the duration of shared ventilation [50].

Table 5.3: Related sentences extracted by S_Covid for the questions in Table 4.2

Chapter 6
Conclusions

The current global pandemic has created a pressing need for efficient search engines that can assist medical experts in quickly finding relevant COVID-19 literature. In response to this need, we present S_Covid, a user-friendly and effective search engine designed to aid in the search for COVID-19-related papers. Our system employs text extraction techniques to provide accurate and relevant answers to specific queries, enabling researchers to make informed decisions about which papers to investigate further.

In this paper, we thoroughly evaluate the performance of both existing COVID-19 literature search engines and our own S_Covid model. Working closely with medical domain experts, we identify the strengths and weaknesses of existing models and highlight how our model addresses their limitations. Our aim is to provide a comprehensive overview of the current state of COVID-19 literature search engines and to demonstrate the capabilities of our own system.

We believe that the contributions we make in this paper will prove to be valuable to the scientific community in the fight against the COVID-19 pandemic. Moreover, we are confident that the techniques and methods we have developed can be applied more broadly to the analysis of scientific literature in general.

6.1 Future Work

Looking ahead, our goal is to provide data scientists with powerful data collection and exploration tools that can help them derive insights from large volumes of scientific literature. We believe that incorporating the following techniques in the future will play an important role in advancing the field of scientific research in the years to come.

-

- Integration with natural language processing (NLP) techniques: One potential area of improvement for S_Covid could be the integration of NLP techniques to improve the accuracy and relevance of search results. We plan to focus on developing data exploration techniques such as query morphing, queries as answers, and query-by-example mechanisms.
- Multilingual support: As the COVID-19 pandemic is a global issue, it would be beneficial to develop multilingual support for S_Covid to assist researchers around the world in accessing COVID-19 related literature in their native language.
- Collaboration with domain experts: Ongoing collaboration with medical domain experts could help to further refine and improve the search capabilities of S_Covid.
- Expansion to other domains: The capabilities developed for S_Covid could potentially be expanded to other domains beyond the COVID-19 pandemic, such as other infectious diseases or broader topics in the medical field.
- Integration with other data sources: Integrating S_Covid with other relevant data sources such as clinical trial data, electronic health records, and genomic data could provide more comprehensive and accurate results for researchers.
- User interface improvements: Further user interface improvements could be made to make the search process more intuitive and user-friendly for subject-matter experts.
- Extension to other forms of literature: The capabilities developed for S_Covid could potentially be extended to other forms of literature, such as patents or technical reports, to assist researchers in other areas of scientific inquiry.

Bibliography

[1] Soni, S. and K. Roberts (2021) "An evaluation of two commercial deep learning-based information retrieval systems for covid-19 literature," *Journal of the American Medical Informatics Association*, **28**(1), pp. 132–137.

[2] Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al. (2020) "CORD-19: The Covid-19 Open Research Dataset," *arXiv preprint arXiv:2004.10706*.

[3] Zhang, E., N. Gupta, R. Nogueira, K. Cho, and J. Lin (2020) "Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned," *arXiv preprint arXiv:2004.05125*.

[4] Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2019) "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*.

[5] Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang (2020) "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, **36**(4), pp. 1234–1240.

[6] Weston, L., V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain (2019) "Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature," *Journal of chemical information and modeling*, **59**(9), pp. 3692–3702.

[7] Chen, Q., Y. Peng, and Z. Lu (2019) "BioSentVec: creating sentence embeddings for biomedical texts," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, pp. 1–5.

[8] Piñero, J., À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong (2016) "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic acids research*, p. gkw943.

[9] Tang, R., R. Nogueira, E. Zhang, N. Gupta, P. Cam, K. Cho, and J. Lin (2020) "Rapidly Bootstrapping a Question Answering Dataset for COVID-19," *arXiv preprint arXiv:2004.11339*.

[10] Řehůřek, R. and P. Sojka (2010) "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50, `http://is.muni.cz/publication/884893/en`.

[11] Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013) "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119.

[12] Hofmann, T. (1999) "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57.

[13] Blei, D. M., A. Y. Ng, and M. I. Jordan (2003) "Latent dirichlet allocation," *Journal of machine Learning research*, **3**(Jan), pp. 993–1022.

[14] Lin, J. (2009) "Is searching full text more effective than searching abstracts?" *BMC bioinformatics*, **10**(1), p. 46.

[15] Neumann, M., D. King, I. Beltagy, and W. Ammar (2019) "Scispacy: Fast and robust models for biomedical natural language processing," *arXiv preprint arXiv:1902.07669*.

[16] Robertson, S. E., S. Walker, M. Beaulieu, M. Gatford, and A. Payne (1996) "Okapi at TREC-4," *Nist Special Publication Sp*, pp. 73–96.

[17] Caspi, G., U. Shalit, S. L. Kristensen, D. Aronson, L. Caspi, O. Rossenberg, A. Shina, and O. Caspi (2020) "Climate effect on COVID-19 spread rate: an online surveillance tool," *medRxiv*.

[18] Kissler, S. M., C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch (2020) "Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period," *Science*, **368**(6493), pp. 860–868.

[19] Coma, E., N. Mora, A. Prats-Uribe, F. Fina, D. Prieto-Alhambra, and M. Medina-Peralta (2020) "Excess cases of influenza suggest an earlier start to the coronavirus epidemic in Spain than official figures tell us: an analysis of primary care electronic medical records from over 6 million people from Catalonia," *medRxiv*.

[20] Chakraborty, I. and P. Maity (2020) "COVID-19 outbreak: Migration, effects on society, global environment and prevention," *Science of the Total Environment*, p. 138882.

[21] Chikina, M. and W. Pegden (2020) "Modeling strict age-targeted mitigation strategies for COVID-19," *arXiv preprint arXiv:2004.04144*.

[22] Goldman, E. (2020) "Exaggerated risk of transmission of COVID-19 by fomites," *The Lancet Infectious Diseases*, **20**(8), pp. 892–893.

[23] Gao, C. X., Y. Li, J. Wei, S. Cotton, M. Hamilton, L. Wang, and B. J. Cowling (2020) "Multi-route respiratory infection: when a transmission route may dominate," *medRxiv*.

[24] Lambert, P.-H., D. M. Ambrosino, S. R. Andersen, R. S. Baric, S. B. Black, R. T. Chen, C. L. Dekker, A. M. Didierlaurent, B. S. Graham, S. D. Martin, et al. (2020) "Consensus Summary Report for CEPI/BC March 12-13, 2020 Meeting: Assessment of Risk of Disease Enhancement with COVID-19 Vaccines," *Vaccine*.

[25] Marcolino, V. A., T. C. Pimentel, and C. E. Barão (2020) "What to expect from different drugs used in the treatment of COVID-19: A study on applications and in vivo and in vitro results," *European Journal of Pharmacology*, **887**, p. 173467.

[26] Kim, L., S. Garg, A. O'Halloran, M. Whitaker, H. Pham, E. J. Anderson, I. Armistead, N. M. Bennett, L. Billing, K. Como-Sabetti, et al. (2020) "Risk factors for intensive care unit admission and in-hospital mortality among hospitalized adults identified through the US coronavirus disease 2019 (COVID-19)-associated hospitalization surveillance network (COVID-NET)," *Clinical Infectious Diseases*.

[27] Reckrey, J. M. (2020) "COVID-19 Confirms It: Paid Caregivers are Essential Members of the Healthcare Team," *Journal of the American Geriatrics Society*.

[28] Martin, A., A. Storto, B. Andre, A. Mallory, R. Dangla, B. Visseaux, and O. Gossner (2020) "High-sensitivity COVID-19 group testing by digital PCR," *arXiv preprint arXiv:2006.02908*.

[29] Ye, Z.-W., S. Yuan, K.-S. Yuen, S.-Y. Fung, C.-P. Chan, and D.-Y. Jin (2020) "Zoonotic origins of human coronaviruses," *International journal of biological sciences*, **16**(10), p. 1686.

[30] Arino, J., N. Bajeux, S. Portet, and J. Watmough (2020) "Assessing the risk of COVID-19 importation and the effect of quarantine," *medRxiv*.

[31] Massad, E., M. Amaku, A. Wilder-Smith, P. C. C. dos Santos, C. J. Struchiner, and F. A. B. Coutinho (2020) "Two complementary model-based methods for calculating the risk of international spreading of anovel virus from the outbreak epicentre. The case of COVID-19," *Epidemiology & Infection*, pp. 1–19.

[32] Stella, L., A. P. Martínez, D. Bauso, and P. Colaneri (2020) "The Role of Asymptomatic Individuals in the COVID-19 Pandemic via Complex Networks," *arXiv preprint arXiv:2009.03649*.

[33] Yuan, S., S. Jiang, Z.-L. Li, et al. (2020) "Do Humidity and Temperature Impact the Spread of the Novel Coronavirus?" *Frontiers in Public Health*, **8**, p. 240.

[34] Elshabrawy, H. A. (2020) "SARS-CoV-2: An Update on Potential Antivirals in Light of SARS-CoV Antiviral Drug Discoveries," *Vaccines*, **8**(2), p. 335.

[35] Goh, H. P., W. I. Mahari, N. I. Ahad, L. Chaw, N. Kifli, B. H. Goh, S. F. Yeoh, and L. C. Ming (2020) "Risk factors affecting COVID-19 case fatality rate: A quantitative analysis of top 50 affected countries," *medRxiv*.

[36] Merzon, E., D. Tworowski, A. Gorohovski, S. Vinker, A. Golan Cohen, I. Green, and M. Frenkel-Morgenstern (2020) "Low plasma 25 (OH) vitamin D level is associated with increased risk of COVID-19 infection: an Israeli population-based study," *The FEBS journal*, **287**(17), pp. 3693–3702.

[37] Cheng, Y., R. Luo, K. Wang, M. Zhang, Z. Wang, L. Dong, J. Li, Y. Yao, S. Ge, and G. Xu (2020) "Kidney disease is associated with in-hospital death of patients with COVID-19," *Kidney international*.

[38] Gao, Y., M. Liu, S. Shi, Y. Chen, Y. Sun, J. Chen, and J. Tian (2020) "Cancer is associated with the severity and mortality of patients with COVID-19: a systematic review and meta-analysis," *medRxiv*.

[39] Zimmerman, P., S. Stroever, T. Burton, K. Hester, M. Kim, R. Fahy, K. Corbitt, J. Petrini, and J. Nicastro (2020) "Mortality Associated With Intubation and Mechanical Ventilation in Patients with COVID-19," *medRxiv*.

[40] Li, X., W. Xu, M. Dozier, Y. He, A. Kirolos, E. Theodoratou, et al. (2020) "The role of children in transmission of SARS-CoV-2: A rapid review," *Journal of global health*, **10**(1).

[41] Chen, Z., L. Tong, Y. Zhou, C. Hua, W. Wang, J. Fu, Q. Shu, L. Hong, H. Xu, Z. Xu, et al. (2020) "Childhood COVID-19: a multicentre retrospective study," *Clinical Microbiology and Infection*, **26**(9), pp. 1260–e1.

[42] Ai, J., J. Chen, Y. Wang, X. Liu, W. Fan, G. Qu, M. Zhang, S. P. Pei, B. Tang, S. Yuan, et al. (2020) "The cross-sectional study of hospitalized coronavirus disease 2019 patients in Xiangyang, Hubei province," *MedRxiv*.

[43] Clark, A., M. Jit, C. Warren-Gash, B. Guthrie, H. H. Wang, S. W. Mercer, C. Sanderson, M. McKee, C. Troeger, K. I. Ong, et al. (2020) "How many are at increased risk of severe COVID-19 disease? Rapid global, regional and national estimates for 2020," *medRxiv*.

[44] Amanat, F. and F. Krammer (2020) "SARS-CoV-2 vaccines: status report," *Immunity*.

[45] Wang, J. and Z. Wang (2020) "Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of China's Prevention and Control Strategy for the COVID-19 Epidemic," *International Journal of Environmental Research and Public Health*, **17**(7), p. 2235.

[46] Laato, S., A. Islam, M. N. Islam, and E. Whelan (2020) "Why do people share misinformation during the Covid-19 pandemic?" *arXiv preprint arXiv:2004.09600*.

[47] Stojkoski, V., Z. Utkovski, P. Jolakoski, D. Tevdovski, and L. Kocarev (2020) "The socio-economic determinants of the coronavirus disease (COVID-19) pandemic," *arXiv preprint arXiv:2004.07947*.

[48] Diao, B., K. Wen, J. Chen, Y. Liu, Z. Yuan, C. Han, J. Chen, Y. Pan, L. Chen, Y. Dan, et al. (2020) "Diagnosis of Acute Respiratory Syndrome Coronavirus 2 Infection by Detection of Nucleocapsid Protein," *medRxiv*.

[49] Smyrlaki, I., M. Ekman, M. Vondracek, N. Papanicoloau, A. Lentini, J. Aarum, S. Muradrasoli, J. Albert, B. Högberg, and B. Reinius (2020) "Massive and rapid COVID-19 testing is feasible by extraction-free SARS-CoV-2 RT-qPCR," *medRxiv*.

[50] Levin, M., M. D. Chen, A. Shah, R. Shah, G. Zhou, E. Kane, G. Burnett, S. Ranginwala, J. Madek, C. Gidiscin, et al. (2020) "Differential ventilation using flow control valves as a potential bridge to full ventilatory support during the COVID-19 crisis," *medRxiv*.