

Course Project Report

Information Retrieval on CORD-19 with Question Answering and Summarization

Submitted By

Naveen Shenoy (191IT134)
Pratham Nayak (191IT241)
Sarthak Jain (191IT145)

as part of the requirements of the course

Information Retrieval (IT458) [Jul - Nov 2022]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Information Technology

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at



DEPARTMENT OF INFORMATION TECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL




JUL-NOV 2022

DEPARTMENT OF INFORMATION TECHNOLOGY
National Institute of Technology Karnataka, Surathkal

C E R T I F I C A T E

This is to certify that the Course project Work Report entitled **Information Retrieval on CORD-19 with Question Answering and Summarization** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
Naveen Shenoy	191IT134	
Pratham Nayak	191IT241	
Sarthak Jain	191IT145	




This report is a record of the work carried out by them as part of the course **Information Retrieval (IT458)** during the semester **Jul - Nov 2022**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology**.

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S

DECLARATION

We hereby declare that the project report entitled **Information Retrieval on COVID-19 with Question Answering and Summarization** submitted by us for the course **Information Retrieval (IT458)** during the semester **Jul-Nov 2022**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Naveen Shenoy	191IT134	
2. Pratham Nayak	191IT241	
3. Sarthak Jain	191IT145	

Place: NITK, Surathkal

Date: 9-11-2022

Information Retrieval on CORD-19 with Question Answering and Summarization

Naveen Shenoy ¹, Pratham Nayak ², Sarthak Jain ³

Abstract—A large number of research articles on COVID-19 have been published over the last few years. More than 400,000 such articles have been collected and are available through the COVID-19 Open Research Dataset (CORD). Effective information retrieval over the dataset can help health workers as well as scientific researchers in their research. In this paper, we present a COVID-19 information retrieval system with question-answering and summarization. Various keyword-based and neural-network-based models are used to efficiently reduce the search space and find relevant sentences from the vast corpus. These sentences are further shortlisted using a combination of various scoring metrics. The retrieved sentences are then used for abstractive summarization. Finally, the summary and the query are used for question answering. We evaluate the proposed model using various metrics on the CovidQA dataset for both natural language queries and keyword queries. The proposed approach achieves 0.177 P@1, 0.244 R@3 and 0.268 MRR for natural language queries, 0.162 P@1, 0.236 R@3 and 0.258 MRR for keyword queries. The proposed model outperforms various unsupervised model baselines in both types of queries for most of the metrics.

Keywords: BioSentVec, BM25, CORD-19, COVID-19, Information retrieval, PageRank, Question-answering, Summarization

I. INTRODUCTION

Due to its high transmission rate, novel coronavirus (COVID-19) has resulted in a pandemic in a very short span. Due to the combined efforts of various organizations, we are continuously getting precise information on various important topics related to COVID. Due to the mutating nature of the virus, it also becomes vital to have updated information. A lot of research is being conducted in order to understand its causes, impact, and precautions to be taken to avoid it. This has resulted in a large number of research articles, and extracting information that is relevant from such a large amount of data is quite challenging. From basic queries that any normal citizen would have to queries that might help in the further development of therapeutics and vaccines, all require a system that can retrieve relevant answers. Finding pertinent information is more important as a result of the numerous false rumours and misunderstandings that people tend to disseminate fast due to ignorance. Therefore, we need a system that can find responses that are relevant to different queries.

With COVID-19 Open Research Dataset (CORD-19) release, Kaggle partnered with the Allen Institute Of for Artificial Intelligence to build up the literature review for COVID-19. CORD-19 consists of 1,000,000 articles with over 400,000 full-text articles about SARS-CoV-2, COVID-19 and related topics. Most promising Jupyter notebooks were then carefully analyzed by a team of medical doctors,

epidemiologists, medical students etc. and a literature review was obtained. A table of semi-structured nature was created with questions and answers as values. From the above-mentioned literature review, the CovidQA dataset is manually created, which consists of 130 question-documents pairs.

In this paper, we propose a framework that is capable of fetching answers to queries from a large corpus of text by retrieving important sentences from the corpus. Queries can be anything from general information regarding COVID-19 to information regarding precautions, vaccines, and therapeutics. The system also provides summaries of the retrieved sentences. Along with a summarizer, the system also provides a question-answering feature, which takes the generated summary as context and the user questions as input to provide answers.

In order to achieve this, documents from COVID-19 Open Research Dataset (CORD-19) dataset are preprocessed and split into sentences using heuristic algorithms. All the sentences are then cleaned using standard preprocessing pipelines. Okapi BM25 model is used to generate scores for each of the sentences. Along with it, BioSentVec is used to convert documents into embeddings, which are then compared with the query embeddings and scored based on cosine similarity. A combination of both these scores is used to shortlist candidate sentences. These sentences are used to construct a graph based on the cosine similarity of BioSentVec embeddings. PageRank is applied to score all the sentences. Finally, a combination of all three scores is used to generate the final rank list. The top sentences are used to generate a summary using the T5 model. Finally, the summary and query are used to find the exact answer using a BioBERT model trained on the SQuAD dataset. For the evaluation of the model, the CovidQA dataset is used. It is a question-answering dataset based on the CORD-19. The dataset contains queries along with the document and answer pairs. This can be used to evaluate the information retrieval model by evaluating the retrieval of sentences from the dataset which contains the answer. Three different metrics are used during this evaluation, Precision, Recall and Mean Reciprocal Rank.

The following summarises the format of the rest of the paper. In section II, we go through most of the recent work that has been done in the domain of feature selection. After describing the proposed work and model architecture in Section III, we demonstrate the experimental setup and results in Section IV. Finally, we conclude with the inferences obtained in Section V.

TABLE I: Summary of Literature Survey

Authors	Methodology	Remarks
(Das et al., 2020)	Information retrieval over a large corpus of documents using BioSentVec embeddings, graph community detection and PageRank.	Performed search space reduction using the network ego-splitting strategies. The final set of documents retrieved and used for question answering was not evaluated using any metrics. Question answering was evaluated using confidence score only.
(Esteva et al., 2021)	COVID-19 information retrieval on the CORD dataset using deep-learning-based models along with support for question answering and summarization.	Used both keyword-based models such as BM-25 as well as deep learning-based models like Siamese BERT embeddings for indexing.
(Tang et al., 2020)	A question answering dataset based on the COVID-19 Open Research Dataset (CORD) is developed. Evaluation of various keyword-based and transformed-based models.	Keyword-based OkapiBM25 model outperforms transformer-based models when the latter is used in an unsupervised manner. Fine-tuning over SQuAD and MS-MARCO improves performance by a significant margin.

II. LITERATURE SURVEY

An Information retrieval system using techniques such as graph community detection over similarity networks built using paper abstracts and text was attempted in (Das et al., 2020). The authors used the COVID-19 Open Research Dataset (CORD) (Wang et al., 2020) for information retrieval. Initially, BioSentVec (Chen et al., 2019) was used to create the initial graph between the documents as well as the common citation information between the two papers. Further, ego-splitting was applied to the graph to generate local clusters. BioBERT (Lee et al., 2020) embeddings were then used to map the query to certain documents and select those clusters reducing the search space. Finally, only sentences from the reduced set were used for information retrieval and question answering. A major drawback of this paper is the lack of evaluation of the proposed question-answering and information retrieval system. Apart from this, no ablation study has been performed to suggest which modules of the model contribute to the effectiveness of the approach.

(Esteva et al., 2021) performed COVID-19 information retrieval on the CORD dataset using deep-learning-based models along with support for question answering and summarization. The authors used a combination of a Siamese BERT deep learning model along with keyword-based models, which included BM25 and TF-IDF.

(Tang et al., 2020) produced CovidQA, which is a question-answering dataset built using the COVID-19 Open Research Dataset (CORD), which consists of natural language questions, their keyword versions, the possible answers from the documents of the dataset along with the document ids. Since the CovidQA dataset contains a small number of question-and-answer pairs, it cannot be used for training neural network-based models. The goal is that the dataset can be used in fine-tuning tasks as well for testing or evaluation purposes. In (Tang et al., 2020), the various baseline models were evaluated on the CovidQA dataset. BM25 performed better than all the unsupervised models, such as the BERT-based models, including vanilla BERT (Devlin et al., 2018), BioBERT (Lee et al., 2020), and SciBERT (Beltagy et al., 2019). Apart from this, the BERT-based models, both unsupervised as well as fine-

tuned ones, performed poorly when compared to the BM25 counterpart for information retrieval on the keyword based queries. This is because BERT is trained to understand the meaning behind text which is often absent in keyword queries. Two datasets, i.e. the Stanford Question-Answering Dataset (SQuAD) (Rajpurkar et al., 2016) and the Microsoft Machine Reading Comprehension Dataset (MS-MARCO) (Nguyen et al., 2016) were used to fine-tune BERT, BioBERT and T5 (Raffel et al., 2020) models for Question-Answering. T5 fine-tuned on the MS-MARCO dataset outperformed other models. A brief summary of the literature survey is shown in Table. I.

Various other question answering datasets in the biomedical domain have been proposed with BioASQ (Tsatsaronis et al., 2015) being one of them. One issue with the provided questions is that they are quite different from the tasks present in the TREC-COVID challenge.

III. METHODOLOGY

The COVID-19 Open Research Dataset (CORD) contains more than 400,000 full-text research articles. On average, documents contain hundreds of sentences. Information retrieval over search a large search space is a difficult and computationally expensive task. It is thus necessary to reduce the search space and select the sentences which are to be used to perform summarization and question-answering.

A. Search Space Reduction

To reduce the search space from the set of all sentences from all documents of the CORD corpus, we use Okapi BM25 model scores for each document as well as the BioSentVec (Chen et al., 2019) scores. Okapi BM25 is a probabilistic retrieval framework based on the bag of words retrieval that ranks documents based on the query terms appearing in them.

The BioSentVec is a deep learning-based model trained on a corpus of over 30 million clinical and bio-medical research articles from PubMed (Canese and Weis, 2013), and MIMIC-III (Johnson et al., 2016) databases, which are publicly available. The BioSentVec embedding for a sentence is a 700-dimensional vector. The BioSentVec score between a query and a sentence is the cosine similarity between the

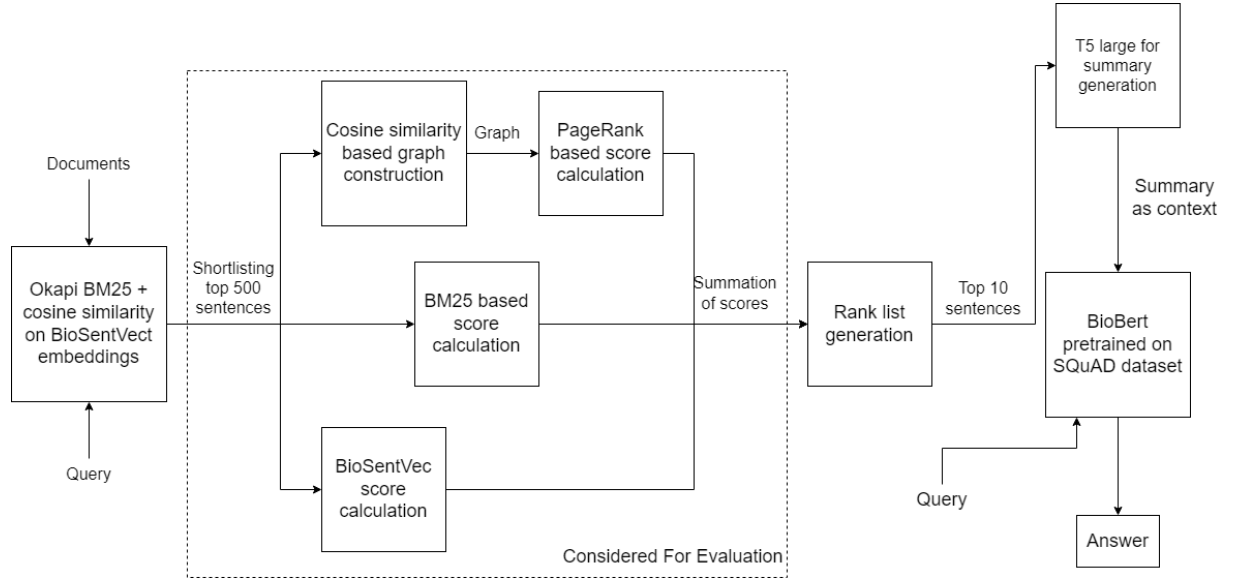


Fig. 1: Model Architecture

BioSentVec embeddings of the query and the sentence. Based on the final score for a sentence, which is the arithmetic sum of the BM25 score and the cosine similarity of the BioSentVec embeddings of the sentence and the query, 500 sentences are shortlisted.

B. Ranking using BM25, BioSentVec and PageRank

In this stage, further shortlisting of the sentences take place. A graph between the sentences is created where an edge between two sentences exists if the cosine similarity between the BioSentVec embeddings of the two sentences, as well as the cosine similarity with that of the query, are above a specific threshold. The value of the threshold should be such that neither the graph should be too dense nor too sparse. A sparse graph can lead to a biased selection of the nodes or sentences.

The PageRank (Xing and Ghorbani, 2004) algorithm on the constructed graph gives scores to each of the nodes or sentences based on how influential they are. Since the graph construction process is slow, the number of sentences on which the PageRank algorithm can be executed is limited. The arithmetic sum of the PageRank score, BM25 score and BioSentVec scores are used to rank the sentences. The top 10 sentences are shortlisted from the generated rank list.

C. Summarization and Question Answering

The T5-large model is used for abstractive text summarization. T5 (Text-to-Text transfer transformer) is a transformer-based model trained on masked language modelling. It can be used to perform various tasks such as text classification, question answering, machine translation, and abstractive summarization. T5-large is used to summarize the top 10 sentences from the previous stage in an abstractive manner to provide a short summary. Abstractive summarization produces novel sentences that are not present in the input.

For question answering, a BioBERT model pre-trained on the SQuAD (Rajpurkar et al., 2016) dataset is used. The Stanford Question Answering Dataset contains a query, context text to answer the query and exact answers. The SQuAD BioBERT model takes the input query along with a limited-sized context as the input. The generated summary from the previous stage is used as context for the BioBERT question-answering model. The output answer provided by the BioBERT model is generally short and concise. The proposed model architecture is shown in Fig. 1.

IV. EXPERIMENTAL SETUP & RESULTS

A. Dataset

For information retrieval along with question answering and summarization, the Covid-19 Open Research Dataset is used. It contains more than 400,000 research articles with all their section such as abstract, body, reference entries, etc. For evaluation, we use the CovidQA dataset (Tang et al., 2020), which is a question-answering dataset created using the COVID-19 Open Research Dataset (CORD).

The dataset comprises questions in the form of natural language queries, keyword queries, the exact answer for the queries and the document in which it occurs. All documents are from the CORD dataset. A single query may have multiple answers from different documents from the CORD corpus. The version of CORD from April 10, i.e. Round-1 of the TREC-COVID challenge, was used. The dataset contains a total of 130 question-answer pairs and 27 unique topics (or questions). The answers to the queries form a set of 85 documents from CORD. On average, each question contains 1.6 answers. The exact answer may be present in several sentences of the given document. The answer can be from the abstract, body text or even the reference entries.

1) *Preprocessing*: For obtaining sentences from the text, a heuristic algorithm is used that covers the cases of text

TABLE II: Summaries and Answers generated by T5 and BioBERT (SQuAD) for some sample queries

Query	Summary	Answer
common symptoms of covid 19	the most common symptoms of covid-19 are fever, cough and tiredness. the most common symptoms of covid-19 are fever, dry cough, and malaise/fatigue. covid screening for that visitor depends on resource availability and symptoms.	fever, cough and tiredness
how to mitigate covid transmission?	social measures for mitigating covid transmission, including social distancing, and sheltering in place, have not been examined. risk of transmission during rugby matches is very low but efforts should be made to further mitigate disease transmission within the environment.	travel bans and sheltering in place
various assistance programmes for covid 19	social assistance programmes such as cash transfers or basic income schemes should also be considered. routine surveillance programmes are planned and conducted on a yearly basis. the absence of robust cost data for handwashing programmes makes financial planning and resource allocation difficult.	social assistance programmes such as cash transfers or basic income schemes

containing abbreviations, hyphens, colons, etc. For further preprocessing, all the sentences are converted to lowercase, and additional whitespaces are removed.

For BM25, additional preprocessing is done. This includes removing normalization, lemmatization and stop word removal. Along with standard English stopwords, around 26 paper-specific stopwords like Elsevier, fig, copyright, reserved, permission, table, org, et, etc., are also considered as stopwords.

B. Baselines

The baseline models for evaluation can be categorized into two categories. The first type consists of models which consider keywords for ranking documents. The second type of models are neural-based, for example, BERT.

1) *Okapi BM25*: For the keyword-based model, we use the Okapi BM25 model (Robertson et al., 1996) implemented in (Tang et al., 2020). It is a probabilistic retrieval framework based on the bag of words retrieval that ranks documents based on the query terms appearing in them.

2) *BERT Models*: For neural network-based models, we consider vanilla BERT (Devlin et al., 2018) as well as SciBERT (Beltagy et al., 2019), and BioBERT (Lee et al., 2020). SciBERT is trained on scientific texts such as SciERC (Luan et al., 2018), which contain computer science abstracts. BioBERT, on the other hand, is trained on biomedical datasets such as MIMIC-III (Johnson et al., 2016) and PubMed (Canese and Weis, 2013). We refer to the baseline values calculated by (Tang et al., 2020).

3) *BioBERT on SQuAD*: The Bio-BERT model fine-tuned on the Stanford Question Answering Dataset (SQuAD) for the question-answering task implemented by (Tang et al., 2020) is using techniques defined by (Nogueira et al., 2020).

C. Evaluation Design

The CovidQA dataset is used for the evaluation of the proposed model on the question-answering task. Given a query and a document containing the answer to the query, the system ranks all the sentences. To be deemed correct, the

chosen sentence must contain the answer in the form of a substring. To evaluate the generated rank list, three evaluation metrics are used. These include precision, recall and mean reciprocal rank (MRR).

- **Precision@1**: For a single query, Precision@1 shows whether the top document in the generated rank list contains the answer or not. The Precision@1 score is the mean of the Precision@1 values across all queries.
- **Recall@3**: For a single query, Recall@3 is the total number of generated sentences in the top 3 of the rank list divided by the total number of sentences containing the answer. The Recall@3 score is the mean of the individual Recall@3 values across all queries.
- **Mean Reciprocal Rank (MRR)**: MRR for a query is the inverse of the rank of the first sentence containing the answer in the generated rank list. The MRR score is the mean of the individual MRR values across all queries.

Since there is large imbalance in the number of questions per topic and per document, micro-averaging is used to compute the final precision, recall and mean reciprocal rank scores.

D. Results

Table III shows the comparison of the proposed approach with the baseline models for natural language queries. Among the baseline models, BM25 gives the second-best performance. It even gives a better performance compared to its deep learning-based BERT counterparts. The best performance amongst the baseline models is achieved by the BioBERT question-answering model, trained on the SQuAD dataset.

The proposed model outperforms all baseline models for the P@1 metric. For R@3 and MRR, the proposed model outperforms all baselines except the SQuAD fine-tuned BioBERT model. The lower performance on R@3 and MRR implies that the proposed model performs well in ranking a relevant sentence at the top position of the rank list but does not perform very well in ranking many relevant sentences in the first three positions.

TABLE III: Performance on Natural Language Queries of CovidQA

Model	P@1	R@3	MRR
Random	0.012	0.034	-
BM25	0.150	0.216	0.243
BERT	0.081	0.117	0.159
SciBERT	0.040	0.056	0.099
BioBERT	0.097	0.142	0.170
BioBERT (SQuAD fine tuned)	0.161	0.403	0.336
BM25 + BioSentVec + PageRank	0.177	0.244	0.268

Table IV shows the comparison of the proposed approach with the baseline models for keyword query. The proposed model outperforms all baseline models for all metrics. Low performance of BERT based models can be because the query lacks natural language semantics around which they heavily rely on.

TABLE IV: Performance on Keyword Queries of CovidQA

Model	P@1	R@3	MRR
Random	0.012	0.034	-
BM25	0.150	0.216	0.243
BERT	0.073	0.164	0.187
SciBERT	0.024	0.064	0.094
BioBERT	0.129	0.145	0.185
BioBERT (SQuAD fine tuned)	0.056	0.093	0.135
BM25 + BioSentVec + PageRank	0.162	0.236	0.258

TABLE V: Ablation study using Natural Language Queries of CovidQA

Model	P@1	R@3	MRR
BioSentVec + PageRank	0.138	0.216	0.247
BM25 + PageRank	0.154	0.215	0.248
BM25 + BioSentVec	0.169	0.244	0.264
BM25 + BioSentVec + PageRank	0.177	0.244	0.268

Table V shows the results of the ablation study performed for natural language queries of the CovidQA dataset by

considering a combination of different subsets of scoring algorithms. Based on the observations, the final model performs optimally. A major performance boost is provided by BM25, followed by BioSentVec and then PageRank. The performance drop is maximum without BM25 and minimum without PageRank.

TABLE VI: Ablation study using Keyword Queries of CovidQA

Model	P@1	R@3	MRR
BioSentVec + PageRank	0.138	0.208	0.248
BM25 + PageRank	0.146	0.215	0.244
BM25 + BioSentVec	0.154	0.236	0.254
BM25 + BioSentVec + PageRank	0.162	0.236	0.258

Table VI shows the results of the ablation study performed for keyword queries of the CovidQA dataset by considering a combination of different subsets of scoring algorithms. The observations are exactly similar to that of the natural language queries.

V. CONCLUSIONS

In this paper, we proposed an information retrieval model using BM25, BioSentVec model and PageRank algorithm. The model outperformed the standard BM25 model as well as BERT-based models (both unsupervised and those fine-tuned on the SQuAD dataset). The search space reduction techniques used help reduce the computational complexity of the method. The ablation study found that the enhanced performance was greatly influenced by the BM25 model, followed by the BioSentVec model and then the PageRank algorithm. Further, it was observed that the performance of the proposed model was consistent with both natural language-based queries as well as keyword-based queries, which is not the case for BERT-based models. Finally, the summarization and question-answering functionalities provided reasonable answers, some of which were showcased in the paper.

REFERENCES

- Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.
- Canese, K. and Weis, S. (2013). Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).
- Chen, Q., Peng, Y., and Lu, Z. (2019). Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.
- Das, D., Katyal, Y., Verma, J., Dubey, S., Singh, A., Agarwal, K., Bhaduri, S., and Ranjan, R. (2020). Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-bert embeddings. In

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., and Socher, R. (2021). Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ digital medicine*, 4(1):1–9.
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.
- Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.
- Nogueira, R., Jiang, Z., and Lin, J. (2020). Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Robertson, S. E., Walker, S., Beaulieu, M., Gatford, M., and Payne, A. (1996). Okapi at trec-4. *Nist Special Publication Sp*, pages 73–96.
- Tang, R., Nogueira, R., Zhang, E., Gupta, N., Cam, P., Cho, K., and Lin, J. (2020). Rapidly bootstrapping a question answering dataset for covid-19. *arXiv preprint arXiv:2004.11339*.
- Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., et al. (2015). An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.
- Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., et al. (2020). Cord-19: The covid-19 open research dataset.
- Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *Proceedings. Second Annual Conference*

APPENDIX

Team15-Naveen Shenoy.pdf

ORIGINALITY REPORT

8%

SIMILARITY INDEX

5%

INTERNET SOURCES

7%

PUBLICATIONS

1%

STUDENT PAPERS

PRIMARY SOURCES

1

"Advances in Information Retrieval", Springer
Science and Business Media LLC, 2018

Publication

1%

2

www.lrec-conf.org

Internet Source

1%

3

deepai.org

Internet Source

1%

4

Sendong Zhao, Aobo Wang, Bing Qin, Fei
Wang. "Biomedical Evidence Engineering for
Data-Driven Discovery", Bioinformatics, 2022

Publication

1%

5

web.archive.org

Internet Source

1%

6

www.arxiv-vanity.com

Internet Source

1%

7

www.ifad.org

Internet Source

1%

8

Deepak Gupta, Swati Suman, Asif Ekbal.
"Hierarchical deep multi-modal network for

1%



Naveen Shenoy (191IT134)



Pratham Nayak (191IT241)



Sarthak Jain (191IT145)