REVIEW3 - QUERY EXPANSION FOR IMPROVING DOCUMENT RANKINGS FOR RELEVANT DOCUMENT RETRIEVAL

NATURAL LANGUAGE PROCESSING(E2+TE2)

TEAM ID - 4

1. HARSH BHARDWAJ (20BCE0198)
2. HARSH VARDHAN (20BCE0049)
3. MEGHA BHATTACHARYA (20BCE0793)
4. SNEHIL SINHA (20BCE2005)

APRIL 01,2023

TOPIC NAME:

Query expansion for Improving Document Rankings for Relevant Document Retrieval

ABSTRACT:

A query is a brief description written in natural language that users frequently send to an information retrieval system. Instead of considering whether the query terms are included in the question, the relevance of a page is determined by its semantics. The IR system simply looks for query terms in documents, which frequently results in term mismatches and significantly reduces performance. This method disregards the document context. The relevance of the content and query is thus the main consideration in document re-ranking.

One of the most used pre-processing methods in the field of information retrieval, query expansion aims to increase the Hit Rate of obtaining the required documents from the database for a given query. It can be difficult to narrow down the right documents from a large batch using the initial query. The efficiency of text retrieval-based search algorithms used for document re-ranking is significantly hampered by mismatches. Query expansion offers efficient techniques to deal with the term mismatch issue by reformulating the queries.

The COVID-19 pandemic is an ongoing coronavirus disease 2019 (COVID 19) epidemic that is brought on by coronavirus 2 that causes severe acute respiratory syndrome (SARS CoV 2). Global social and economic upheaval brought on by the epidemic includes the worst recession since the Great Depression.

The COVID-19 Open Scientific Dataset was created in response to the COVID-19 epidemic by the White House and a consortium of top research organisations (CORD-19). The CORD-19 database contains more than 158,000 research publications regarding COVID-19, SARS-CoV-2, and similar coronaviruses, including more than 75,000 full-text articles. We create a corpus of 5900 publications by filtering the CORD-19 dataset to only include papers relevant to COVID-19/SARS-CoV-2 that were published after December 2019.

The COVID-19 corpus will be searched using a Search Engine (SE) that takes into consideration language phenomena like synonymy and polysemy. Additionally, a web application will be developed to enhance the search function. There isn't a single programme available right now that effectively finds the papers.

The NLP application we are creating will be used to carry out document retrieval and improve document ranking. For this project, the BM25 approach, which can obtain the Document with the highest query relevance following query expansion, will be used. Using the foundation of Word2Vec embedding models and Natural Language Pre-processing (NLP) tools, we suggest a number of query expansion approaches. Exploratory data analysis (EDA) methods may also be used by the model to find hidden patterns.

Four phases make up our method: data pre-processing, model training, query expansion, and document ranking.

Keywords: **Document Re-Ranking, Ad Hoc Information Retrieval, Query Expansion, NLP Tools, Word2Vec, BM25.**

INTRODUCTION:

Information retrieval encompasses all aspects of information representation, storage, organization, and access. Users must have easy access to the information items they require, and their demands must be expressed in a way that the search engine can understand (or IR system). The translation is displayed as a list of keywords (or index phrases) that summarizes the information that the user is interested in and represents their search.

Typically, using only terms to retrieve information is not very efficient. In general information about a particular subject may be described using a number of keywords, some of which might not precisely match the user's search terms. Papers may still be helpful because they contain other terms with the same meaning even though they may not contain all of the keywords from the user query.

In the realm of text retrieval, an approach for improving search accuracy is known as query expansion, or QE. The main idea is to rephrase the first question in light of the returned information and run a second round of searches to obtain more accurate results.

Using query expansion (QE) techniques, a query is reformulated to improve retrieval performance and obtain additional relevant documents by expanding the original query with additional relevant terms and reweighting the terms in the expanded query. Query expansion techniques are widely used for improving the efficiency of textual information retrieval systems, helping to overcome vocabulary mismatch issues including words in queries with the same or related meaning.

Using query expansion (QE) techniques, a query can be extended to discover more relevant documents more quickly by adding more pertinent terms to the original search and reweighting the terms in the expanded search. Query expansion methods are frequently used to improve the efficiency of textual information retrieval systems and address issues with vocabulary mismatch. These strategies involve using words with the same or related meanings in queries.

The BM25 model will be put to the test using the COVID-19 corpus (CORD -19), which contains more than 45,000 study articles about COVID-19, SARS-CoV-2, and related coronaviruses. The BM25 function will tokenize the corpus and send it to the BM25Okapi, which will then generate results showing the top n relevant results. Finally, in order to create a generalized model that meets the needs of our clients, we will attempt to combine our query expansion model with a website.

PROBLEM STATEMENT:

The continuing global pandemic of the coronavirus disease 2019 (COVID-19) is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2). The epidemic has caused widespread societal and economic upheaval, including the worst recession since the Great Depression.

The need for in-depth research on the widespread disease has led to an abundance of research papers being made accessible online globally. Despite the fact that the websites hosting the study papers typically provide useful information, due to the vast amount of content, it is not always used. The White House and a coalition of leading academic institutions have developed the COVID-19 Open Academic Dataset in response to the COVID-19 virus. (CORD-19). There are more than 158,000 study articles in the CORD-19 database that cover

But if end users don't always know what they're looking for until they discover it, finding relevant information becomes more difficult. Even if they do, they might not always know how to ask the right questions in some circumstances. This problem served as the impetus for our endeavor. Due to this, the problem we are addressing is to create an algorithm over the COVID-19 corpus that would change the original query into a new query, accounting for linguistic phenomena, synonymy, and polysemy, in order to increase the accuracy of retrieving the required document and ranking the results in order of relevance. Our goal is to develop a Search Engine (SE) over the COVID-19 corpus that considers language phenomena, synonymy, and other synonymy, andpolysemy and transforms the original question into a new query, increasing the accuracy of retrieving the needed document with the results from the new query.

LITERATURE REVIEW:

Title: Document ranking with a pretrained sequence-to-sequence model. (Published on 14th March,2022)

Author: Nogueira, Rodrigo, Zhiying Jiang, and Jimmy Lin,

Description:

Keywords: Document Retrieval, Document Ranking, Multi-Stage Document Ranking, Sequence-to-sequence model.

This study ranks the texts using a sequence-to-sequence model that has already been trained. The strategy suggested in this study is unconventional since the most often applied technique is based on designs of encoder-only pertained transformers, such as BERT. They create "Target words"—words related to the query—using a sequence-to-sequence model. By essentially broadening the user's query, the appropriate documents are found. In the situation when the training data is substantially inferior, that is, there are less training instances, the authors of this study assert that their technique is more suited than the other alternatives.

This study puts into practise a T5-based re-ranking method. This sequence-to-sequence model pre-trains its encoder-decoder architecture using a masked language modelling aim that is comparable to BERT. With the document and query as input, this is trained to return either "True" or "False," with "True" being returned when the corresponding document is relevant to the user-provided question.

We only use a SoftMax on the logits of the "true" and "false" tokens when computing probability for each query-document combination at inference time (in a reranking setting). We thus reranked the texts using the probabilities assigned to the "true" token. We tried a few other approaches before settling on this one.

Other approaches, such as reranking texts after computing the softmax using the logits of all tokens and the logit of the "true" token, proved ineffectual, with nearly nil metrics for retrieval.

The MS MARCO passage and Robust04 datasets were utilised by the authors.

Result:

The main contribution of this study is the development of a novel generation-based technique for document grading task using pre-trained sequence-to-sequence models. Our models outperform a classification-based approach, particularly in a data-poor context with little training data. We attempt to provide an explanation for these results in terms of beliefs about the knowledge that a model picks up from pretraining as opposed to fine-tuning task-specific data.

Title: Limited Query Expansion with Voting Integration for Document Retrieval and Ranking(Published on 5th August,2022)

Author: Rana, Ashish, Pujit Golchha, Roni Juntunen, Andreea Coajă, Ahmed Elzamarany, Chia-Chien Hung, and Simone Paolo Ponzetto,

Description:

Keywords: Comparative Question Answering, Document Retrieval, Document Ranking, Multi-Stage Document Ranking.

The overabundance of knowledge on a certain topic that is available online and the fact that it costs so much to obtain it are the first issues the article tackles. As a result, this research suggests a strategy for rating papers for optimal information retrieval.

This is accomplished using the paper's three-step recommended technique, the LeviRANK system.

- Information retrieval
- Document ranking
- Stance prediction.

Information retrieval: This is accomplished with the use of a user-provided enlarged first query. The user's inquiry is often brief, consisting on average of two nouns and one adjective. This translates to a significant drift in the results with even a tiny modification in the query. As a result, the method in this work (BM25) is given 9 distinct requests, each of which has one word changed. Each document only appears once in our list of documents after the nine distinct searches return nine different lists of documents.

Document Ranking: The recovered documents are scored in this subtask using mono-T5 alone and mono-T5-duo-T5. The model generates a condensed single relevance score using SoftMax of the true/false label logits for each text at inference time for probability computations.

Stance Prediction: In this stage, the findings from the rated papers in the previous step are utilised to forecast the stance. This subtask is formulated as a two-stage binary classifier problem, with the first classifier separating and forecasting the documents and labels based on the input sequence, and the second classifier forecasting the attitude. Roberta-Large-MNLI is used in this step.

Result:

This three-stage process produces a more accurate rating, particularly after the Stance prediction in the third phase. The LeviRANK system ranks the document relevance task with the highest mean nDCG@5 score of 0.758, the document quality task with the second-highest nDCG@5 score of 0.744, and the stance prediction job with the second-highest Macro-F1 score of 0.301.

Title: Query Expansion in Text Information Retrieval with Local Context and Distributional Model (Published on 9th October,2019)

Author: da Silva, Fabiano Tavares, and José Everardo Bessa Maia.,

Description:

Keywords: Distributional Semantic Model, Information Retrieval, Local Context Analysis

The study provides an explanation of effective information retrieval via query expansion. The central thesis of this study is that queries should be subjected to a posterior filter based on the constrained vocabulary of these documents in a closed dataset of documents. The authors of this study state that this was created and tested using publicly accessible benchmarks, and the results were encouraging.

It begins by outlining the conventional information retrieval system and fundamental query extension strategies. Two models—the global model and the local model—are used to broadly categorise the procedures.

In the Global model, the complete corpus of documents or an external corpus is used for query expansion. While the local system heavily relies on the accuracy of the top results because the query is formulated based on some retrieved documents of the search of the original query, the most fundamental technique used in this case is the creation of a thesaurus that defines all the synonyms of the words. When the query expansion step is reached in our algorithm, these words are also searched in the documents.

In more recent times, statistical and semantic-lexical approaches have been integrated with embedding techniques. A global statistical thesaurus was built utilising the word2vec trained representation and a mix of ontology and word embedding, together with local context analysis and word embedding. In each of these instances, the publications state that the findings were improved by the embedding representation. It is a well-known occurrence that a straightforward query extension can boost recall at the expense of decreased accuracy in the retrieved documents.

This study combines the Semantic Distribution model with Local Context Analysis (LCA). This method suggests using the first results of a query to build a representation by the co-occurrence of ideas (groups of nouns), and by comparing these results to the query, discover candidates to be aggregated to the query expansion, combining local and global analysis for QE.

Result:

The authors employed three widely accessible datasets from the fields of electrical engineering, medicine, and libraries to evaluate their suggested approach. Four measures utilised by the TREC community and established in the Grainfield paradigm were employed in the evaluation of performance: The Precision-Recall curve, MRR, MAP, and BP

The work employs a number of well-known techniques for information retrieval, including VSM, BM25, WordNet, word2vec, etc. We'll utilise BM25 in our project because it produces better outcomes than the other algorithms.

Title: Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents. (Published on 3rd November 2012)

Author: Qing T. Zeng, Doug Redd, Thomas Rindflesch, and Jonathan Nebeker

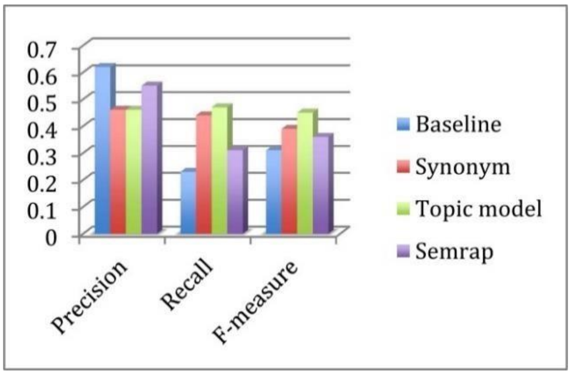
Description:

Three query expansion strategies were developed and put to the test in this investigation to retrieve clinical information. Finding relevant documents in a huge healthcare data warehouse could be challenging. We originally used a synonym expansion method using a few pre-selected vocabulary to address this issue. In order to uncover related terms for query expansion, we secondly used a vast library of clinical literature to train a topic model. Third, utilising a substantial predicate database derived from Medline abstracts, we retrieved pertinent terms for query expansion. The three expansion methods were investigated using a set of clinical notes. All three tactics achieved higher average recalls and average F-measures as compared to the baseline method. However, all expansions resulted in a drop in the average precision and precision at 10. The topic model-based strategy outperformed the other two expansion techniques in terms of recall and F-measure.

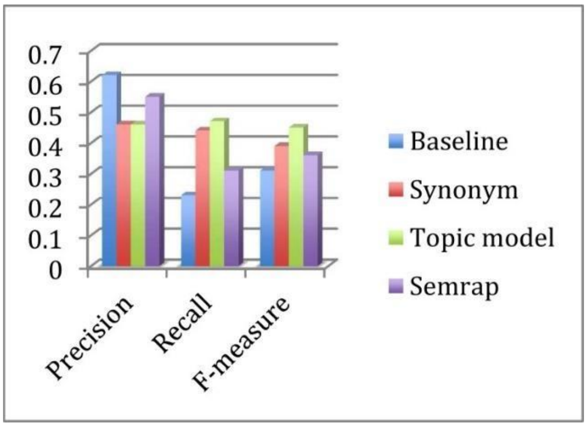
Result:

When compared to queries without expansion, all three query expansion methods exhibited a discernible improvement in the average F-measure. (Note that rather than utilising the average accuracy and recall, the average F-measure was obtained by averaging the F-measures of each of the queries.) This improvement in performance as a whole was brought about by a substantially greater recall rate because all methods had poorer accuracy. The topic-model expansion, the synonym expansion, and SemRep all generated the highest recall and F-measure. The least accuracy loss was evident in SemRep.

Treating Possibly Relevant as Relevant



Treating Possibly Relevant as Irrelevant



Title: Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval (Published on 2nd March,2014)

Author: A. R. Rivas, E. L. Iglesias, and L. Borrajo

Description:

The primary objective of information retrieval is to locate documents from a large document collection whose content matches a user query. Since the majority of users find it difficult to develop effective searches, query expansion is necessary to retrieve relevant information. To improve the efficiency of textual information retrieval systems, techniques for extending queries are commonly utilised. These methods help to get over the problems caused by vocabulary mismatch by reweighing the terms in the expanded question and adding more pertinent terms to the initial inquiry. This study integrates several text preparation and query extension strategies to improve the quality of the documents that are first returned by a query in a scientific documental database.

The Cystic Fibrosis corpus, a component of MEDLINE, is the information source. The efficiency attained by conventional inquiries is significantly increased by the recommended combinations of approaches, according to experimental data.

Result:

We have developed and tested preprocessing and query expansion methods for extracting material from the Cystic Fibrosis corpus of MEDLINE records, a corpus of biomedical journals. We assess the importance of stemming and stopwords in the production of texts and inquiries based on the study of prior writers.

Studies comparing the weighting algorithms Okapi BM25 and TF-IDF available in the Lemur tool showed that the latter delivers superior results when applying the formula supplied by the BM25 approximation.

Using all three fields—Abstract, MeSH, and Title—as opposed to just one of them—seems to be more effective when searching for documents. Additionally, the provision of pertinent feedback, a technique that is commonly used by researchers in this field, considerably improves the retrieval of scientific content. It is feasible to achieve good results while boosting MAP and other metrics thanks to the Rocchio algorithm. We undertake research to better search expanding queries utilising MeSH keywords. We have upgraded searches that find Entry words in them and acquire MeSH Headings from PubMed in order to extend the original query and map it with the documents. The results are good when compared to baseline methods, and they improve the list of retrieved articles by employing the Title, Abstract, and MeSH variables.

In this work, a rather simple procedure was used to calculate the values for the BM25 parameters. Tuning the BM25 free parameters presents a complex and computationally expensive problem that requires advanced multidimensional optimization methodologies. More advanced parameterization approaches can be used to improve retrieval accuracy.

Title: Document Re-Ranking Model for Machine-Reading and Comprehension (Published on 27th October 2020)

Author: Ben He and Iadh Ounis,

Description:

Recently, the performance of machine-reading and comprehension (MRC) systems has significantly improved. High-performance text retrieval models are required by MRC systems due to the requirement that text passages containing answer phrases be prepared beforehand. To improve the efficiency of the text retrieval models powering MRC systems, we provide a re-ranking model based on artificial neural networks that consists of a query encoder, a passage encoder, a phrase modelling layer, an attention layer, and a similarity network. The proposed approach learns the degrees of links between questions and textual passages using dot products between words that make up questions and passages. In tests using the MS-MARCO dataset, the recommended model outperformed most of the preceding models with mean reciprocal ranks (MRRs) of 0.8%p-13.2%p.

Result:

The end result is BM25, which improves the usability of MRC models. The proposed model consisted of five subnetworks: a phrase modelling layer, an attention layer, a similarity network, a query encoder, and a passage encoder. By calculating the mutual information of the phrase unit for questions and passages, it was possible to reflect the passage scores for inquiries effectively. In trials utilising the MS-MARCO dataset, the recommended model showed better MRRs, 0.8%p-13.2%p, than the preceding models. We conclude that these efficiencies are essential engineering components in the development of a functional MRC system for several concurrent users.

Title: Query Expansion for Effective Retrieval Results of Hindi–English Cross-Lingual IR (Published on 8th April 2019)

Author: Youngjin Jang 1 and Harksoo Kim 2,

Description:

Information retrieval is the study of finding documents or subdocuments inside a database or collection of information (IR). The process of obtaining information in a language other than the query language is known as cross-lingual information retrieval (CLIR). Monolingual information retrieval is the process of getting data in the query language. Low retrieval performance caused by query mismatching, multiple representations of query words, and untranslated query phrases is one of the biggest problems with CLIR compared to monolingual IR.

Query expansion is the practise or method of extending the original inquiry in order to formulate a new query (QE). The translated queries were expanded using term selection value after the texts were sorted using Okapi BM25.

Result:

The main problem with CLIR is poor performance brought on by incorrectly matched query phrases, untranslated query words, various query term formats, etc. QE aids in addressing the problem of ambiguity in CLIR by incorporating the proper terms in a query to get better results. The QE strategy improves a search engine's efficacy as well. The primary focus of this study is on the QE term selection. The mean average accuracy of the Hindi-English CLIR must be increased by broadening the user's inquiries through word selection.

Title: Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval (Published on 2nd March 2014)

Author: A. R. Rivas, E. L. Iglesias,* and L. Borrajo,

Description:

Information retrieval's primary objective is to locate documents from a huge document collection whose content matches a user query. In order to extract relevant information, query expansion is necessary because the majority of users find it difficult to develop well-designed searches. Textual information retrieval systems commonly make use of techniques for widening queries to improve their performance. These techniques aid in resolving vocabulary mismatch issues by re-weighting the terms in the expanded question and adding more pertinent terms to the original inquiry. Various text preparation and query extension strategies are incorporated in this study to improve the quality of the documents that are originally returned by a query in a scientific document database. The suggested combinations of tactics greatly boost the efficiency gained by ordinary enquiries ,according to experimental results.

Result:

Studies comparing the weighting algorithms Okapi BM25 and TF-IDF accessible in the Lemur tool show that TF-IDF using the tf formula provided by the BM25 approximation is preferable in its results.

Using all three fields—Abstract, MeSH, and Title—as opposed to just one of them—seems to be more effective when searching for documents. Additionally, the provision of pertinent feedback, a technique that is commonly used by researchers in this field, considerably improves the retrieval of scientific content.

Tuning the BM25 free parameters is a difficult and expensive computational job that requires advanced multidimensional optimization methodologies (k 1, b, and k 3). More advanced parameterization approaches can be used to improve retrieval accuracy.

Title: Studying Query Expansion Effectiveness (Published on 19th October 2009)

Author: Ganesh Chandra & Sanjay K. Dwivedi,

Description:

Through query expansion, the retrieval efficiency for ad hoc retrieval may be improved. However, query expansion mistakes might also occur, which would worsen retrieval efficiency. In this study, we aim to expand our understanding of query expansion by performing an empirical inquiry on the variables that might impact query expansion and how these variables affect query expansion. We look at the connection between query quality as measured by first-pass retrieval performance and the efficiency of query expansion. Results only show a weak correlation between the two, indicating that first-pass retrieval has little impact on query expansion effectiveness.

The results also imply that the feedback materials need to be sincere about the topic in addition to being relevant.

Result:

We investigate the two factors—poor query quality and subject drift—that could have prevented query expansion from being successful. Our experimental results show a weak association between query expansion efficiency and query quality as measured by retrieval performance on the first run. When the feedback is of actual worth and its significance is understood, the feedback papers should also exhibit a strong interest in the topic.

Our findings provide several directions for more research. If there is a lot of interest in the topic, we might be able to utilise the Entropy measure, for example, to choose appropriate feedback documents for query expansion. By looking at how query terms and possible expansion terms co-occur in sentences when the query terms are used frequently, we may also discover appropriate expansion terms.

Title: A contemporary combined approach for query expansion (Published on 3rd July 2020)

Author: Dilip Kumar Sharma, Rajendra Pamula and D. S. Chauhan ,

Description:

This paper offers a technique for choosing the optimal phrases for query improvement. First, baseline methods to query expansion are evaluated in relation to the effects of abbreviation resolution, lexical variation, synonyms, n-gram pseudo-relevance feedback, and co-occurrence technique. The Okapi BM25 algorithm's application for ranking has been described in this work. Concept words were previously handled through concept-based normalisation. results that are better than the baseline strategy. For the purpose of improving queries, a novel combination approach that combines lexical variety, synonyms, and n-gram pseudo relevance feedback has been mentioned.

Result:

The strategies described in this research, which are based on Query Expansion methods, greatly improve IR performance for small datasets. However, the improvement is negligible for big datasets. Thus, it will be necessary to take into account further massive datasets in the future. The suggested approaches do not take into account abbreviation for query expansion. One noteworthy aspect of the research described in this paper is that it does not take into account any such external environmental aspects, such as attitudes or the user's state of mind while retrieving documents.

Title: A Hybrid Model of Query Expansion using Word2Vec (Published on 18th December 2021)

Author: Abhishek Kumar Shukla; Sujoy Das,

Description:

Some of the techniques for query expansion are mentioned in this publication. For document retrieval from a vast collection of documents, pseudo relevance feedback is not particularly effective. In the suggested approach, a hybrid way of query expansion employing the word embedding technique is used to minimise the vocabulary mismatch. To forecast the expansion terms, the suggested method employs both Word2Vec and a local technique. Additionally, the suggested model is contrasted with the initial inquiry and the BM25 model.

Result:

In this work, a Query Expansion method based on Word2Vec embedding models and Natural

Language Pre-processing methods is suggested.

Title: An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search (BMC) (Published on 2nd March 2021)

Author: **Ganesh Chandra & Sanjay K. Dwivedi,**

Description:

The retrieval approach discussed in this research uses extended word and co-word analysis, together with Cuckoo Search to improve retrieval function parameters. It mostly focuses on how to find biomedical papers that discuss therapies in their abstracts. The BM25 algorithm is used by the methods described in this study to determine the abstract score. The scores of enlarged words and co-words are computed to create a composite retrieval function, which is then optimised using the Cuckoo Search, according to an enhanced version of BM25 that is described.

Result:

This research suggests a BM25-based technique for retrieving biomedical publications that includes co-word analysis. Combining the co-word score with the gene appearance weight to improve the BM25 algorithm and use it to calculate the score of enlarged words. The evaluation functions are optimised using the Cuckoo Search Algorithm. According to optimization results, weighting the "word list" more heavily might significantly raise the ranking of the connected texts. There has been less emphasis placed on embracing linked data since the query extension utilised in this research is straightforward.

Title: Query Expansion Based on NLP and Word Embeddings (Published on 15th November 2018)

Author: **Saeid Balaneshin-kordan, Alexander Kotov**

Description:

One of the crucial information retrieval techniques known as "query extension" is adding new relevant terms to the initial query in order to more precisely find applicable documents.

Various query expansion strategies are covered in this work. These methods rely on Natural Language Pre-processing (NLP) tools and Word2Vec embedding models. Using the names of TREC topics, we select words that are semantically related to the query.

The first four processes are data pre-processing, model training, query expansion, and document ranking.

Result:

They proposed a Query Expansion approach in this study that is based on Word2Vec embedding models and Natural Language Pre-processing techniques. Overall results are in accordance with their expectations.

80% of our themes for our best run outperformed the TREC median scores, according to the data. We trained Word2Vec models on the whole TREC Washington Post Corpus for our experiments.

Future study will look into query-specific local Word2Vec model training, since earlier experiments have produced good results.

Title: Embedding-based Query Expansion for Weighted Sequential Dependence Retrieval Model (Published on 8th August 2018)

Author: **Billel Aklouche, Ibrahim Bounhas and Yahya Slimani**

Description:

It is known how to effectively account for term dependencies in query expansion techniques based on pseudo-relevance feedback (PRF) for retrieval models of this kind, despite the fact that term dependencies are taken into account by information retrieval models based on Markov Random Fields (MRF), such as Sequential Dependence Model and Weighted Sequential Dependence Model (WSDM). In this paper, we propose Semantic Weighted Dependence Model (SWDM), a PRF-based query expansion method for WSDM, using distributed low-dimensional word representations (i.e., word embeddings). By applying this technique, the closest unigrams to each query word are identified in the top documents and embedding space and then immediately added to the WSDM retrieval function.

Research employing the TREC datasets demonstrates that SWDM statistically substantially outperforms state-of-the-art MRF retrieval models, PRF techniques for MRF retrieval models, and embedding based query expansion methods for bag of words retrieval models.

Result:

In this paper, the authors describe the Semantic Weighted Dependence Model, which addresses the vocabulary gap in the Weighted Sequential Dependence Model by using distributed word representations (also known as word embeddings) in two distinct methods. Word embeddings are one method that distributional similarity is computed to find keywords that are semantically related to query terms for query expansion. However, as features, they are used to assess the importance of query ideas. We also provide an augmentation of SWDM that incorporates words from the most frequently searched articles together with phrases that are semantically related.

Title: An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search (Published in 2020)

Author: **Zicheng Zhang,**

Description:

The extraction of gene and disease information from a big library of scholarly papers to provide doctors with clinical decision assistance is one of the primary study topics in precision medicine. In addition to Cuckoo Search, the novel article retrieval approach being proposed incorporates extended word and co-word analysis to improve retrieval function parameters. Finding biological papers with medicines mentioned in their abstracts is the main goal. The methods in this article calculate the abstract score using the BM25 algorithm. We do, however, provide a more effective version of BM25 that computes the expanded word and co-word scores to produce a composite retrieval function that is then optimised with the Cuckoo Search.

The recommended method looks for details on both illnesses and genes in a single biomedical paper's abstract. The goal of doing this is to raise article relevancy and, consequently, score. We also examine how different characteristics affect the retrieval process and talk about how they meet diverse retrieval needs.

Result:

Co-word implementation and Cuckoo Search are both used in the recommended method, which is an improved version of the BM25 algorithm and has been demonstrated to yield superior results on various testing sets.

This paper also employs a rather simple query expansion approach. Future study will focus on ontology and semantic networks to expand the query vocabulary.

Comparison Table:

Paper Title/ Journal Details	Method / Algorithm	Challenges	Observations
Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents	**Synonym Expansion, Topic Modelling, Predication-based Expansion, Document Ranking,BM-25**	• Failed to assess the techniques for expansion using the TREC Medical Records 2011 data. • A lower volume of papers • Our hypothesis is that using word or concept characteristics will raise the subject model's calibre.	The F-measure and recall were successfully increased by the Query expansion methods, but accuracy was decreased. The increases of 8 to 24 percentage points in recall are more noteworthy, even if the gains of 5 to 14 percentage points in F-measure were not insignificant. The subject model-based expansion outperformed the other two expansion techniques in terms of recall and F-measure. The method with the lowest recall and highest precision was the baseline approach.
Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval	**Okapi BM25 Weighting Algorithm, Query Expansion, Pseudo relevance Feedback, Use of mesh to Expand Queries**	• Lower retrieval accuracy as a result of less sophisticated parametrization techniques • Value approximation	It appears more efficient to get documents based on the Abstract, mesh, and Title fields than than looking at each of these fields separately. Additionally, the retrieval of scientific materials is greatly enhanced by the use of relevant feedback, a method that is frequently employed by researchers in this field. The Rocchio algorithm makes it possible to get decent outcomes while enhancing MAP and other metrics.
Document Re-Ranking Model for Machine-Reading and Comprehension	**Passage re-ranking; passage retrieval; machine-reading comprehension,ANN -based model,BM-25,MRC model**	• Bert models occasionally perform better than the suggested model • Extensive computational work	The suggested model outperformed the other models in the trials using the MS-MARCO dataset, 0.8%p-13.2%p, with the exception of the BERT-based models. Although the suggested model showed fewer mrrs than the BERT-based models, it showed to be much more efficient than the latter in terms of memory consumption and reaction time (about 8 times less memory usage and around 3.7 times faster response time). We come to the conclusion that these efficiencies are crucial engineering elements in the creation of a workable MRC system for many concurrent users.
Query Expansion for Effective Retrieval Results of Hindi–English Cross-Lingual IR	**Cross lingual information retrieval (CLIR) , Query expansion (QE) , Okapi BM25**	• Poor retrieval performance as a result of query mismatches, various query term formats,	By include the proper terms in a query, QE helps CLIR overcome the problem of ambiguity and produce better results.

		and untranslated query phrases.	
Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval	**Information retrieval (IR) , Query expansion (QE) , Indexing , Matching , Stemming**	• Produce well-designed searches, deal with vocabulary mismatch issues, and deal with the computationally expensive tweaking of the BM25 free parameters	The results of TF-IDF utilising the tf formula provided by the BM25 approximation are superior, and employing all three of the fields—Abstract, MeSH, and Title—appears to be more effective than using just one of them.
Studying Query Expansion Effectiveness	**Query expansion (QE) , Information retrieval (IR) , Okapi BM25**	• Poor query quality and subject drift; by expanding the query, the retrieval performance for ad hoc retrieval can be improved. However, query expansion mistakes might also occur, which would worsen retrieval efficiency.	When the feedback is of actual value and its significance is understood, the documentation should show a keen interest in the topic. The first-pass retrieval performance, which measures the quality of the query, has a moderate correlation with the efficiency of query expansion.
A contemporary combined approach for query expansion	**Okapi BM25 algorithm, Concept-based normalization, n-gram pseudo relevance feedback for query enhancement**	• For small datasets, this QE method improves IR performance; but, for big datasets, the improvement is negligible.	It doesn't take into account any such external environmental aspects like feelings or the user's attitude while retrieving documents.
A Hybrid Model of Query Expansion using Word2Vec	**Tagging and lemmatization query, Proposing replacement for each lemma in query, Paraphrasing queries, Probability of a paraphrase, Retrieving documents for each query**	• Selecting the wrong keywords by the user might lead to subpar results.	A decision graph-based analysis reveals factors that affect retrieval performance; a query reduction based on these factors greatly improves retrieval performance; and a query expansion followed by a query reduction produces even more significant gains in retrieval performance.

An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search	**BM25 algorithm**	• Low linked data availability	The ranking of the associated texts is successfully improved by raising the "word listscore "'s weight.
Document ranking with a pretrained sequence-to-sequence model	**Uses a pretrained Sequence to Sequence model known as T5 in order to re-rank the documents**	• There are several sequence-to-sequence models that may be utilised, such as T5, yielding varying accuracy in various situations.	When compared to other algorithms like BERT, it performs well when dealing with smaller datasets.
Limited Query Expansion with Voting Integration for Document Retrieval and Ranking	**The paper proposes a three step approach containing retrieval ranking and stance prediction for the query.** **Uses BM25, monoT5, monoT5-DuoT5 for step1 and Step2 where as pretrained RoBERTA-Large-MNLI is used in step3.**	• The usage of numerous algorithms or the repetition of a single method throughout the whole process adds to its complexity.	The last phase, posture prediction, is a highly helpful information retrieval feature that was skillfully implemented in this research.
Query Expansion in Text Information Retrieval with Local Context and Distributional Model	**Local Context and Distribution Model for query expansion and multiple algorithms like BM25 and VSM**	• The information retrieval technique is significant here since the original query's top results heavily influence subsequent rankings.	The suggested technique is effective for both retrieving particular information and general information.
Query Expansion Based on NLP and Word Embeddings	**Word2Vec,Okapi – BM25. Skip-Gram or CBOW models**	• The retrieval mechanism of Okapi BM25 often unnecessarily penalises extremely lengthy documents. • The method employed is to	• Surprisingly, using query reweighting generated the greatest results and greatly beat the other strategies. In every case, the Word2Vec Skip-Gram model outperformed the CBOW model. Furthermore, acquiring expansion terms by comparing similarity to the complete question generated better results than comparing similarity to individual

		express our text in a vector space so that we may locate related documents for a certain document or query, increasing the complexity of the space.	query phrases, which is consistent with our hypothesis.
Embedding-based Query Expansion for Weighted Sequential Dependence Retrieval Model	**SWDM+ outperforms EQE1+RM1**	• Costly in terms of computation, the model weakens as the amount of fine-tuning data increases, and the method is less successful than alternatives that do not make use of the entire labelled semantic dependency network	• Using word embeddings for query expansion in conjunction with a bag-of-words retrieval model enhances retrieval accuracy, just as it does when taking into account the sequential links between query terms. • Findings show that SWDM outperforms EQE1 in terms of retrieval accuracy because it takes into consideration dependencies between expansion terms as well as relationships between expansion terms and query terms. • EQE1+RM1 is outperformed by SWDM+.
An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search	**BM-25**	• In the optimization, the cuckoo search algorithm's strength, convergence rate, and convergence accuracy may be improved. • The Okapi BM25 retrieval algorithm usually unfairly penalises extremely lengthy documents.	Medical papers presented in the Clinical Decision Support (CDS) Tracks of the 2017-2019 Text Retrieval Conference (TREC): Precision Medicine provided the data for this study. In all, there are 120 subjects. Three indicators are utilised to compare the techniques, which were selected solely based on the BM25 algorithm and its modified version, in order to perform comparable experiments. The results showed that the recommended method generates better results.

PROPOSED SYSTEM:

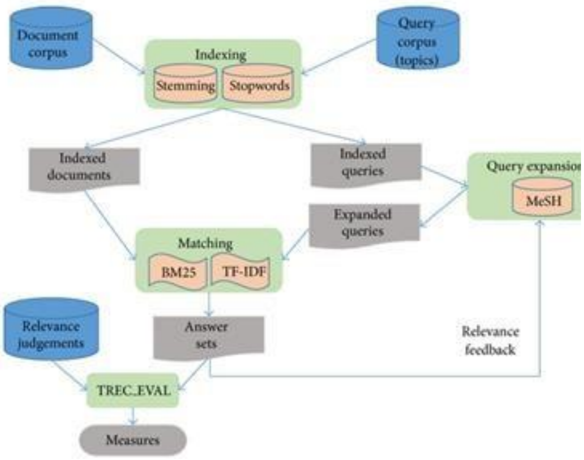
OBJECTIVES:

1) Improving document rankings for relevant document retrieval is our key goal while expanding the query.

2) Document rating for a specific query, which will aid in quickly locating the data and enhancing its quality as support for more study.

3) Attempt to link our query expansion model with a website to create a generalized model that meets the needs of our clients.

ARCHITECTURE DIAGRAM:



Okapi BM25 Weighting Algorithm

In information retrieval the basic idea is to find document that contains the keyword or sorts the documents that are relevant to what we are searching in the form of query to achieve this task we usually use something known as scoring function

Scoring function takes document from corpus and query from user and return a numeric score that can tell us how good of the match the document is for our input query

Typically, higher score means bigger match

Once we have scores for all those documents then we can rank those documents and return the results

Okapi BM25, or BM25, is a weighting function used to rank documents according to their relevance to a given query. Many researchers apply the BM25 function in different corpus to retrieve relevant documents. It searches our input query against document database.

BM stands for best match, the system where this model was first implemented was called okapi

this function is used by search engines to estimate the relevance of documents to a given query

it is based on probabilistic retrieval framework

BM25 extends the scoring function for the binary independence model to include document and query term weights

Let's say we want to search ***"novel coronavirustreatment"*** in CORD-19 corpus of documents

BM25 scoring function is simply a linear weighted combination of scores for each of the words that make up the query i.e., say ***q1, q2, q3*** be separate score values for ***novel, coronavirus, treatment*** respectively and then we add these scores (q1+q2+q3) and this gives us the final score.

So, when are taking query as input and takingdocument from document database three factors can affect the weight of each query term:

1. TF: how frequently the query appeared in the document

for example: suppose there is no ***novel*** word in the document then term frequency will be zero and the document will supposedly be irrelevant.

2. IDF: Some words are supposedly more frequent in documents as compared to others so such words must be given low weightage

for example: as the corpus we are using is CORD 19 so the word ***coronavirus*** will be more frequent so it will be given less weightage as it has low inverse document frequency but for word like ***treatment*** the weightage will be high because it has high inverse document frequency.

3.Length of the document:

Suppose a document contains 10000 words and 2 occurrences of word ***novel, treatment*** and on the other hand we have another document that contains only 1000 words but with single occurrences of word ***novel, treatment*** so the shorter document will be a better choice as the large in length documents are supposed to mention all query words but that may not actually be about the query instead, a short document that mentions all words is a much better candidate.

WORKING BM25:

BM25 is a probabilistic model, where the weight of a search term is assigned based on its frequency within the document and the frequency of the query term. The simplest score (Retrieval Status Value) for document d is just idf weighing of the query terms present in the document

$$BM25 = \sum_{t \in q} \log \left[\frac{N}{df(t)} \right]$$

Improve this formula by factoring the term frequency and document length

$$BM25 = \sum_{t \in q} \log \left[\frac{N}{df(t)} \right] \cdot \frac{(k_1 + 1) \cdot tf(t, d)}{k_1 \cdot \left[(1 - b) + b \cdot \frac{dl(d)}{dl_{avg}} \right] + tf(t, d)}$$

tf --> term frequency document d

dl(d) –> length of document d

There are some variations of the scoring function for BM25, the most common form is

$$\sum_{i \in Q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1) f_i}{K + f_i} \cdot \frac{(k_2 + 1) q f_i}{k_2 + q f_i}$$

- ri is the number of relevant documents containing term I,
- ni is the number of documents containing term i
- N is the total number of documents in the collection
- R is the number of relevant documents for this query
- r and R are set of 0 if there is no relevant information
- fi is the frequency of term I in the document
- qfi is the frequency of term I in the query;
- k1, k2, and K are parameters whosevalues are set empirically

K is a more complicated parameter that normalizes the tf component by document length Specifically

$$K = K1(\underline{(1-b)} + \underline{b\ .L_d\ /L_{ave}})$$

the constant b regulates the impact of the length normalization where b=0 corresponds to no length normalization,

b=1 is full normalization

After the scores are generated by this function

From here on we don’t get whether document is relevant or not

we just generate scores for all documents, sort all these scores in descending order and thus we obtain ranks of documents. The document with highest score will be highly relevant to this query.

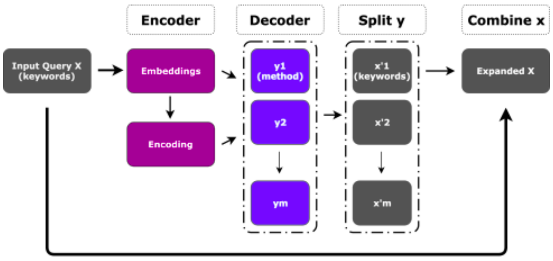
This is how bm25 works in ranking the documents

Query Expansion:

Document, corpus (all source codes), and di (one code), Vk = Vocabulary MT (all method names tokenized at camelCase and snake case), Vm = Vocabulary M (all method names in D), Set of keywords from Vk called X, and a ranked list of documents called R (di)

As a result, the system returns R for a given query X and lets dx be the anticipated ground truth document. As a consequence, we define rank as (dx, R), where a lower rank for dx simply indicates a better search result. Consider rank as the priority order of the search results, where more priority is indicated by a lower rank. Determining that the rank of the anticipated document dx with an enlarged query (E) should be lower than the original query(O) is the purpose of the expansion system.

The authors' proposed NQE model is an encoder-decoder that accepts a query X as input and outputs a list of methods with names Y = y1,..., my, where each yi Vm. Once we have a list of method names, they use the split functions camelCase and snake case to find keywords for an extended query (Xexp).



As shown in the above picture, we use a trainable embedding layer to obtain dense representations of each keyword for each input X. To obtain a representation of the complete X, we add up all of the embeddings at the column level. This thought vector then traverses a stack of GRU units that are stacked horizontally, producing a softmax distribution across method names at each time step until the sequence token is formed.

The Rocchio formulation for feedback on pseudo relevance is implemented in the Lemur toolbox that we employed in our studies. The standard retrieval model is used to obtain the first m documents for the supplied query q. The extended query q' is calculated as the following given the retrieved documents and the initial query

$$q' = q + \frac{\alpha}{M}\sum_{i=1}^{M} d_i,$$

where M is the total number of documents found after running the specified query and is the parameter used to assess the significance of the documents found.

Proposed system:

Prior to **indexing**, documents and queries are processed to identify keywords (also known as terms; pertinent phrases that may be used in the query). In order to condense related words to their stem, base, or root form at this stage, it is crucial to take into account the use of stemming and stopword lists. Affix removal is used to eliminate words that do not include information pertinent to the document and adapt many derivational or inflectional versions of the same term to a single indexing form.

By weighting phrases, the process of **matching** determines how similar documents and queries are, with the BM25 algorithm being the most widely used method. When a retrieval system receives a query, it typically responds with a ranked document list, with the documents that the system considers to be most comparable to the query appearing first.

Different query expansion strategies might be used once the initial response set has been received. For instance, the query may be modified to include the most pertinent terms from the top papers that were previously obtained in order to rerank the documents. We call this procedure relevance feedback. By changing the queries' wording and utilising additional keywords that are more indicative of the document's content, the retrieval may be improved still more.

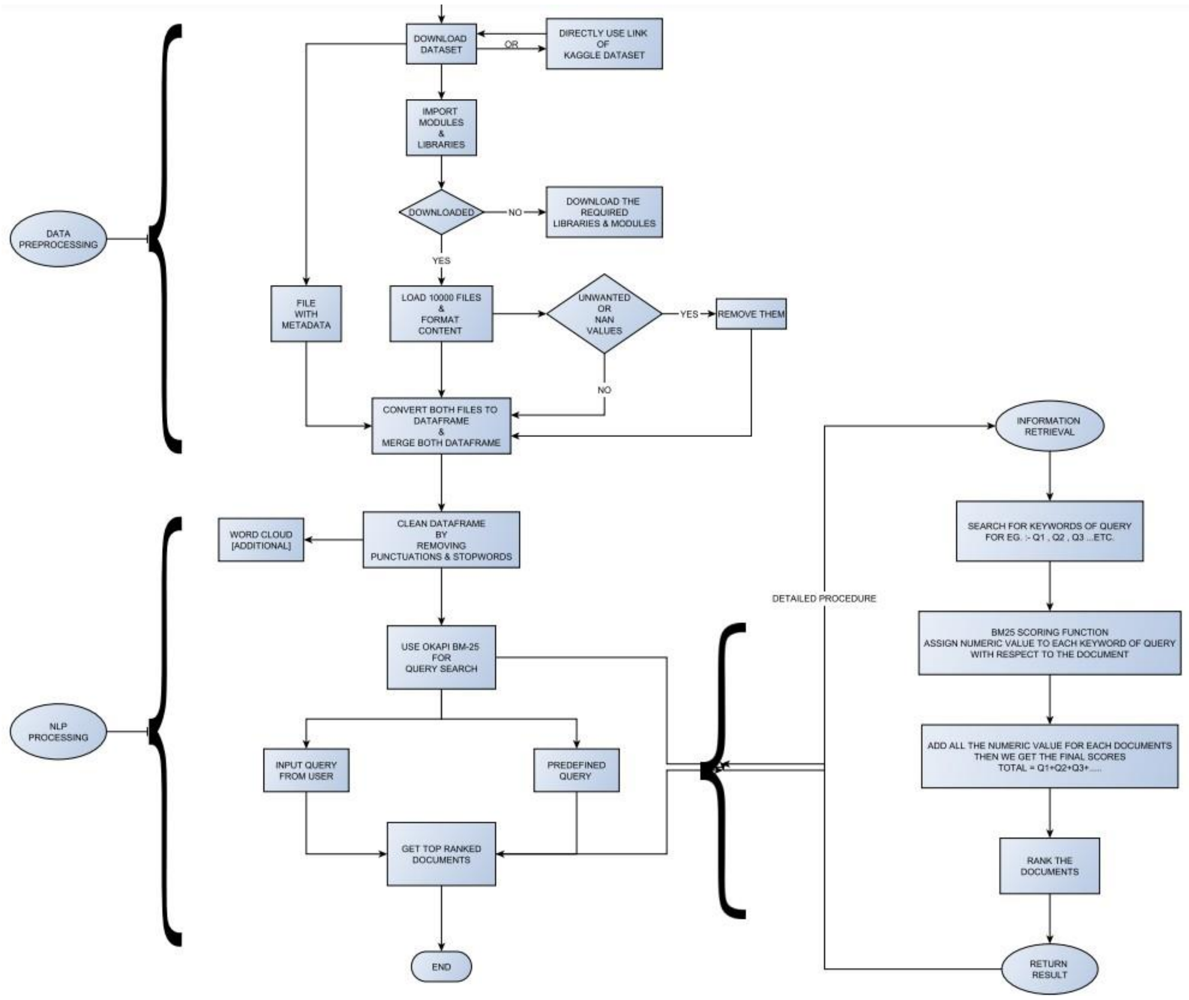
Stemming is the process of removing affixes from words to reduce them to their stem, base, or root form. To convert several derivational or inflectional spellings of the same word into a single indexing form is its goal.

Stopwords: A document is indexed by the frequency of its terms in information retrieval. This procedure' statistical analysis reveals that some terms are used less frequently than others. For instance, and, of, and the are commonly used in publications that lack important details. Stopwords are the name given to this group of words. Stopwords can be eliminated to greatly reduce indexing structure size, speed up calculation, and improve accuracy.

Tokenization: Tokenization techniques will be used to turn a document's body content (or abstract) into keywords.

The Okapi BM-25 indexing method and query expansion may then be used to obtain the desired result.

FLOW DIAGRAM:

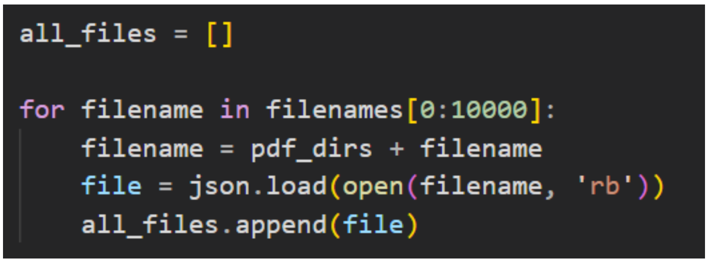

DOWNLOAD DATASET
OR
DIRECTLY USE LINK OF KAGGLE DATASET
IMPORT MODULES & LIBRARIES
DOWNLOADED
NO
DOWNLOAD THE REQUIRED LIBRARIES & MODULES
YES
DATA PREPROCESSING
FILE WITH METADATA
LOAD 10000 FILES & FORMAT CONTENT
UNWANTED OR NAN VALUES
YES
REMOVE THEM
NO
CONVERT BOTH FILES TO DATAFRAME & MERGE BOTH DATAFRAME
INFORMATION RETRIEVAL
WORD CLOUD [ADDITIONAL]
CLEAN DATAFRAME BY REMOVING PUNCTUATIONS & STOPWORDS
SEARCH FOR KEYWORDS OF QUERY FOR EG. :- Q1 , Q2 , Q3 ...ETC.
DETAILED PROCEDURE
USE OKAPI BM-25 FOR QUERY SEARCH
BM25 SCORING FUNCTION ASSIGN NUMERIC VALUE TO EACH KEYWORD OF QUERY WITH RESPECT TO THE DOCUMENT
NLP PROCESSING
INPUT QUERY FROM USER
PREDEFINED QUERY
ADD ALL THE NUMERIC VALUE FOR EACH DOCUMENTS THEN WE GET THE FINAL SCORES TOTAL = Q1+Q2+Q3+.....
GET TOP RANKED DOCUMENTS
RANK THE DOCUMENTS
END
RETURN RESULT

- START
- Start with downloading the dataset or using the link to it directly.
- Next, we will import or download the libraries and modules required for data preprocessing.
- Out of all the files we will take a small chunk of data files and process it by removing the unwanted and NAN values from it.
- Next, we will take the file with the metadata and the above formatted datafile, convert both of them to dataframes i.e., csv format files and then merge both of them.
- Further we continue with the NLP processing by cleaning the above merged dataframe by removing the punctuations and stopwords.
- Additionally, we can show the keywords in the form of Wordcloud if wanted.
- Next, we will use the Okapi-BM-25 Model in order to perform the search query which can epredefined or user input.
- Output will consist of Top 10 documents related to the searched query.
- END

PSEUDOCODE:

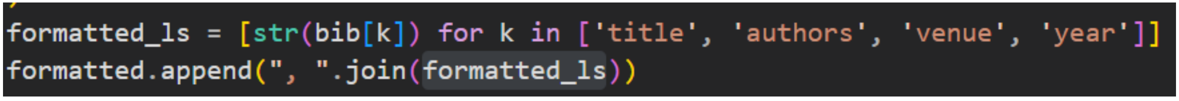
For NLP part:

Step1:



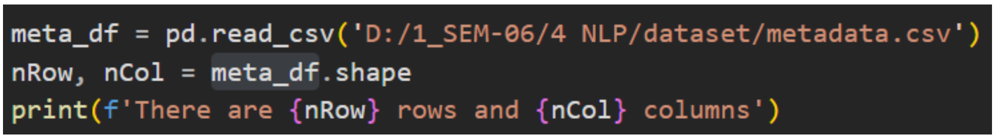
```
all_files = []

for filename in filenames[0:10000]:
    filename = pdf_dirs + filename
    file = json.load(open(filename, 'rb'))
    all_files.append(file)
```

Step2:



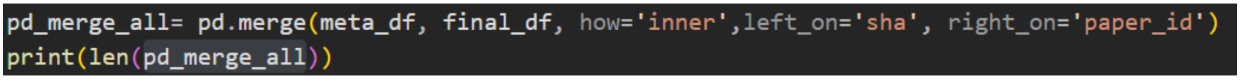
```
formatted_ls = [str(bib[k]) for k in ['title', 'authors', 'venue', 'year']]
formatted.append(", ".join(formatted_ls))
```

Step3:



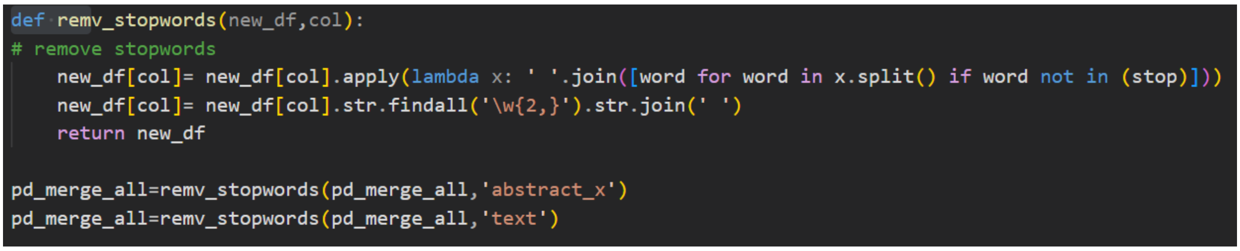
```
meta_df = pd.read_csv('D:/1_SEM-06/4 NLP/dataset/metadata.csv')
nRow, nCol = meta_df.shape
print(f'There are {nRow} rows and {nCol} columns')
```

Step4:



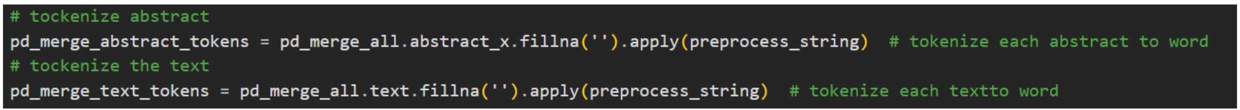
```
pd_merge_all= pd.merge(meta_df, final_df, how='inner',left_on='sha', right_on='paper_id')
print(len(pd_merge_all))
```

Step5:

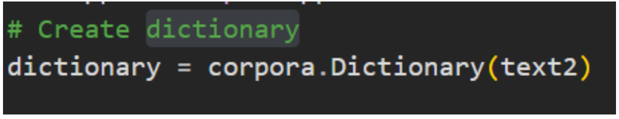


```
def remv_stopwords(new_df,col):
# remove stopwords
    new_df[col]= new_df[col].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))
    new_df[col]= new_df[col].str.findall('\w{2,}').str.join(' ')
    return new_df

pd_merge_all=remv_stopwords(pd_merge_all,'abstract_x')
pd_merge_all=remv_stopwords(pd_merge_all,'text')
```

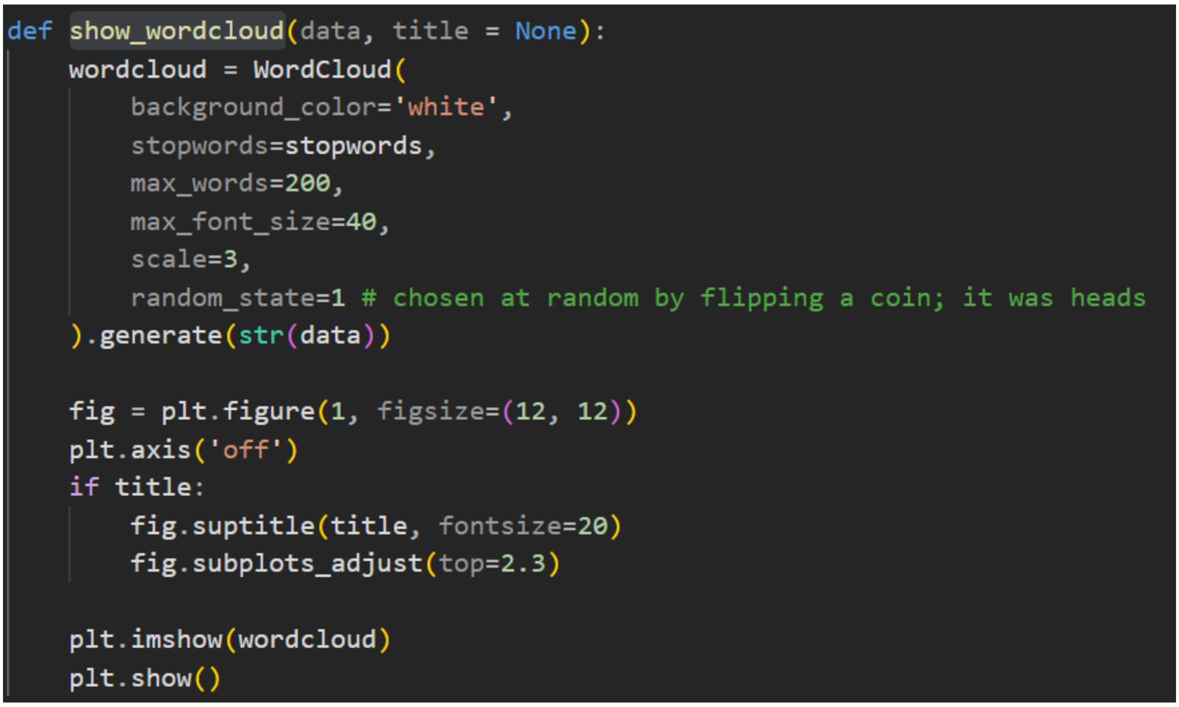
Step6:



```
# tockenize abstract
pd_merge_abstract_tokens = pd_merge_all.abstract_x.fillna('').apply(preprocess_string)  # tokenize each abstract to word
# tockenize the text
pd_merge_text_tokens = pd_merge_all.text.fillna('').apply(preprocess_string)  # tokenize each textto word
```

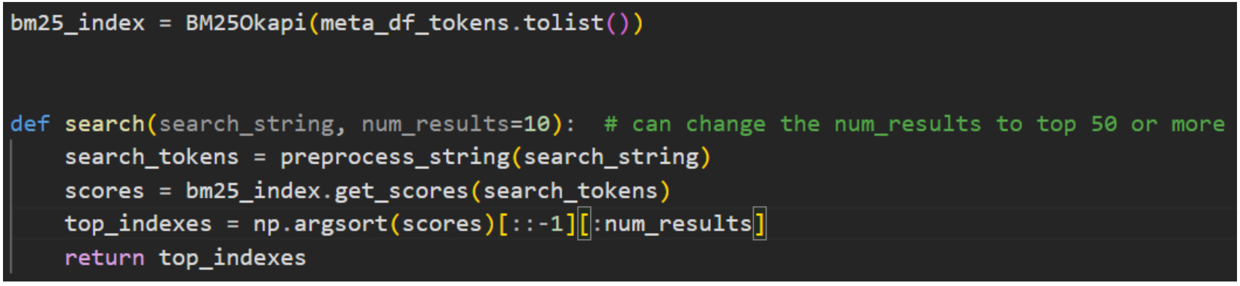


```
# Create dictionary
dictionary = corpora.Dictionary(text2)
```



```
def show_wordcloud(data, title = None):
    wordcloud = WordCloud(
        background_color='white',
        stopwords=stopwords,
        max_words=200,
        max_font_size=40,
        scale=3,
        random_state=1 # chosen at random by flipping a coin; it was heads
    ).generate(str(data))

    fig = plt.figure(1, figsize=(12, 12))
    plt.axis('off')
    if title:
        fig.suptitle(title, fontsize=20)
        fig.subplots_adjust(top=2.3)

    plt.imshow(wordcloud)
    plt.show()
```

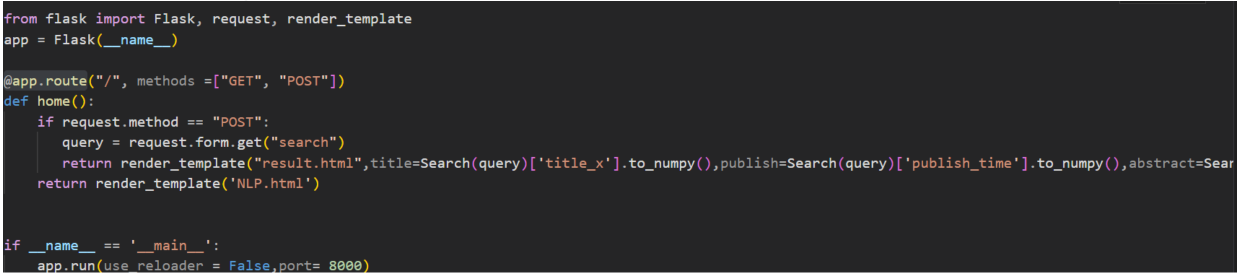
Step7:



```
bm25_index = BM25Okapi(meta_df_tokens.tolist())


def search(search_string, num_results=10):  # can change the num_results to top 50 or more
    search_tokens = preprocess_string(search_string)
    scores = bm25_index.get_scores(search_tokens)
    top_indexes = np.argsort(scores)[::-1][:num_results]
    return top_indexes
```

Step8: RUN the program to get the result.

For connecting with the website:



```
from flask import Flask, request, render_template
app = Flask(__name__)

@app.route("/", methods =["GET", "POST"])
def home():
    if request.method == "POST":
       query = request.form.get("search")
       return render_template("result.html",title=Search(query)['title_x'].to_numpy(),publish=Search(query)['publish_time'].to_numpy(),abstract=Sear
    return render_template('NLP.html')


if __name__ == '__main__':
    app.run(use_reloader = False,port= 8000)
```

NOVELTY IN OUR PROJECT / FEASIBILITY:

To add uniqueness to our model, we plan to use the JavaScript framework React.js to build a website that houses all metadata or documents related to a particular topic on the backend. Users can enter their queries in a search bar and get results once processing is complete. We will use the user's data to populate the backend, and perform pre-processing to provide relevant query results. Our process involves taking user input, rating documents based on query-related information, extracting chunks from these documents, re-ranking documents, and displaying connected queries on the same page. Additionally, we may offer query expansion results to the user, allowing them to find the necessary documents from the expansion results instead of the initial query results. In this case, they can select the appropriate result and access the relevant documents. Our system aims to enhance the user's experience by providing relevant and useful results.

Eg: Suppose initially user searched for document on covid 19 stats. So, he got some relevant documents like coivd 19 stats of world in 2020. Now the query expansion results are following [covid 19 stats of India (2020-2021), covid 19 stats of world in 2019, etc.]

Now suppose the user exactly wanted to search for the document of covid 19 stats of India in 2020 so now in query expansion results he got the exact query for which he can search for hence making his search even easier.

APPLICATIONS:

- For example, in cutting-edge research on COVID 19. An enormous amount of research has been done on this recently prevalent condition, which has resulted in inferior management of it. A query's document rating will aid in quickly finding the material and enhance its quality to support ongoing study.
- Supporting the educational sector. Numerous details about a subject may be found on research paper websites or in school notes. Going through them is a tiresome chore. By retrieving the most pertinent materials based on the query, our system will significantly aid students and teachers.

  • enhancing search results. The present search engine version's "Document searching" feature may be improved. The user-provided query is the sole thing the search engine looks for, and it gives results preferably in the order of popularity of websites. This may be improved by using the phrases related to the search, which we have already implemented while searching the papers.

In any situation when there are many papers, the algorithm will assist in obtaining and extracting the most pertinent documents and rating them in order.

EXPERIMENT RESULTS:

Dataset:

The White House and a group of leading academic institutions developed the COVID-19 Open Scientific Dataset in response to the COVID-19 outbreak (CORD-19). The CORD-19 database contains more than a million research articles on COVID-19, SARS-CoV-2, and other coronaviruses, including more than 400,000 in full text.

The continuing COVID-19 pandemic is caused by the coronavirus 2 that causes severe acute respiratory syndrome (COVID 19), which is a coronavirus disease (SARS CoV 2).

By restricting the CORD-19 dataset to only include papers pertinent to COVID-19/SARS-CoV-2 that were published after December 2019, we construct a corpus of 5900 publications. Dataset

Link: https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge?resource=download

Methodology with the Dataset:

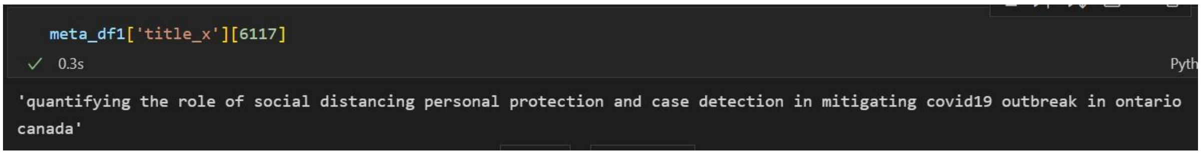
Our dataset consists mostly of a collection of about 5 lakh papers (Publishing, Research Papers, Journals, etc.). When a query is given to us, our goal is to quickly identify the appropriate document from this mass of documents.

We make an effort to search for the precise query in the dataset as a document title.

Therefore, based on our technique and approach, we ought to obtain the same document as the top papers that were implemented.

As our search term, let's use the title of a random document.

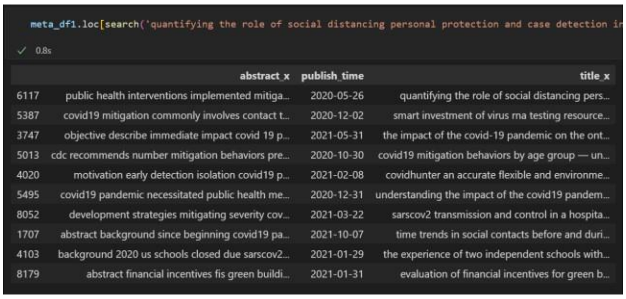
Index of the random document: 6117



```
meta_df1['title_x'][6117]
✓ 0.3s                                                                    Pyth
'quantifying the role of social distancing personal protection and case detection in mitigating covid19 outbreak in ontario
canada'
```

Search Query: 'quantifying the role of social distancing personal protection and case detection in mitigating covid19 outbreak in ontario canada'

After Implementation We get the desired Result:



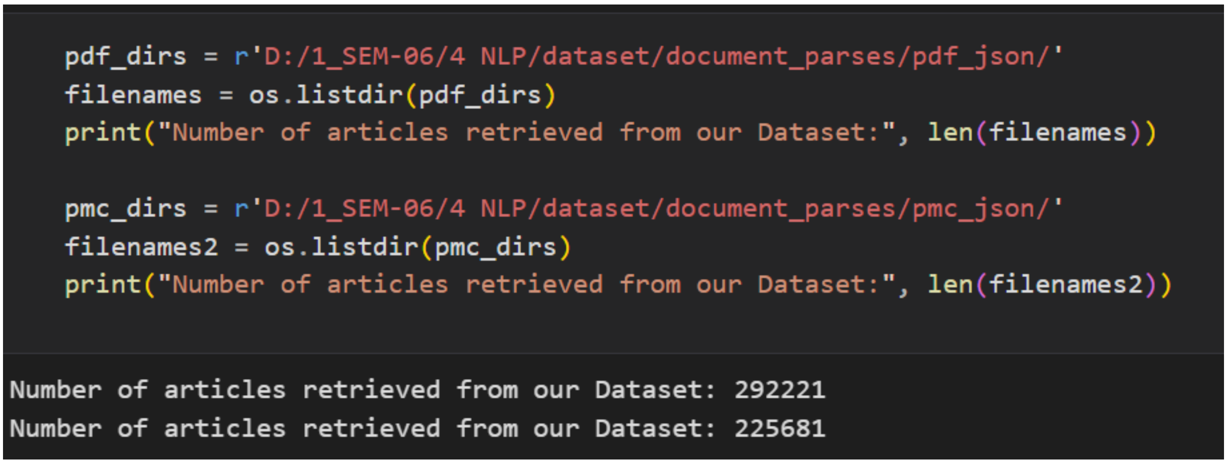
```
meta_df1.loc[search('quantifying the role of social distancing personal protection and case detection i
✓ 0.8s
```

	abstract_x	publish_time	title_x
6117	public health interventions implemented mitiga...	2020-05-26	quantifying the role of social distancing pers...
5387	covid19 mitigation commonly involves contact t...	2020-12-02	smart investment of virus rna testing resource...
3747	objective describe immediate impact covid 19 p...	2021-05-31	the impact of the covid-19 pandemic on the ont...
5013	cdc recommends number mitigation behaviors pre...	2020-10-30	covid19 mitigation behaviors by age group — un...
4020	motivation early detection isolation covid19 p...	2021-02-08	covidhunter an accurate flexible and environme...
5495	covid19 pandemic necessitated public health me...	2020-12-31	understanding the impact of the covid19 pandem...
8052	development strategies mitigating severity cov...	2021-03-22	sarscov2 transmission and control in a hospita...
1707	abstract background since beginning covid19 pa...	2021-10-07	time trends in social contacts before and duri...
4103	background 2020 us schools closed due sarscov2...	2021-01-29	the experience of two independent schools with...
8179	abstract financial incentives fis green buildi...	2021-01-31	evaluation of financial incentives for green b...

We get top index as 6117 which was the required result.

Sample Output Screen:

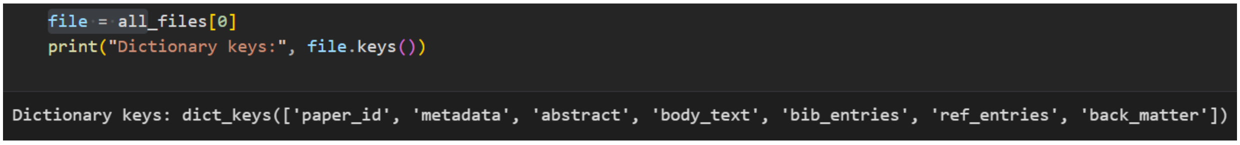
Number of articles from different sources in our Dataset:



```
pdf_dirs = r'D:/1_SEM-06/4 NLP/dataset/document_parses/pdf_json/'
filenames = os.listdir(pdf_dirs)
print("Number of articles retrieved from our Dataset:", len(filenames))

pmc_dirs = r'D:/1_SEM-06/4 NLP/dataset/document_parses/pmc_json/'
filenames2 = os.listdir(pmc_dirs)
print("Number of articles retrieved from our Dataset:", len(filenames2))
```

```
Number of articles retrieved from our Dataset: 292221
Number of articles retrieved from our Dataset: 225681
```

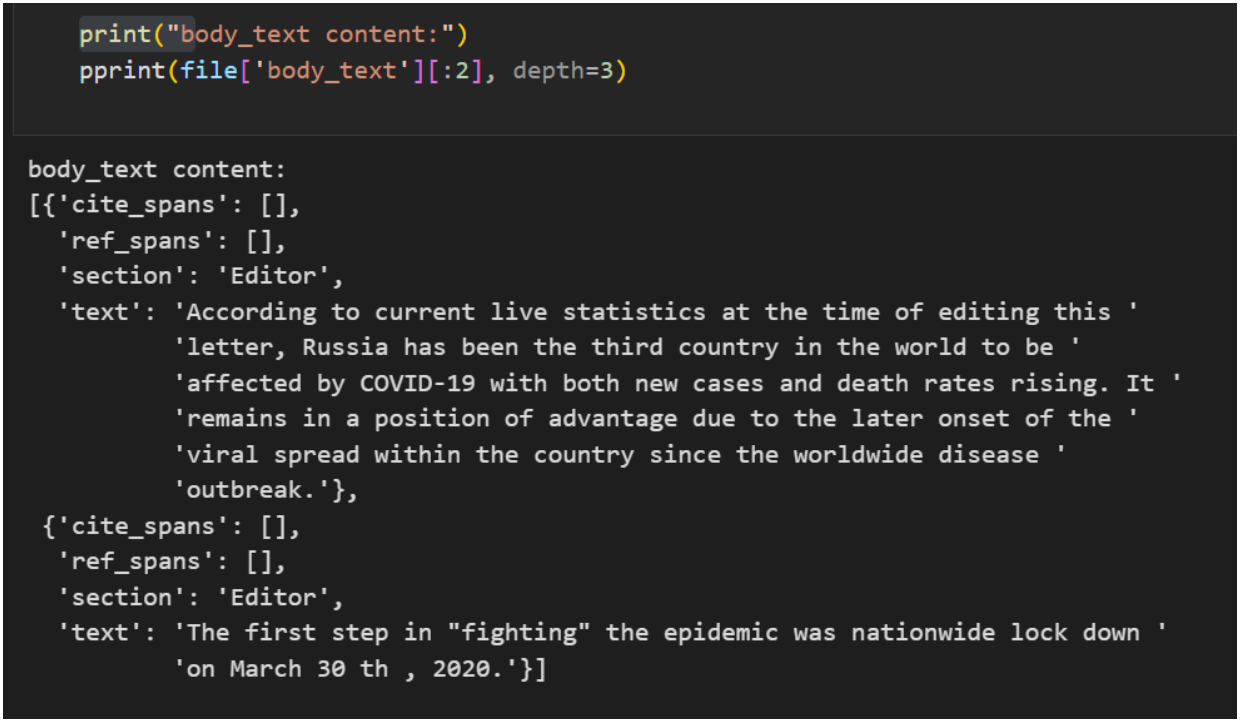
The Keys(Attributes) of a particular document:



```
file = all_files[0]
print("Dictionary keys:", file.keys())
```

```
Dictionary keys: dict_keys(['paper_id', 'metadata', 'abstract', 'body_text', 'bib_entries', 'ref_entries', 'back_matter'])
```

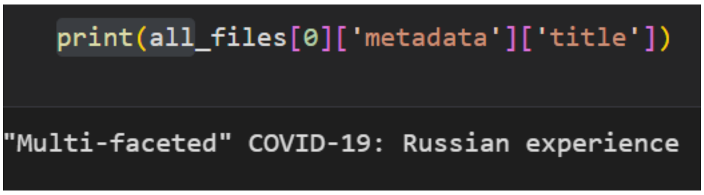
Sample Body text of a random document:



```
print("body_text content:")
pprint(file['body_text'][:2], depth=3)
```

```
body_text content:
[{'cite_spans': [],
  'ref_spans': [],
  'section': 'Editor',
  'text': 'According to current live statistics at the time of editing this '
          'letter, Russia has been the third country in the world to be '
          'affected by COVID-19 with both new cases and death rates rising. It '
          'remains in a position of advantage due to the later onset of the '
          'viral spread within the country since the worldwide disease '
          'outbreak.'},
 {'cite_spans': [],
  'ref_spans': [],
  'section': 'Editor',
  'text': 'The first step in "fighting" the epidemic was nationwide lock down '
          'on March 30 th , 2020.'}]
```

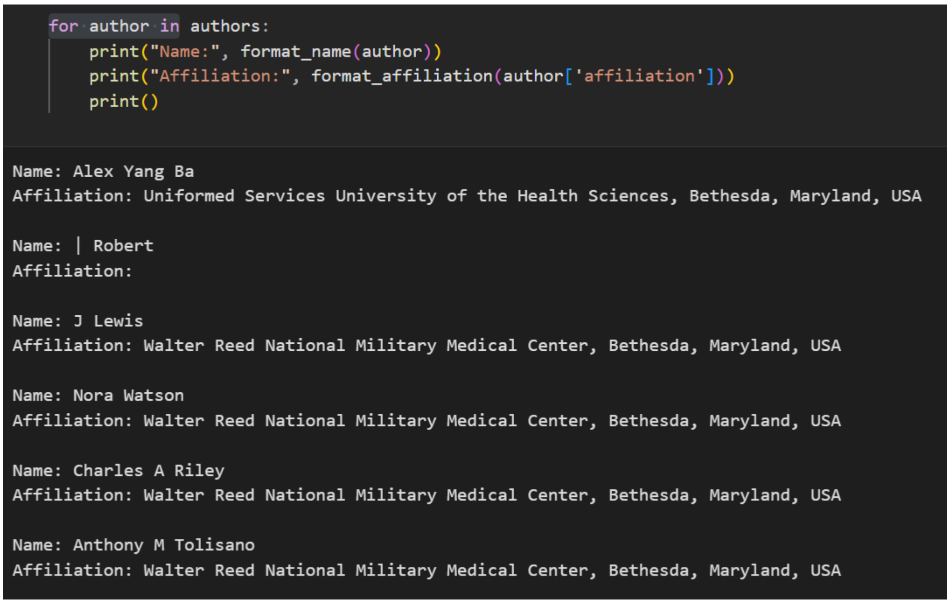
Title of a random Document:



```
print(all_files[0]['metadata']['title'])
```

```
"Multi-faceted" COVID-19: Russian experience
```

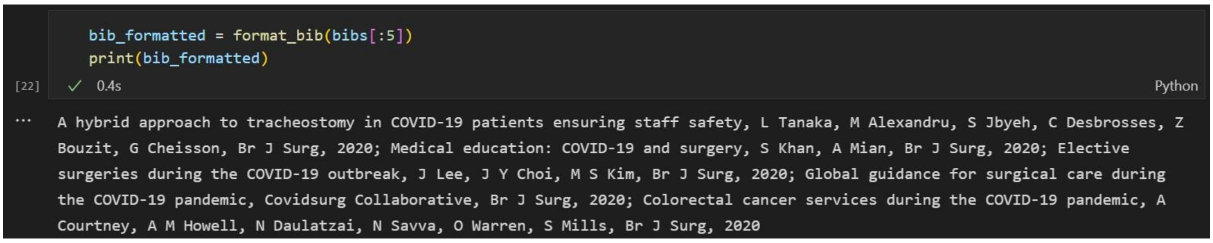
Formatted author names of a particular document:



```
for author in authors:
    print("Name:", format_name(author))
    print("Affiliation:", format_affiliation(author['affiliation']))
    print()
```

```
Name: Alex Yang Ba
Affiliation: Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

Name: | Robert
Affiliation:

Name: J Lewis
Affiliation: Walter Reed National Military Medical Center, Bethesda, Maryland, USA

Name: Nora Watson
Affiliation: Walter Reed National Military Medical Center, Bethesda, Maryland, USA

Name: Charles A Riley
Affiliation: Walter Reed National Military Medical Center, Bethesda, Maryland, USA

Name: Anthony M Tolisano
Affiliation: Walter Reed National Military Medical Center, Bethesda, Maryland, USA
```

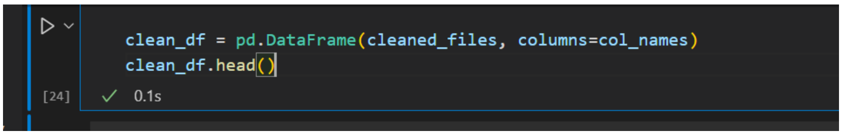
Sample Bibliography of a particular document:



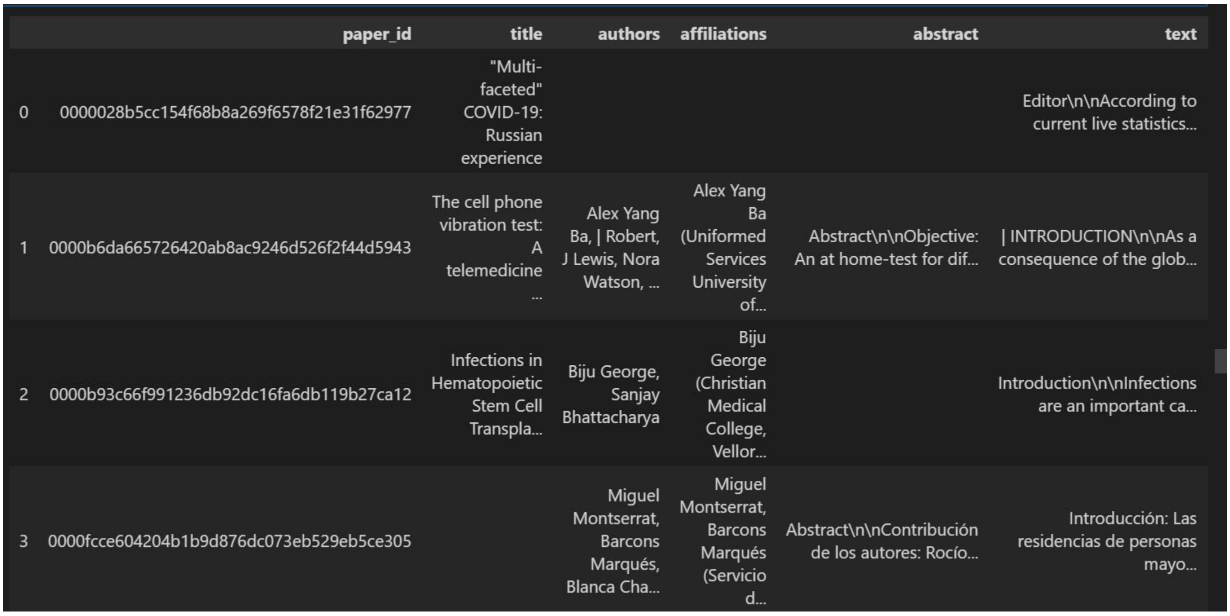
```
bib_formatted = format_bib(bibs[:5])
print(bib_formatted)
```

```
A hybrid approach to tracheostomy in COVID-19 patients ensuring staff safety, L Tanaka, M Alexandru, S Jbyeh, C Desbrosses, Z
Bouzit, G Cheisson, Br J Surg, 2020; Medical education: COVID-19 and surgery, S Khan, A Mian, Br J Surg, 2020; Elective
surgeries during the COVID-19 outbreak, J Lee, J Y Choi, M S Kim, Br J Surg, 2020; Global guidance for surgical care during
the COVID-19 pandemic, Covidsurg Collaborative, Br J Surg, 2020; Colorectal cancer services during the COVID-19 pandemic, A
Courtney, A M Howell, N Daulatzai, N Savva, O Warren, S Mills, Br J Surg, 2020
```

Converting The documents into a CSV File:

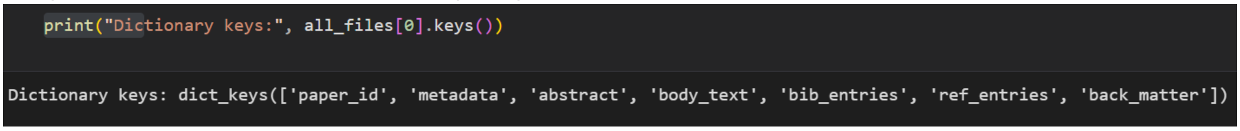


```
clean_df = pd.DataFrame(cleaned_files, columns=col_names)
clean_df.head()
```



	paper_id	title	authors	affiliations	abstract	text
0	0000028b5cc154f68b8a269f6578f21e31f62977	"Multi-faceted" COVID-19: Russian experience				Editor\n\nAccording to current live statistics...
1	0000b6da665726420ab8ac9246d526f2f44d5943	The cell phone vibration test: A telemedicine ...	Alex Yang Ba, | Robert, J Lewis, Nora Watson, ...	Alex Yang Ba (Uniformed Services University of...	Abstract\n\nObjective: An at home-test for dif...	| INTRODUCTION\n\nAs a consequence of the glob...
2	0000b93c66f991236db92dc16fa6db119b27ca12	Infections in Hematopoietic Stem Cell Transpla...	Biju George, Sanjay Bhattacharya	Biju George (Christian Medical College, Vellor...		Introduction\n\nInfections are an important ca...
3	0000fcce604204b1b9d876dc073eb529eb5ce305		Miguel Montserrat, Barcons Marqués, Blanca Cha...	Miguel Montserrat, Barcons Marqués (Servicio d...	Abstract\n\nContribución de los autores: Rocío...	Introducción: Las residencias de personas mayo...

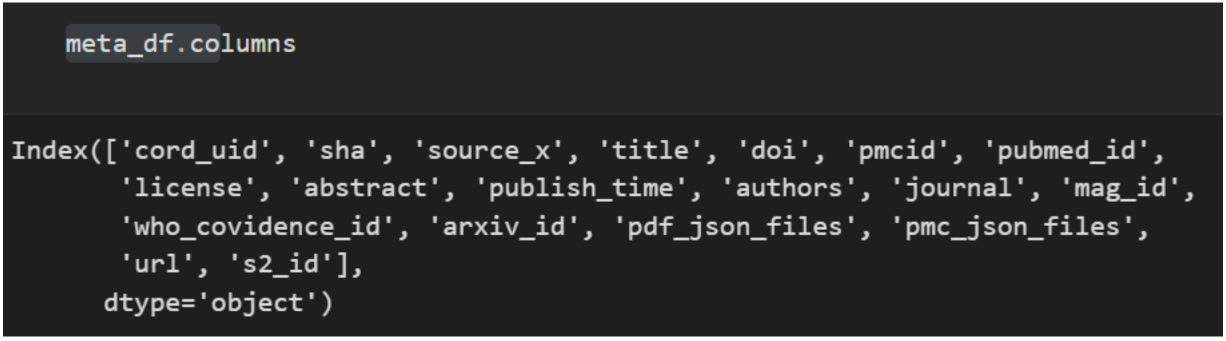
Keys of document selected randomly (Keys of the CSV file):



```
print("Dictionary keys:", all_files[0].keys())
```

```
Dictionary keys: dict_keys(['paper_id', 'metadata', 'abstract', 'body_text', 'bib_entries', 'ref_entries', 'back_matter'])
```

Keys of the metadata:



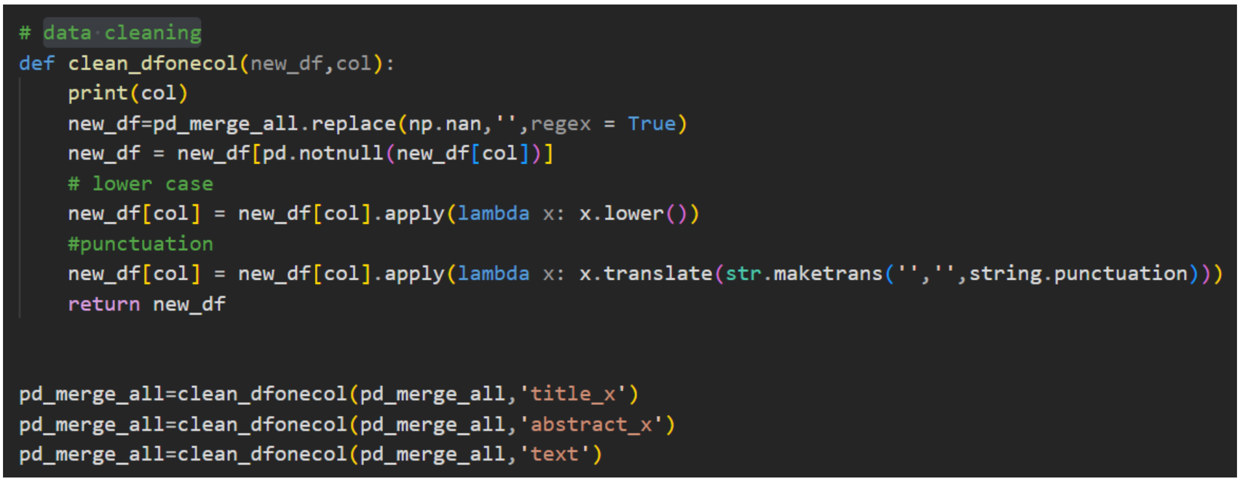
```
meta_df.columns
```

```
Index(['cord_uid', 'sha', 'source_x', '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'],
      dtype='object')
```

Converting the title, abstract and body Text of a document into lower case

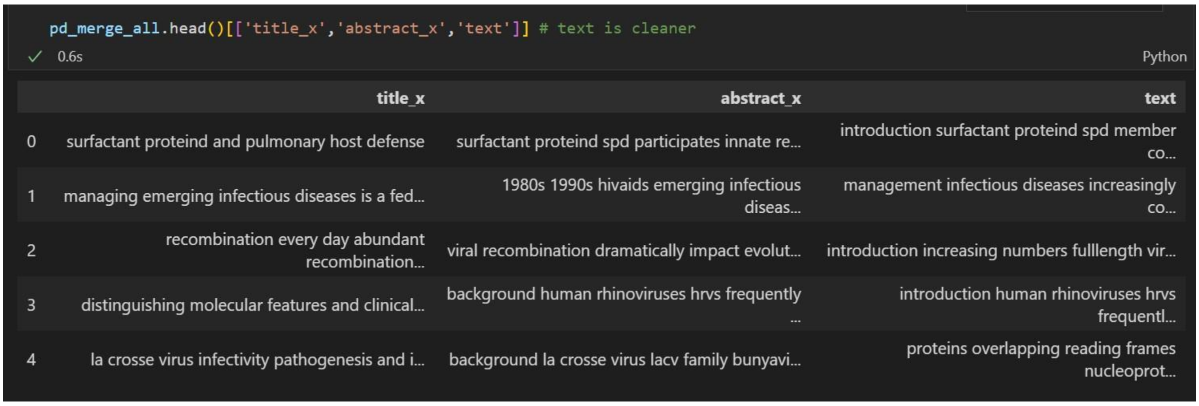
Removing the punctuation from title,abstract and bodyText of a document.

Code:



```
# data cleaning
def clean_dfonecol(new_df,col):
    print(col)
    new_df=pd_merge_all.replace(np.nan,'',regex = True)
    new_df = new_df[pd.notnull(new_df[col])]
    # lower case
    new_df[col] = new_df[col].apply(lambda x: x.lower())
    #punctuation
    new_df[col] = new_df[col].apply(lambda x: x.translate(str.maketrans('','',string.punctuation)))
    return new_df


pd_merge_all=clean_dfonecol(pd_merge_all,'title_x')
pd_merge_all=clean_dfonecol(pd_merge_all,'abstract_x')
pd_merge_all=clean_dfonecol(pd_merge_all,'text')
```

Output:



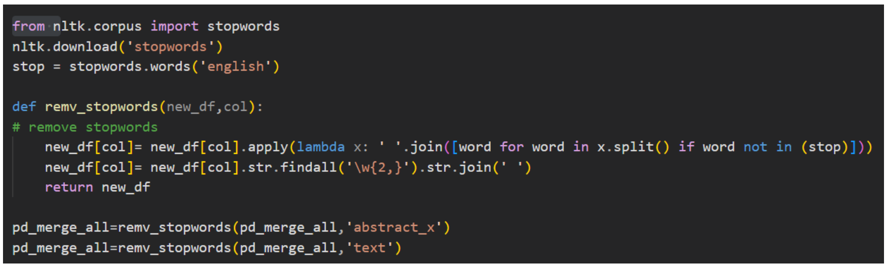
```
pd_merge_all.head()[['title_x','abstract_x','text']] # text is cleaner
```

✓ 0.6s Python

	title_x	abstract_x	text
0	surfactant proteind and pulmonary host defense	surfactant proteind spd participates innate re...	introduction surfactant proteind spd member co...
1	managing emerging infectious diseases is a fed...	1980s 1990s hivaids emerging infectious diseas...	management infectious diseases increasingly co...
2	recombination every day abundant recombination...	viral recombination dramatically impact evolut...	introduction increasing numbers fulllength vir...
3	distinguishing molecular features and clinical...	background human rhinoviruses hrvs frequently ...	introduction human rhinoviruses hrvs frequentl...
4	la crosse virus infectivity pathogenesis and i...	background la crosse virus lacv family bunyavi...	proteins overlapping reading frames nucleoprot...

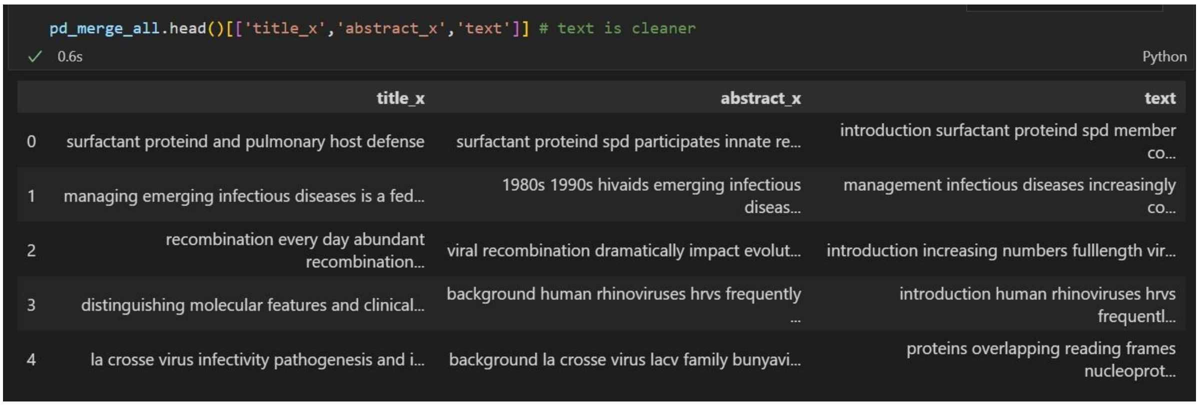
Removing Stop Words:

Code:



```python
from nltk.corpus import stopwords
nltk.download('stopwords')
stop = stopwords.words('english')

def remv_stopwords(new_df,col):
# remove stopwords
    new_df[col]= new_df[col].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))
    new_df[col]= new_df[col].str.findall('\w{2,}').str.join(' ')
    return new_df

pd_merge_all=remv_stopwords(pd_merge_all,'abstract_x')
pd_merge_all=remv_stopwords(pd_merge_all,'text')
```

Output:

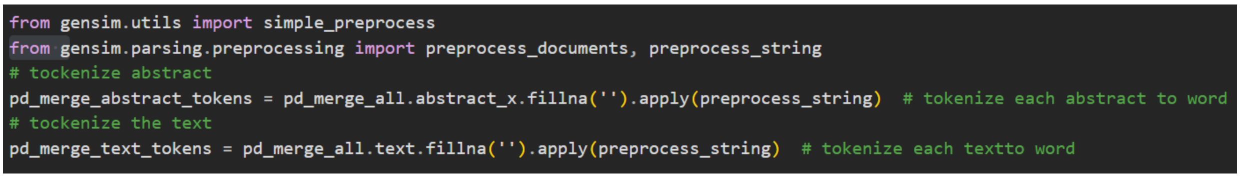


```
pd_merge_all.head()[['title_x','abstract_x','text']] # text is cleaner
✓ 0.6s                                                                Python
```

	title_x	abstract_x	text
0	surfactant proteind and pulmonary host defense	surfactant proteind spd participates innate re...	introduction surfactant proteind spd member co...
1	managing emerging infectious diseases is a fed...	1980s 1990s hivaids emerging infectious diseas...	management infectious diseases increasingly co...
2	recombination every day abundant recombination...	viral recombination dramatically impact evolut...	introduction increasing numbers fulllength vir...
3	distinguishing molecular features and clinical...	background human rhinoviruses hrvs frequently ...	introduction human rhinoviruses hrvs frequentl...
4	la crosse virus infectivity pathogenesis and i...	background la crosse virus lacv family bunyavi...	proteins overlapping reading frames nucleoprot...

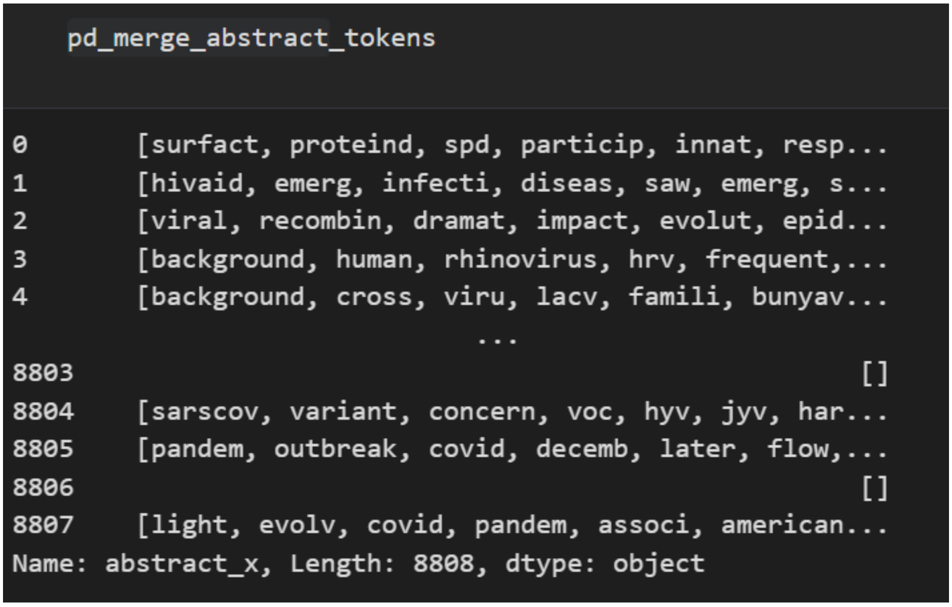
Tokenizing the abstract and text of the Document:

Code:



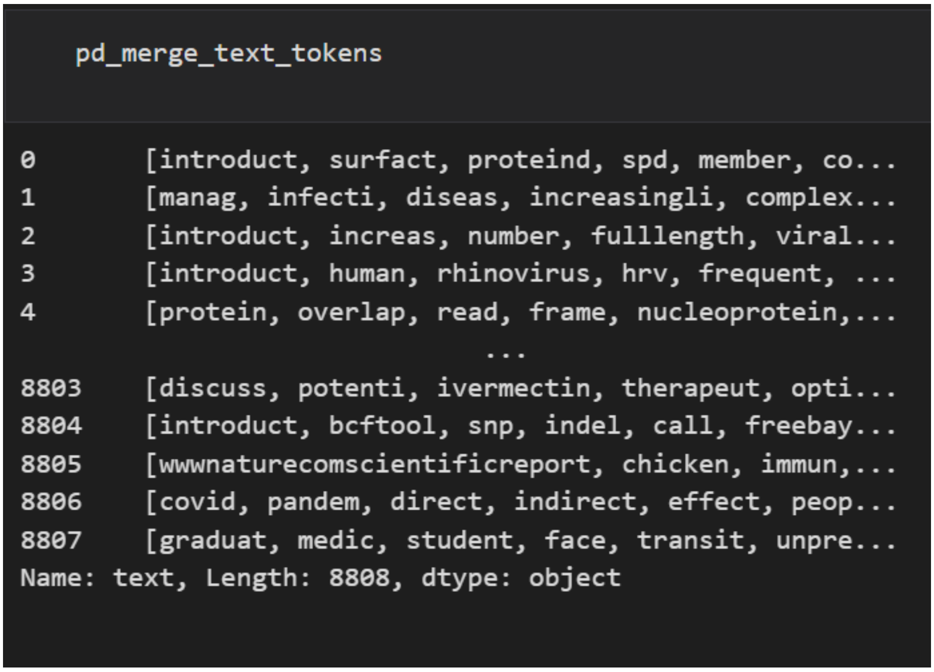
```python
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import preprocess_documents, preprocess_string
# tockenize abstract
pd_merge_abstract_tokens = pd_merge_all.abstract_x.fillna('').apply(preprocess_string)  # tokenize each abstract to word
# tockenize the text
pd_merge_text_tokens = pd_merge_all.text.fillna('').apply(preprocess_string)  # tokenize each textto word
```

Output (Abstract)



```
pd_merge_abstract_tokens

0       [surfact, proteind, spd, particip, innat, resp...
1       [hivaid, emerg, infecti, diseas, saw, emerg, s...
2       [viral, recombin, dramat, impact, evolut, epid...
3       [background, human, rhinovirus, hrv, frequent,...
4       [background, cross, viru, lacv, famili, bunyav...
                              ...
8803                                                   []
8804    [sarscov, variant, concern, voc, hyv, jyv, har...
8805    [pandem, outbreak, covid, decemb, later, flow,...
8806                                                   []
8807    [light, evolv, covid, pandem, associ, american...
Name: abstract_x, Length: 8808, dtype: object
```

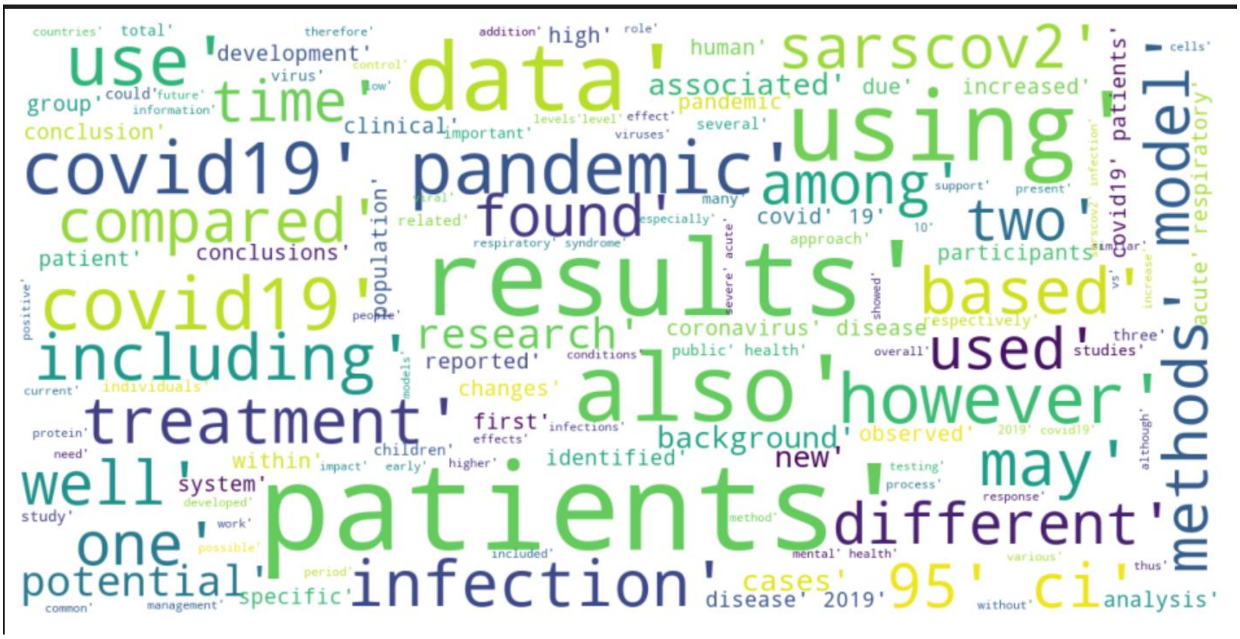
Output(Text)



```
pd_merge_text_tokens
```

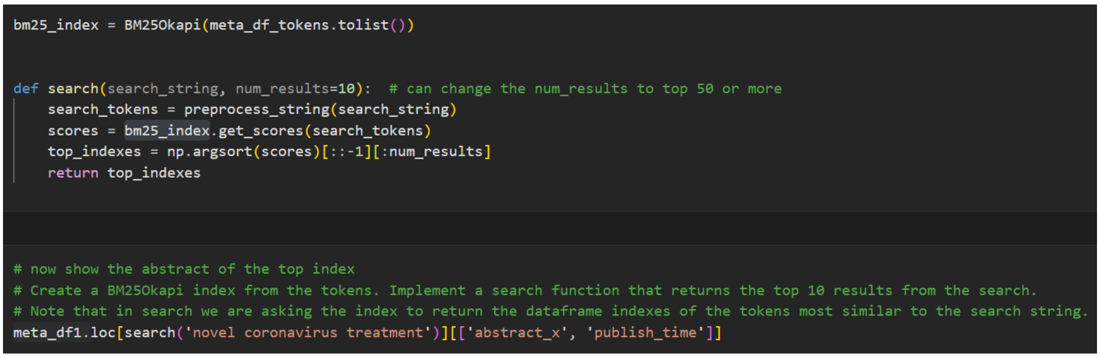
```
0       [introduct, surfact, proteind, spd, member, co...
1       [manag, infecti, diseas, increasingli, complex...
2       [introduct, increas, number, fulllength, viral...
3       [introduct, human, rhinovirus, hrv, frequent, ...
4       [protein, overlap, read, frame, nucleoprotein,...
                                ...
8803    [discuss, potenti, ivermectin, therapeut, opti...
8804    [introduct, bcftool, snp, indel, call, freebay...
8805    [wwwnaturecomscientificreport, chicken, immun,...
8806    [covid, pandem, direct, indirect, effect, peop...
8807    [graduat, medic, student, face, transit, unpre...
Name: text, Length: 8808, dtype: object
```

A WordCloud of All the Tokens:



Using Okapi BM-25 for Search Query Optimization:

Code:

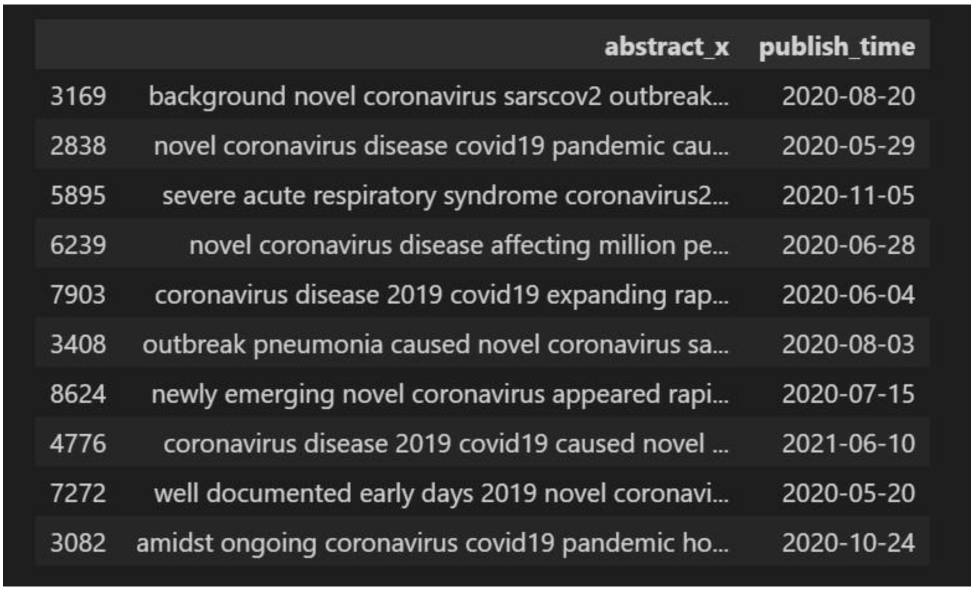


```
bm25_index = BM25Okapi(meta_df_tokens.tolist())


def search(search_string, num_results=10):  # can change the num_results to top 50 or more
    search_tokens = preprocess_string(search_string)
    scores = bm25_index.get_scores(search_tokens)
    top_indexes = np.argsort(scores)[::-1][:num_results]
    return top_indexes
```

```
# now show the abstract of the top index
# Create a BM25Okapi index from the tokens. Implement a search function that returns the top 10 results from the search.
# Note that in search we are asking the index to return the dataframe indexes of the tokens most similar to the search string.
meta_df1.loc[search('novel coronavirus treatment')][['abstract_x', 'publish_time']]
```

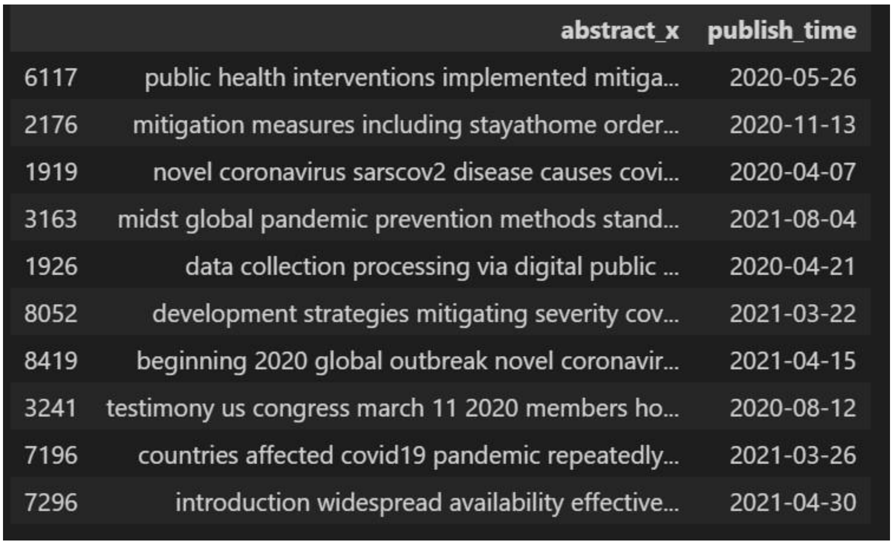
Search Query : novel coronavirus treatment

Result Documents:



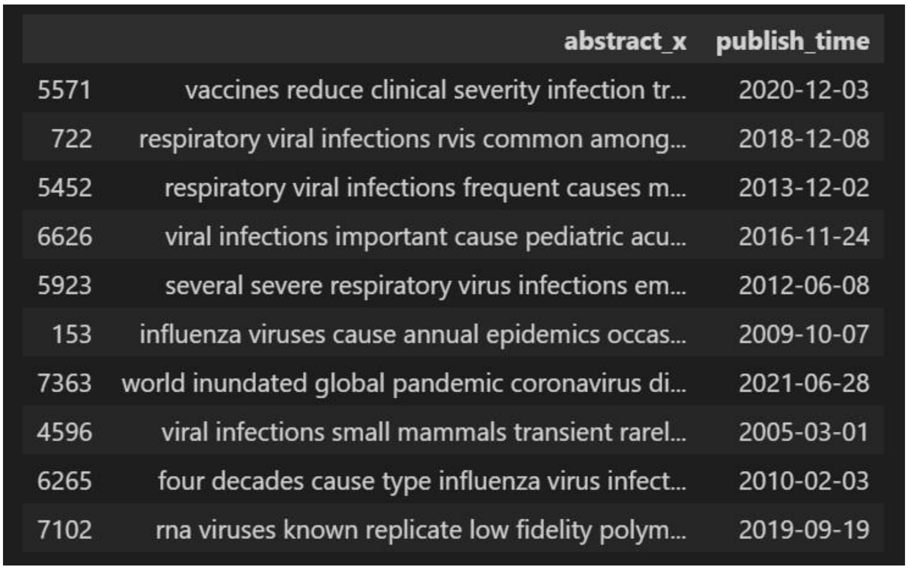
	abstract_x	publish_time
3169	background novel coronavirus sarscov2 outbreak...	2020-08-20
2838	novel coronavirus disease covid19 pandemic cau...	2020-05-29
5895	severe acute respiratory syndrome coronavirus2...	2020-11-05
6239	novel coronavirus disease affecting million pe...	2020-06-28
7903	coronavirus disease 2019 covid19 expanding rap...	2020-06-04
3408	outbreak pneumonia caused novel coronavirus sa...	2020-08-03
8624	newly emerging novel coronavirus appeared rapi...	2020-07-15
4776	coronavirus disease 2019 covid19 caused novel ...	2021-06-10
7272	well documented early days 2019 novel coronavi...	2020-05-20
3082	amidst ongoing coronavirus covid19 pandemic ho...	2020-10-24

Search Query: Public health mitigation measures that could be effective for control



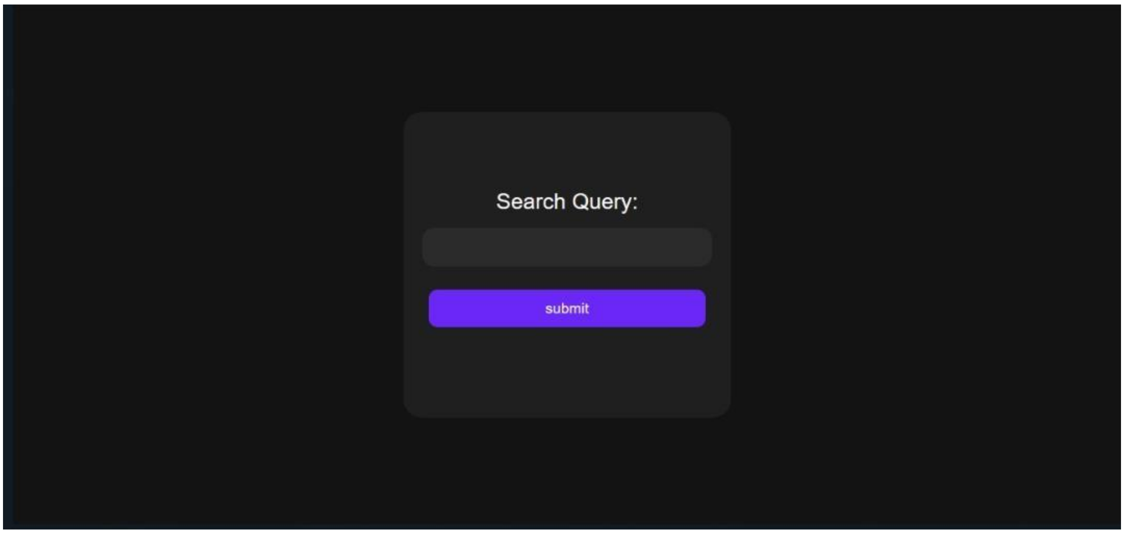
	abstract_x	publish_time
6117	public health interventions implemented mitiga...	2020-05-26
2176	mitigation measures including stayathome order...	2020-11-13
1919	novel coronavirus sarscov2 disease causes covi...	2020-04-07
3163	midst global pandemic prevention methods stand...	2021-08-04
1926	data collection processing via digital public ...	2020-04-21
8052	development strategies mitigating severity cov...	2021-03-22
8419	beginning 2020 global outbreak novel coronavir...	2021-04-15
3241	testimony us congress march 11 2020 members ho...	2020-08-12
7196	countries affected covid19 pandemic repeatedly...	2021-03-26
7296	introduction widespread availability effective...	2021-04-30

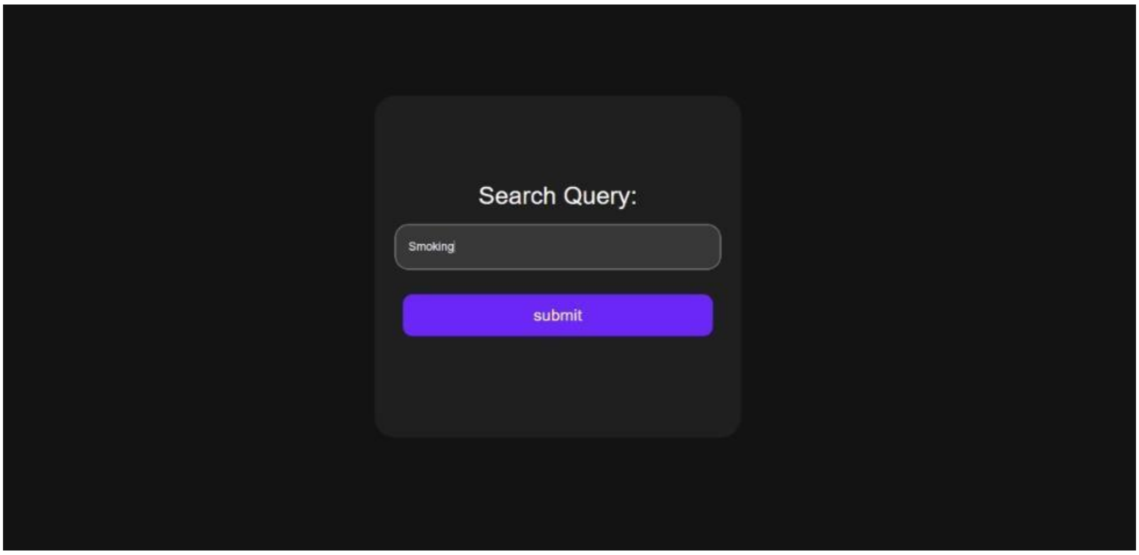
Search Query: Co-infections (determine whether co-existing respiratory/viral infections make the virus more transmissible or virulent) and other co-morbidities



	abstract_x	publish_time
5571	vaccines reduce clinical severity infection tr...	2020-12-03
722	respiratory viral infections rvis common among...	2018-12-08
5452	respiratory viral infections frequent causes m...	2013-12-02
6626	viral infections important cause pediatric acu...	2016-11-24
5923	several severe respiratory virus infections em...	2012-06-08
153	influenza viruses cause annual epidemics occas...	2009-10-07
7363	world inundated global pandemic coronavirus di...	2021-06-28
4596	viral infections small mammals transient rarel...	2005-03-01
6265	four decades cause type influenza virus infect...	2010-02-03
7102	rna viruses known replicate low fidelity polym...	2019-09-19

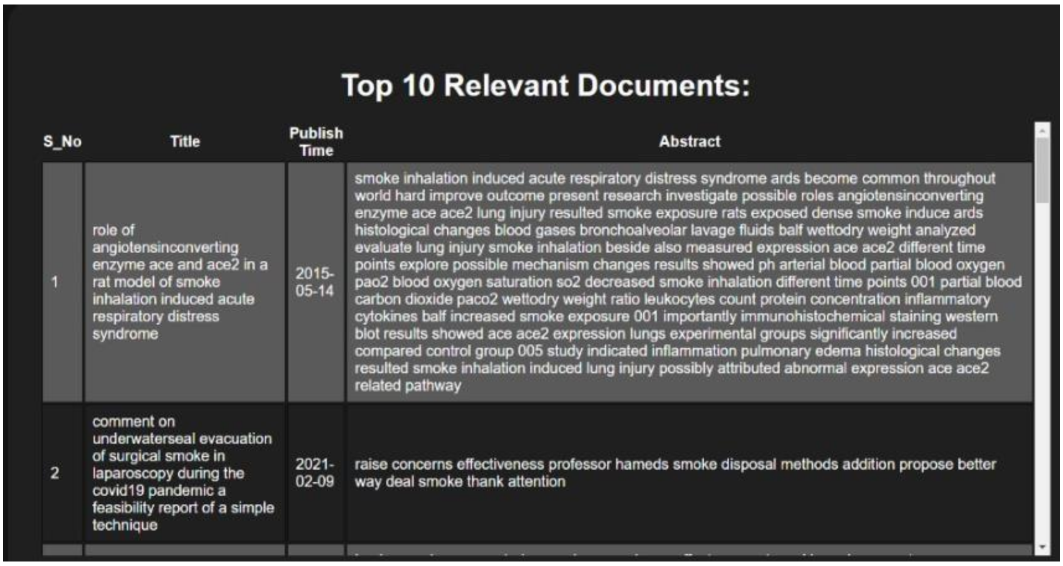
After implementing the site:





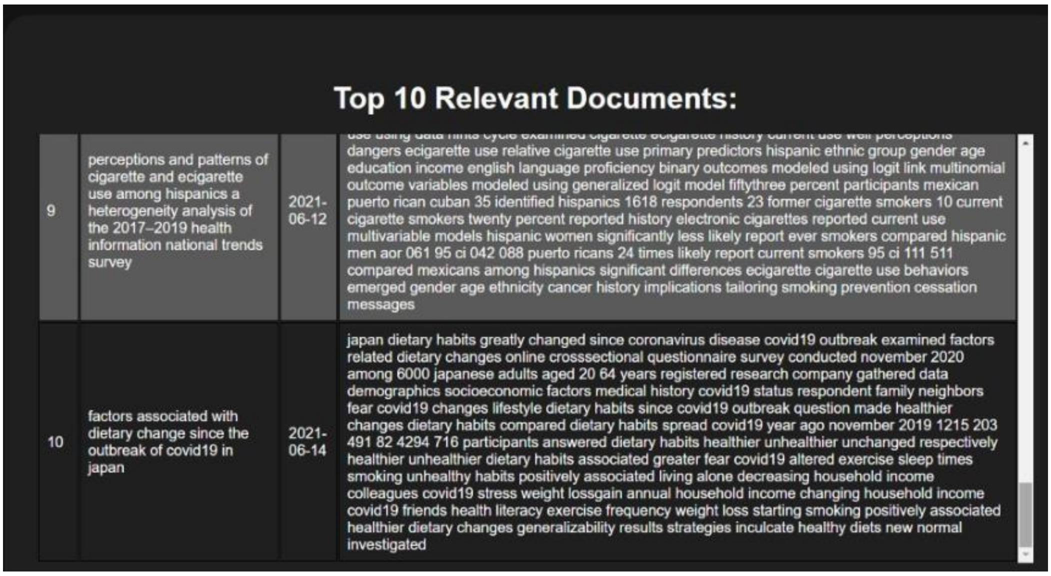
Smoking

Results given by the website after implementing the python code which is responsible for the NLP part:



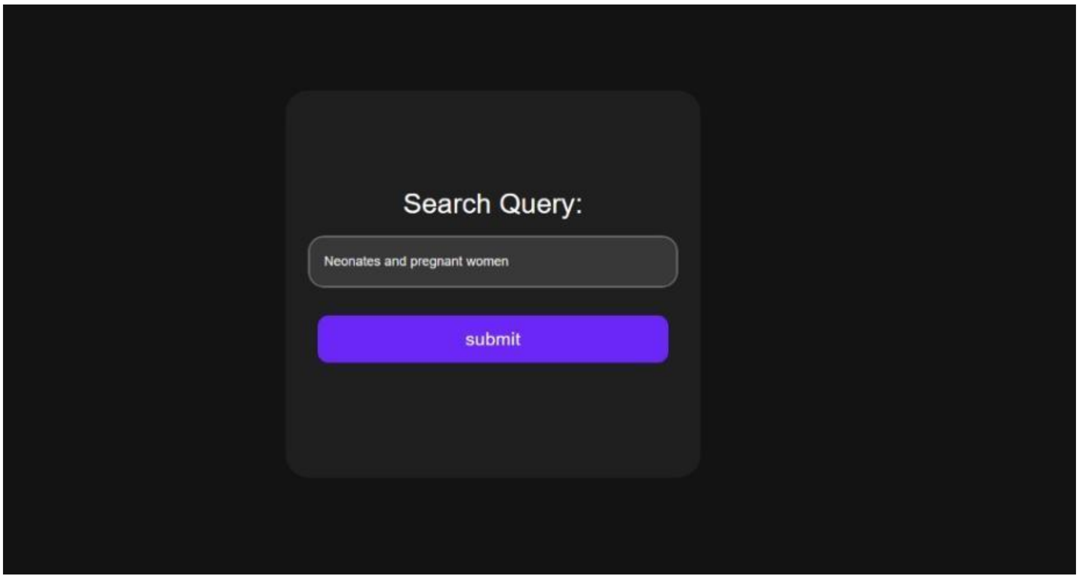
Top 10 Relevant Documents:

S_No	Title	Publish Time	Abstract
1	role of angiotensinconverting enzyme ace and ace2 in a rat model of smoke inhalation induced acute respiratory distress syndrome	2015-05-14	smoke inhalation induced acute respiratory distress syndrome ards become common throughout world hard improve outcome present research investigate possible roles angiotensinconverting enzyme ace ace2 lung injury resulted smoke exposure rats exposed dense smoke induce ards histological changes blood gases bronchoalveolar lavage fluids balf wettodry weight analyzed evaluate lung injury smoke inhalation beside also measured expression ace ace2 different time points explore possible mechanism changes results showed ph arterial blood partial blood oxygen pao2 blood oxygen saturation so2 decreased smoke inhalation different time points 001 partial blood carbon dioxide paco2 wettodry weight ratio leukocytes count protein concentration inflammatory cytokines balf increased smoke exposure 001 importantly immunohistochemical staining western blot results showed ace ace2 expression lungs experimental groups significantly increased compared control group 005 study indicated inflammation pulmonary edema histological changes resulted smoke inhalation induced lung injury possibly attributed abnormal expression ace ace2 related pathway
2	comment on underwaterseal evacuation of surgical smoke in laparoscopy during the covid19 pandemic a feasibility report of a simple technique	2021-02-09	raise concerns effectiveness professor hameds smoke disposal methods addition propose better way deal smoke thank attention



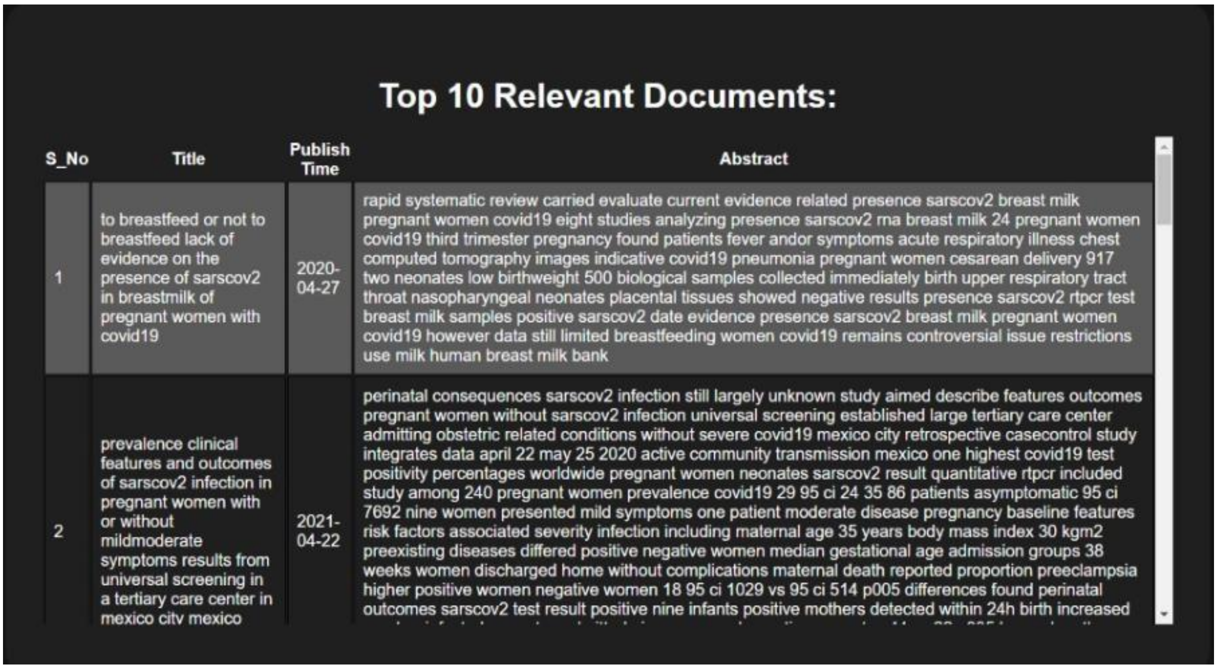
Top 10 Relevant Documents:

9	perceptions and patterns of cigarette and ecigarette use among hispanics a heterogeneity analysis of the 2017–2019 health information national trends survey	2021-06-12	use using data hints cycle examined cigarette ecigarette history current use well perceptions dangers ecigarette use relative cigarette use primary predictors hispanic ethnic group gender age education income english language proficiency binary outcomes modeled using logit link multinomial outcome variables modeled using generalized logit model fiftythree percent participants mexican puerto rican cuban 35 identified hispanics 1618 respondents 23 former cigarette smokers 10 current cigarette smokers twenty percent reported history electronic cigarettes reported current use multivariable models hispanic women significantly less likely report ever smokers compared hispanic men aor 061 95 ci 042 088 puerto ricans 24 times likely report current smokers 95 ci 111 511 compared mexicans among hispanics significant differences ecigarette cigarette use behaviors emerged gender age ethnicity cancer history implications tailoring smoking prevention cessation messages
10	factors associated with dietary change since the outbreak of covid19 in japan	2021-06-14	japan dietary habits greatly changed since coronavirus disease covid19 outbreak examined factors related dietary changes online crosssectional questionnaire survey conducted november 2020 among 6000 japanese adults aged 20 64 years registered research company gathered data demographics socioeconomic factors medical history covid19 status respondent family neighbors fear covid19 changes lifestyle dietary habits since covid19 outbreak question made healthier changes dietary habits compared dietary habits spread covid19 year ago november 2019 1215 203 491 82 4294 716 participants answered dietary habits healthier unhealthier unchanged respectively healthier unhealthier dietary habits associated greater fear covid19 altered exercise sleep times smoking unhealthy habits positively associated living alone decreasing household income colleagues covid19 stress weight lossgain annual household income changing household income covid19 friends health literacy exercise frequency weight loss starting smoking positively associated healthier dietary changes generalizability results strategies inculcate healthy diets new normal investigated



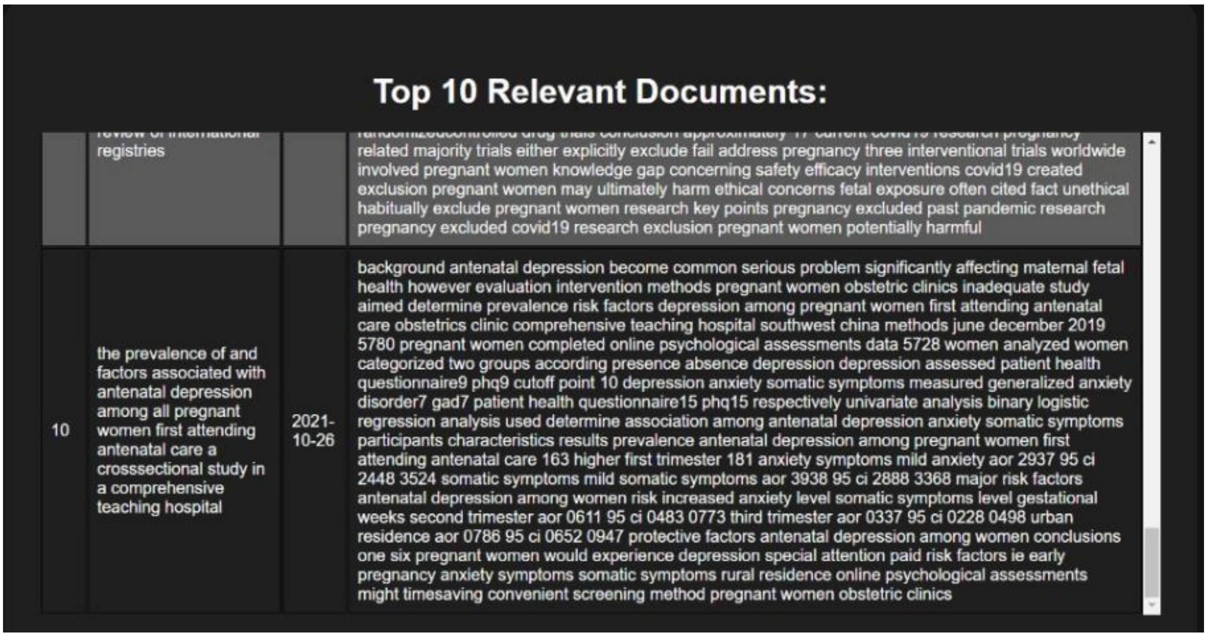
Search Query:

Neonates and pregnant women



Top 10 Relevant Documents:

S_No	Title	Publish Time	Abstract
1	to breastfeed or not to breastfeed lack of evidence on the presence of sarscov2 in breastmilk of pregnant women with covid19	2020-04-27	rapid systematic review carried evaluate current evidence related presence sarscov2 breast milk pregnant women covid19 eight studies analyzing presence sarscov2 rna breast milk 24 pregnant women covid19 third trimester pregnancy found patients fever andor symptoms acute respiratory illness chest computed tomography images indicative covid19 pneumonia pregnant women cesarean delivery 917 two neonates low birthweight 500 biological samples collected immediately birth upper respiratory tract throat nasopharyngeal neonates placental tissues showed negative results presence sarscov2 rtpcr test breast milk samples positive sarscov2 date evidence presence sarscov2 breast milk pregnant women covid19 however data still limited breastfeeding women covid19 remains controversial issue restrictions use milk human breast milk bank
2	prevalence clinical features and outcomes of sarscov2 infection in pregnant women with or without mildmoderate symptoms results from universal screening in a tertiary care center in mexico city mexico	2021-04-22	perinatal consequences sarscov2 infection still largely unknown study aimed describe features outcomes pregnant women without sarscov2 infection universal screening established large tertiary care center admitting obstetric related conditions without severe covid19 mexico city retrospective casecontrol study integrates data april 22 may 25 2020 active community transmission mexico one highest covid19 test positivity percentages worldwide pregnant women neonates sarscov2 result quantitative rtpcr included study among 240 pregnant women prevalence covid19 29 95 ci 24 35 86 patients asymptomatic 95 ci 7692 nine women presented mild symptoms one patient moderate disease pregnancy baseline features risk factors associated severity infection including maternal age 35 years body mass index 30 kgm2 preexisting diseases differed positive negative women median gestational age admission groups 38 weeks women discharged home without complications maternal death reported proportion preeclampsia higher positive women negative women 18 95 ci 1029 vs 95 ci 514 p005 differences found perinatal outcomes sarscov2 test result positive nine infants positive mothers detected within 24h birth increased



Top 10 Relevant Documents:

	review of international registries		randomizedcontrolled drug trials conclusion approximately 17 current covid19 research pregnancy related majority trials either explicitly exclude fail address pregnancy three interventional trials worldwide involved pregnant women knowledge gap concerning safety efficacy interventions covid19 created exclusion pregnant women may ultimately harm ethical concerns fetal exposure often cited fact unethical habitually exclude pregnant women research key points pregnancy excluded past pandemic research pregnancy excluded covid19 research exclusion pregnant women potentially harmful
10	the prevalence of and factors associated with antenatal depression among all pregnant women first attending antenatal care a crosssectional study in a comprehensive teaching hospital	2021-10-26	background antenatal depression become common serious problem significantly affecting maternal fetal health however evaluation intervention methods pregnant women obstetric clinics inadequate study aimed determine prevalence risk factors depression among pregnant women first attending antenatal care obstetrics clinic comprehensive teaching hospital southwest china methods june december 2019 5780 pregnant women completed online psychological assessments data 5728 women analyzed women categorized two groups according presence absence depression depression assessed patient health questionnaire9 phq9 cutoff point 10 depression anxiety somatic symptoms measured generalized anxiety disorder7 gad7 patient health questionnaire15 phq15 respectively univariate analysis binary logistic regression analysis used determine association among antenatal depression anxiety somatic symptoms participants characteristics results prevalence antenatal depression among pregnant women first attending antenatal care 163 higher first trimester 181 anxiety symptoms mild anxiety aor 2937 95 ci 2448 3524 somatic symptoms mild somatic symptoms aor 3938 95 ci 2888 3368 major risk factors antenatal depression among women risk increased anxiety level somatic symptoms level gestational weeks second trimester aor 0611 95 ci 0483 0773 third trimester aor 0337 95 ci 0228 0498 urban residence aor 0786 95 ci 0652 0947 protective factors antenatal depression among women conclusions one six pregnant women would experience depression special attention paid risk factors ie early pregnancy anxiety symptoms somatic symptoms rural residence online psychological assessments might timesaving convenient screening method pregnant women obstetric clinics

CONCLUSION:

Our project aims to improve information retrieval and prioritize papers based on their significance, which is crucial given the overwhelming amount of available information that often impedes knowledge acquisition. Our primary objective is to enable researchers, academics, and scientists to effectively utilize vast information resources and transform them into meaningful insights. This initiative was inspired by the extensive research conducted during the COVID-19 pandemic, where physicians and scientists collaborated to develop a vaccine for the illness, generating an extraordinary amount of data. Our project not only benefits ongoing studies but will also prove useful in future crises similar to the COVID-19 pandemic, serving both experts and students alike.

References:

[1] Nogueira, Rodrigo, Zhiying Jiang, and Jimmy Lin. "Document ranking with a pretrained sequence-to-sequence model." *ArXiv preprint arXiv:2003.06713* (2020).

[2] Rana, Ashish, Pujit Golchha, Roni Juntunen, Andreea Coajă, Ahmed Elzamarany, Chia-Chien Hung, and Simone Paolo Ponzetto. "LeviRANK: Limited Query Expansion with Voting Integration for Document Retrieval and Ranking." *Working Notes Papers of the CLEF* (2022).

[3] da Silva, Fabiano Tavares, and José Everardo Bessa Maia. "Query Expansion in Text Information Retrieval with Local Context and Distributional Model." *J. Digit. Inf. Manag.* 17, no. 6 (2019): 313.

[4] Zeng, Qing T., Doug Redd, Thomas Rindflesch, and Jonathan Nebeker. "Synonym, topic model and predicate-based query expansion for retrieving clinical documents." In AMIA Annual Symposium Proceedings, vol. 2012, p. 1050. American Medical Informatics Association, 2012.

[5] Rivas, Andreia Rodrıguez, Eva Lorenzo Iglesias, and L. Borrajo. "Study of query expansion techniques and their application in the biomedical information retrieval." The Scientific World Journal 2014 (2014).

[6] Jang, Youngjin, and Harksoo Kim. "Document re-ranking model for machine-reading comprehension." Applied Sciences 10, no. 21 (2020): 7547.

[7] He, Ben, and Iadh Ounis. "Studying query expansion effectiveness." In European Conference on Information Retrieval, pp. 611-619. Springer, Berlin, Heidelberg, 2009.

[8] Rivas, Andreia Rodrıguez, Eva Lorenzo Iglesias, and L. Borrajo. "Study of query expansion techniques and their application in the biomedical information retrieval." The Scientific World Journal 2014 (2014).

[9] Chandra, Ganesh, and Sanjay K. Dwivedi. "Query expansion for effective retrieval results of Hindi–English cross-lingual IR." Applied Artificial Intelligence 33, no. 7 (2019): 567-593.

[10] Sharma, Dilip Kumar, Rajendra Pamula, and D. S. Chauhan. "A contemporary combined approach for query expansion." Multimedia Tools and Applications (2020): 1-27.

[11] A. K. Shukla and S. Das, "A Hybrid Model of Query Expansion using Word2Vec," 2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES), 2021, pp. 1-6, doi: 10.1109/TRIBES52498.2021.9751673.

[12] Zhang, Zicheng. "An Improved BM25 for Clinical Decision Support in Precision Medicine Based on Co-word Analysis and Cuckoo Search." (2020).

[13] Zhang, Zicheng. "An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search." BMC Medical Informatics and Decision Making 21, no. 1 (2021): 1-15.

[14] Aklouche, Billel, Ibrahim Bounhas, and Yahya Slimani. "Query Expansion Based on NLP and Word Embeddings." In TREC. 2018.

[15] Balaneshin-kordan, Saeid, and Alexander Kotov. "Embedding-based query expansion for weighted sequential dependence retrieval model." In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1213-1216. 2017.