Search and retrieval service for scientific articles on COVID-19

Cristian E. Munoz Villalobos
Department of Electrical Engineering
PUC-Rio
Rio de Janeiro, Brazil
crisstrink@gmail.com

Leonardo A. Forero Mendoza
Department of Electrical Engineering
Rio de Janeiro State University
Rio de Janeiro, Brazil
leofome@eng.uerj.br

Renato Sayão da Rocha
Department of Electrical Engineering
PUC-Rio
Rio de Janeiro, Brazil
rrsayao@puc-rio.br

Jose Eduardo Ruiz
Department of Electrical Engineering
PUC-Rio
Rio de Janeiro, Brazil
joseruiz1989@hotmail.com

Harold D. de Mello Junior
Department of Electrical Engineering
Rio de Janeiro State University
Rio de Janeiro, Brazil
harold@eng.uerj.br

Marco Aurélio C. Pacheco
Department of Electrical Engineering
PUC-Rio
Rio de Janeiro, Brazil
marco@ele.puc-rio.br

***Abstract*—The COVID-19 pandemic was a global health crisis that lasted until May 4, 2023, affecting millions of people and raising many questions about transmission, diagnosis, treatment, vaccine development, and viral pathogens. Unfortunately, misinformation created more socioeconomic damage than the disease itself. To address this problem, we have developed Cognitive Search, a user-friendly application service that uses the latest advances in Natural Language Processing (NLP) to retrieve information from CORD-19, a resource for scholarly articles on COVID-19 and related pathogens. This system uses a combination of Term-Frequency, Semantic Neural Research, and Hybrid Term-Neural algorithms to improve document retrieval performance. The Hybrid Term-Neural approach also considers temporal information in documents to provide more accurate search results. With an intuitive interface, this application can generate valuable insights to help combat outbreaks.**

***Index Terms*—BERT, BM25, coronavirus, search engine, cosine similarity**

I. Introduction

COVID-19, which mainly attacks the respiratory system, started in the Wuhan region of China and was classified by the World Health Organization (WHO) as a global emergency. The outbreak caused not only a public health crisis but also an information crisis. The unprecedented number of new publications in the field of coronavirus had the potential to boost scientific progress, facilitating instant and direct access to the latest scientific findings. However, simultaneously, it left scientists with the challenge of dealing with a large amount of data to keep up.

Due to the significance and immediacy of this task, the NLP research community has shown a growing interest in biomedical text mining, information retrieval (IR), and biomedical question answering (biomedical QA). As a result, various biomedical IR and QA datasets, challenges, and competition have emerged alongside many systems, models, architectures, and techniques tailored specifically to biomedical IR and QA [1].

This paper presents the Cognitive Search[1] - a service that retrieves information from the COVID-19 Open Research Dataset (CORD-19). This database includes articles about the coronavirus SARS-CoV-2, more precisely, on control measures, vaccine development, mitigation impacts, genetic analysis, economic impact, etc. With so much scientific literature available, it is impractical - if possible - to analyze all of this manually. Thus, the system was built to satisfy information needs during the pandemic. With dedicated CPUs and GPUs to make queries in real-time, the system allows:

- query: a short keyword or a sentence query;
- question: a more precise natural language question;
- narrative: a more extended description that further elaborates on the question, often providing specific types of information intent

The significant contribution of this work is to provide a search engine that concatenates different information retrieval approaches. The system enables rending documents retrieval by Term-Frequency and Neural-Similarity and the Hybrid Term-Neural. The retrieval performance can often be improved significantly by combining several different retrieval algorithms and their results in contrast to just one. Additionally, the Hybrid Term-Neural approach supports the exploitation of temporal information in documents and the usage of such information to anchor search results along a well-defined timeline. Then it can generate insights through an intuitive and easy-to-use interface.

The remainder of the paper divides into seven sections. Section II looks into the fundamentals of IR. Section III describes the whole system. Section IV presents the methodology adopted for IR. The CORD-19 database and the web interface are detailed in Sections V e VI, respectively. Section VII shows

[1] http://www.iacontracovid.com.br/buscador-cognitivo/
This website is unavailable as its computational resources are allocated to other projects.

the results obtained in our experiments for each retrieval core. Finally, Section VIII summarizes the conclusion.

II. Fundamentals of Information Retrieval

Information Retrieval (IR), or more precisely, Text Information Retrieval, is a branch of computer science that deals with processing collections of documents, such as scientific papers. Information Retrieval starts when a user creates any query into the system through some graphical interface. In IR, a query matches with several collections of data objects from which the most relevant document is considered for further evaluation. A ranking is done to find the most related document to the given query. Once the core system generates results retrieval, some graphical user interface returns it to the user. Therefore, the overall IR system comprises the indexing system and the [2] consultation system. The first (blue) creates indexes, while the second (green) represents the ranking process, as can see in the scheme of Figure 1.

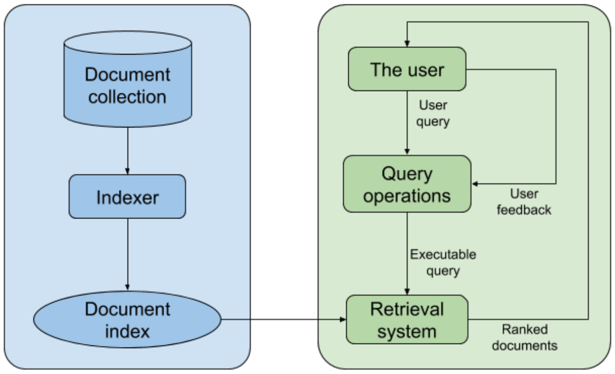


Fig. 1. Natural Language Information Retrieval Process.

A search engine is an IR system that returns textual documents d_j from a set of documents D ordered by their relevance to a query q sent by the user. The engine combines queries and indexed documents to return a sub-collection ranked as "best matches." So, let q be a query for which a set of relevant documents belongs to D. The ranking function is defined as the ϕ function such that $\phi(q, d_j)$, representing the score amongst q and d_j. This process can be iterative if the user wants to refine a query. The process repeats, and results are modified until the user is satisfied with what he is looking for.

Document ranking helps users prioritize the examination of search results. And this is also to bypass the difficulty in determining absolute relevance. So, this further suggests that the main technical challenge of a search engine is the practical design of the ranking function. In other words, we need to define the ϕ value on the query and document pair. Each recovery strategy incorporates a specific function for document ranking. These functions can be categorized based on their mathematical basis:

- **Set-theoretic models:** represent documents as sets of words or phrases. The similarities are generally derived from set theory operations [3], [4]:
- **Algebraic models:** represent documents and queries generally as vectors, matrices, or tuples. The similarity between a query vector and a document vector is defined as a scalar value [5], [6].
- **Probabilistic models:** treat the document retrieval process as probabilistic inference. The scores are computed as probabilities of a document to be or not be relevant to a specific query. Based on probabilistic theorems, such as Bayes' theorem [7], [8].
- **Neural models:** use shallow or deep neural networks to rank search results in response to a query. Learn language representations from raw text, bridging the gap between the query and document vocabulary [9].

III. Overall Cognitive-Search System

We use an application interface capable of receiving, interpreting, and answering questions requested by the user. The application programming interface (API) developed in Python enables users to retrieve documents based on a query. The user can refine the research by the type of data set: abstracts or paragraphs of the full text, the year of publication from which the query should be restricted, and the type of search engine (detailed in section IV). The response messages to the query, expressed in JSON, are exposed over the Internet.

The search engine is available on a web page with a translation for two idioms besides the Portuguese (original): English and Spanish.

The front end was developed in HTML (HyperText Markup Language), CSS (Cascading Style Sheets), and JS (JavaScript), allowing a quick execution with the help of libraries and structures like jQuery - which facilitates the implementation of JS code on the website - and bootstrap 4.0 - one of the HTML, CSS and JS libraries to customize responsive websites fastly.

The service built by the back-end was developed in PHP (Hypertext Preprocessor), Python, and Sh (Shell Command Language). It was conceived for its ease of implementation in PHP, compatibility with all libraries and frameworks used, and the generation of a log file to control the queries on the website and create the database with MongoDB. As the search tool was implemented in Python, and the server operating system was Unix, it was necessary to use scripts in GNU Bash or simply Bash (Bourne-again shell) - a Unix shell command language to communicate with the server. The basic workflow patterns in Figure 2.

IV. Ranking Methods

A. Term-Frequency Search

Probabilistic models are widely used to measure the probability of relevance of a document, combining the frequency and specificity of the term.

In information retrieval, Okapi BM25 [7] is a ranking function used to estimate the relevance of documents to a given search query. It includes the name of the first system to use, the Okapi information retrieval system. BM25 represents state-of-the-art TF-IDF functions used in document retrieval.

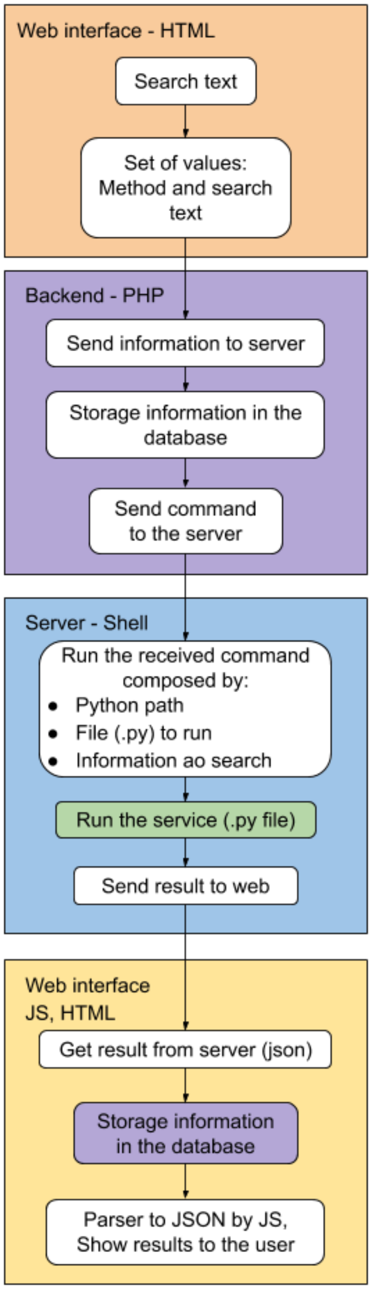

Web interface - HTML
Search text
Set of values: Method and search text
Backend - PHP
Send information to server
Storage information in the database
Send command to the server
Server - Shell
Run the received command composed by:
• Python path
• File (.py) to run
• Information ao search
Run the service (.py file)
Send result to web
Web interface JS, HTML
Get result from server (json)
Storage information in the database
Parser to JSON by JS, Show results to the user

Fig. 2. System operation flowchart.

Given a query Q, containing keywords $q_1, ..., q_n$, the BM25 score of a document D is:

$$\text{BM25(D,Q)} = \sum_{i=1}^{n} f(IDF(q_i)) \frac{f(q_i,D)(k_1+1)}{f(q_i,D)+k_1(1-b+b\frac{|D|}{avgdl})} \quad (1)$$

where:

- $f(D,Q)$ is the number of times that the term q_i occurs in document D;
- $|D|$ is the number of words in document D;
- $avgdl$ is the average number of words per document;
- b and k_1 are hyperparameters of BM25: b controls how the length of a document affects the score, and k_1 controls how much the frequency of terms each term corresponding to the final score for a query/document pair.
- $f(IDF(q_i))$ is the weight of the IDF (inverse document frequency) of the query term calculated as $log\frac{Nn(q_i)+0.5}{n(q_i)+0.5}$. N is the number of documents in the corpus and $n(q_i)$ is the number of documents in the corpus that contains q_i.

While there are advantages in using probabilistic models, searching the corpus of scientific literature on COVID-19 with such approaches makes it difficult to identify evidence relevant to complex queries. In addition, they also require several unrealistic simplifying assumptions, such as independence between terms and documents. Users must choose query words that belong to the vocabulary of the articles. A more robust notion of similarity must consider the semantic content of the query and documents.

B. Semantic Neural Search

In the past, search engines could only provide results based on exact matches of terms. But as people have started using more natural language to express their search needs, search engines have had to get smarter and incorporate semantic search principles to rank content. Semantic search systems aim to improve accuracy by understanding the intention and context behind the query and the meaning of terms within the searchable data space.

An effective way to develop a semantic system is by utilizing neural embedding techniques like those described in [10]–[12]. In particular, [12] suggests using paragraph vectors, which represent each document with a dense vector trained to predict words in that document. We concatenate the sentence vector with several word vectors from a document to predict the next word in a given context. Both word and paragraph vectors are trained using stochastic gradient descent and back-propagation, with the word vectors being shared across documents while the paragraph vectors are unique. The paragraph vectors are inferred at prediction time by fixing the word vectors and training the new paragraph vector until convergence.

One of the key benefits of using paragraph vector, as explained in [12], is that it considers the internal structure of words while still generating numeric representations. This makes it particularly useful for morphologically complex vocabularies in specialized fields like biomedicine. By indexing the embedding and evaluating the relevance of the vector, it becomes possible to assess the similarity between documents beyond just a word-level comparison. The study used a document representation vector with 700 dimensions, which was chosen through empirical testing. The pre-trained model can be easily loaded and used to encode queries into vectors.

To effectively process textual information, it is important to determine how similar or different two pieces of text are in meaning, regardless of their syntax or vocabulary. This is known as semantic similarity [13] and can apply to words, sentences, or paragraphs. Cosine similarity is a metric that compares or ranks documents based on a given vector of query words. Let two vectors d and q, the angle θ is obtained by the scalar product and the norm of the vectors:

$$similarity = \cos\theta = \frac{\mathbf{d} \cdot \mathbf{q}}{\|\mathbf{d}\| \, \|\mathbf{q}\|} \tag{2}$$

Since the $\cos\theta$ value ranges from -1 to 1: -1 means strongly opposite vectors, 0 means independent (orthogonal) vectors, and 1 means similar (positive co-linear) vectors. Intermediate values determine the degree of similarity.

Thus, semantic search engines can comprehend language nuances beyond just keywords. This means that even if there isn't an exact match to a search query, the system can still retrieve relevant records. However, the downside to this approach is that the cosine similarity between each document and the query needs to be calculated. To address this issue, in this work, all paragraph embeddings in the database were indexed and pre-computed to minimize the problem.

C. Hybrid Term-Neural Search

When we compare short search queries to multiple documents, we often encounter ambiguity. If the user doesn't provide enough context, the search engine may struggle to provide an accurate result or rank it appropriately. To enhance the system, we offer a hybrid term-neural retrieval model.

BERT (Bidirectional Encoder Representations) [14] is a bidirectional transformer that can learn language representations from large amounts of unlabeled text data. It can be fine-tuned for NLP tasks using left and right sentence context. Its bidirectional transformer [15], the vast amounts of data to pre-train, and Google's computing power contribute to its performance. With just one additional output layer, BERT can create advanced models for language processing tasks, such as TransformerXL [16], GPT-2 [17], and XLNet [18].

We use a pre-trained BERT model adjusted in the Stanford Question Answering Dataset [2] to re-rank the abstracts retrieval by BM25. The breakthrough is a BERT model that considers the full context of a word by looking at the words that come before and after to understand the intention of the search query and search perspective.

The query question and text segment obtained from abstract retrieval are inputted into the model as the first and second text sequences to improve BERT for question-answering search engines. To determine the starting position of the text span, a fully-connected layer transforms the BERT representation of each token from the text segment at position i into a scalar score s_j. These scores are then put through the softmax operation to create a probability distribution, assigning each token position i in the segment a probability p_i of being the start of the text span. During the re-ranking stage, we calculate $Relevance(q, si)(i = 1, \ldots, j)$ for the user's query and all non-zero ranked documents retrieved with BM25. The pairs with higher ranks are then used as search results. The goal is to identify the most relevant portion of the abstract retrieval for a given query.

V. COVID-19 Open Research Dataset (CORD-19)

CORD-19 is an open research data set for the new coronavirus created through collaboration between several organizations, including the Chan Zuckerberg Initiative, the National Library of Medicine, Microsoft, and the Center for Security and Emerging Technology at Georgetown University.

The dataset, launched in March 2020, has over 30,000 scientific papers and is regularly updated. Elsevier's FTP displays about 8 million publications without full text. Figure 3 shows the distribution of documents per year.

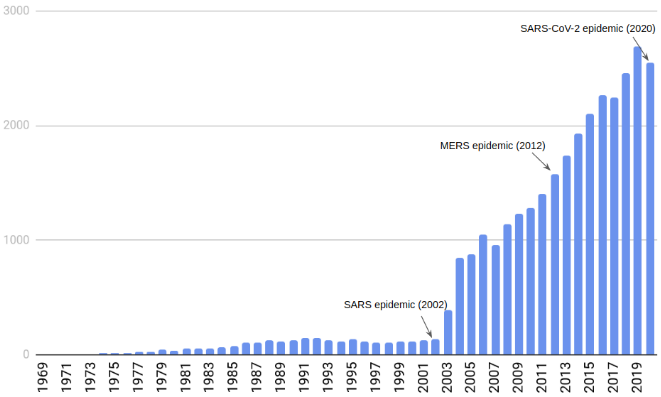


Fig. 3. The distribution of articles per year.

The COVID-19 pandemic is the biggest outbreak since the Severe Acute Respiratory Syndrome (SARS) in 2002 and the Middle East Respiratory Syndrome (MERS) that was discovered in 2012. Although COVID-19 is related to other respiratory syndromes, it has distinct pathogenetic, epidemiological, and clinical characteristics. However, we can apply the lessons we learned from the SARS and MERS epidemics to combat this new disease.

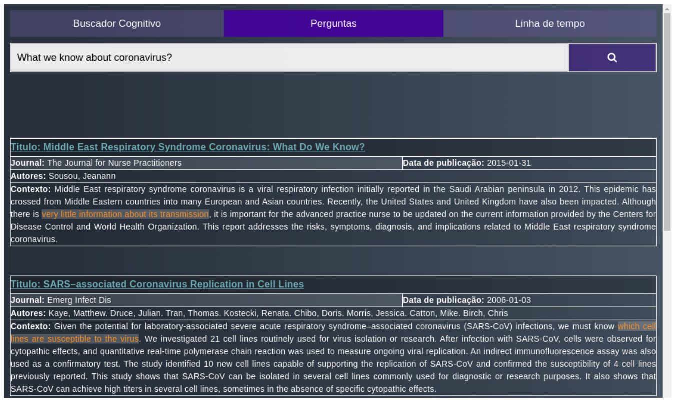
VI. Interface

Figure 4 shows how the Cognitive Search Engine works. The user types in their query and the engine shows results based on frequency (which is explained in detail in section IV.A), similarity (in section IV.B), or question-answer (in section IV.C). Users can also choose to search by paragraphs or abstracts. Thus, users can search for keywords or enter entire paragraphs to find articles with relevant content.

The researcher can also filter the results for a specific year regarding similarity and frequency. For example, a filter for publications starting in 2019 when searching about SARS-Cov-2 is acceptable.

The user can still click on the title to be redirected to the article webpage and read the entire content.

[2] SQuAD is a dataset consisting of Wikipedia articles and a set of question and answer pairs for each article.



What we know about coronavirus?

Titulo: Middle East Respiratory Syndrome Coronavirus: What Do We Know?

Journal: The Journal for Nurse Practitioners | **Data de publicação:** 2015-01-31

Autores: Sousou, Jeanann

Contexto: Middle East respiratory syndrome coronavirus is a viral respiratory infection initially reported in the Saudi Arabian peninsula in 2012. This epidemic has crossed from Middle Eastern countries into many European and Asian countries. Recently, the United States and United Kingdom have also been impacted. Although there is very little information about its transmission, it is important for the advanced practice nurse to be updated on the current information provided by the Centers for Disease Control and World Health Organization. This report addresses the risks, symptoms, diagnosis, and implications related to Middle East respiratory syndrome coronavirus.

Titulo: SARS–associated Coronavirus Replication in Cell Lines

Journal: Emerg Infect Dis | **Data de publicação:** 2006-01-03

Autores: Kaye, Matthew. Druce, Julian. Tran, Thomas. Kostecki, Renata. Chibo, Doris. Morris, Jessica. Catton, Mike. Birch, Chris

Contexto: Given the potential for laboratory-associated severe acute respiratory syndrome–associated coronavirus (SARS-CoV) infections, we must know which cell lines are susceptible to the virus. We investigated 21 cell lines routinely used for virus isolation or research. After infection with SARS-CoV, cells were observed for cytopathic effects, and quantitative real-time polymerase chain reaction was used to measure ongoing viral replication. An indirect immunofluorescence assay was also used as a confirmatory test. The study identified 10 new cell lines capable of supporting the replication of SARS-CoV and confirmed the susceptibility of 4 cell lines previously reported. This study shows that SARS-CoV can be isolated in several cell lines commonly used for diagnostic or research purposes. It also shows that SARS-CoV can achieve high titers in several cell lines, sometimes in the absence of specific cytopathic effects.

Fig. 4. Web Cognitive Search Engine in Action.

VII. Discussion of Results

When developing the COVID-19 Research Explorer, we encountered an obstacle due to the use of a specialized language in biomedical literature. As a result, some articles may not cater to the interests of general users. Additionally, as our collection includes both peer-reviewed and pre-print work, the results may not always be as rigorous.

Two main components must be considered to determine the effectiveness of information retrieval systems: the user's queries and the system's ability to provide relevant responses. Therefore, we will comprehensively examine the CORD-19 database and evaluate our web service qualitatively.

A. CORD-19 Exploration

To improve the search process, humans must recognize the specific query terms relevant to a document. To help understand the information available in the çcollected CORD-19 corpus, we have made LDA (Latent Directory Allocation) [19] accessible on the full text from CORD-19 [3]. LDA is a topic model that aims to determine the high-level meaning of documents through a set of representative words based on co-occurrence patterns of words in documents. These words are used to identify the subject area of a document. Figure 5 provides examples of topic words generated through LDA. We hypothesize that the topics of retrieved documents can help identify important aspects of user information needs.

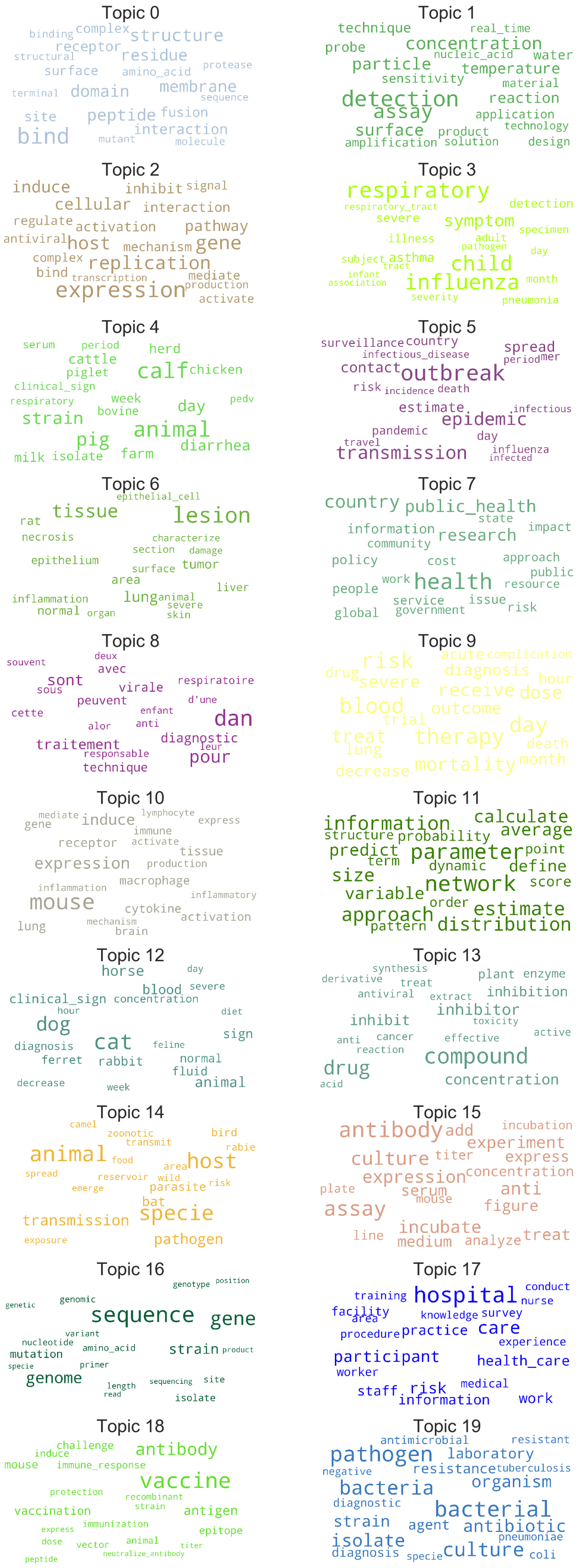


Fig. 5. LDA on CORD-19.

[3] http://www.iacontracovid.com.br/analise-de-dados/

We could identify twenty topics. For example, topic 5 suggests a cluster of documents about transmission, and topic 16 about coronavirus genome information. Each document is related to a degree of similarity for each topic. In Figure 6, notice the cluster of topics and the similarity that they share amongst them.

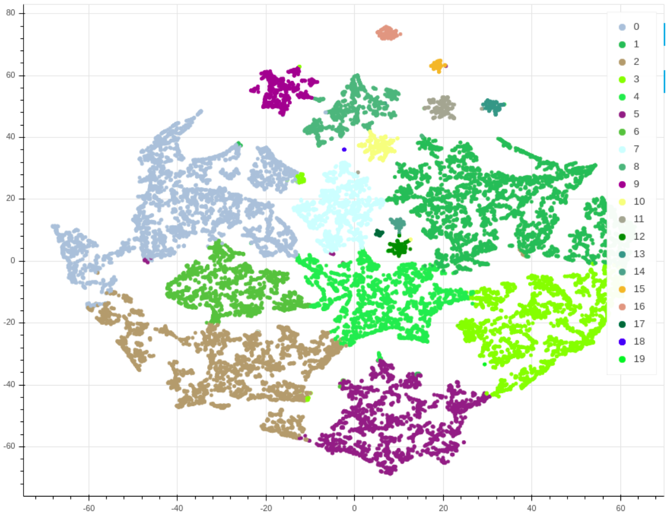


Fig. 6. t-SNE clustering of 20 LDA topics.

Computers can use Artificial Intelligence to mimic how people assign meanings or concepts to words. This feature lets users analyze vast amounts of data and discover relationships and analogies between words.

The accuracy of these analyses largely depends on the source of information. Through exploration, knowledge can be extracted from numerous texts with minimal specialist intervention. This type of analysis represents words as vectors with multiple dimensions, highlighting the resources defining each word's concept. The key on Word2Vec [20] is identifying words that share common contexts in the corpus and placing them in close proximity compared to other words in the vector space. Trained word2vec models with 50, 100, and 300 dimensions are available on the internet [4], including vectors and metadata for responses on the Embedding Projector [21] [5]. These models are accessible in binary files for use with Gensim [22].

In Figure 7, one can observe clusters of words that are associated with interleukin-6, chemicals, and COVID-19, along with clusters of authors.

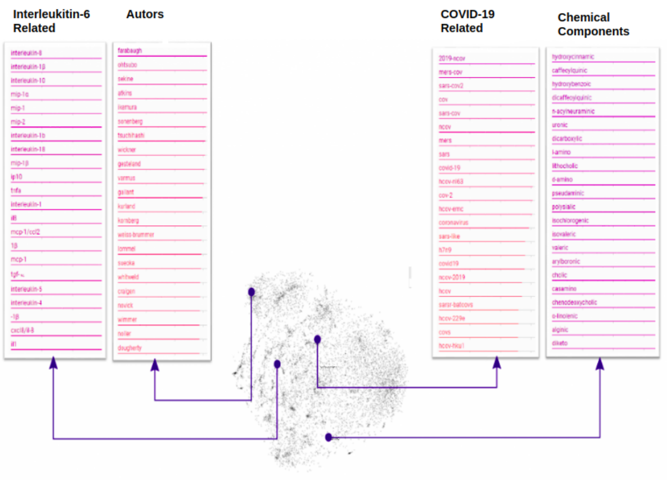


Fig. 7. A t-SNE projection of Word2vec to visualize words in CORD-19

B. System Evaluation

For a user evaluation of the web service, we assign a search task and compare the results using the three approaches described in session IV. Table I has shown the results of the query ”How does COVID-19 spread?”:

In the example in Table I, we noticed variations in how search engines disseminate information about the COVID-19 outbreak. However, we should expect such discrepancies as each search engine has its own set of priorities when it comes to ranking results.

We have observed that queries that are not well-defined can lead to unexpected responses. This is particularly true in Neural Search, where cosine similarity rank functions are used. Since this type of search can be imprecise and not necessarily point to a specific document in the vector space, the query embedding vector may be far away from the embedding vector of all documents in the data space. In contrast, when searching through the existing corpus of COVID-19 scientific literature using BM25, it can be challenging to find relevant evidence for the same queries, which may result in the engine providing insufficient or excessive results.

The overall conclusion that can be drawn from the experiments is that BERT-based re-ranking seems to overcome the other ones returning a more self-contained answer. We hypothesize that this approach works because combining different strengths from BERT and BM25. The BERT benefits from deep semantic understanding based on next-sentence prediction context, while BM25 identifies higher relevance through query term repetition.

Although the contextualized language models yield ranking improvements, users generally do not know well about the information they seek. And in that case, even if the Hybrid Term-Neural Search is accurate, the user also can still want to rend documents retrieval by Term-Frequency and Neural-Similarity, regardless of relevance is a matter of degree. For example, for COVID-19 spread, the user could also want to know about the range of incubation periods for the disease,

[4] http://www.iacontracovid.com.br/modelos-covid/
[5] https://projector.tensorflow.org/

TABLE I
RESULTS FOR THE QUERY ”HOW DOES COVID-19 SPREAD?”

Term-Frequency	***Neural Semantic Search***	***Hybrid Term-Neural Search***
Title: COVID-19, SARS, and MERS: are they closely related? **Journal:** Clinical Microbiology and Infection **Publish Year:** 2020 **Content:** The 2019 novel coronavirus (SARS-CoV-2) is a new human coronavirus which is spreading with epidemic features in China and other Asian countries with cases reported worldwide. This novel Coronavirus Disease (**COVID-19**) is associated with a respiratory illness that may cause severe pneumonia and acute respiratory distress syndrome (ARDS) (...) The gastrointestinal route of transmission of SARS-CoV-2, which has been also assumed for SARS-CoV and MERS-CoV, cannot be ruled out and needs to be further investigated. Implications There is still much more to know about **COVID-19**, especially as concerns mortality and capacity of spreading on a pandemic level. Nonetheless, all of the lessons we learned in the past from SARS and MERS epidemics are the best cultural weapons to face this new global threat.	**Title:** The Risk and Prevention of Novel Coronavirus Pneumonia Infections Among Inpatients in Psychiatric Hospitals **Journal:** Neuroscience Bulletin **Publish Year:** 2020 **Content:** Since the middle of December 2019, human-to-human transmission of novel coronavirus pneumonia (NCP, also called **COVID-19**) has occurred among close contacts [1]. After the outbreak on January 21, 2020, it was swiftly included among the Class B infectious diseases stipulated in the Law of the People’s Republic of China on the Prevention and Control of Infectious Diseases, and measures for prevention and control of Class A infectious diseases were adopted. At 21:27 on February 12, 2020, the China News Network updated information to include epidemic data from the National Health Commission and official channels in Hong Kong, Macao, and Taiwan regions: the highest death rate was in Wuhan City (Table 1). Overload of inpatients at hospitals may play a negative role in the overall therapeutic effect and contribute to the death rate.	**Title:** Coronavirus Disease 2019 (COVID-19): Protecting Hospitals From the Invisible **Journal:** Annals of Internal Medicine **Publish Year:** 2020 **Content:** Coronavirus disease 2019 (**COVID-19**) is optimized to **spread** widely: Its signs and symptoms are largely indistinguishable from those of other respiratory viruses. This commentary specifically addresses the best ways to protect our hospitals against **COVID-19**.

seasonality of transmission, immune response, or maybe the role of the environment in the spread.

The Hybrid Term-Neural Search approach also allows using temporal information in documents to anchor search results along a timeline. This means that the retrieval of text relevant to a specific time can be enabled by extracting and representing temporal events mentioned in a document. However, the idea of time in the retrieval context is often associated with how a piece of information changes over time, which promotes freshness in results.

When the user asks an initial question, the engine returns a set of documents and highlights pieces from the article that are potential answers. The user can review the pieces and quickly decide whether or not that article is worth reading. Table II shows the results retrieved for the query “Can animals transmit 2019-nCoV?”.

From Table II, it is evident that the response remains consistent over time. To summarize our findings, we can conclude that the coronavirus can spread from a wild animal to an animal, from an animal to a human, and among animals. Examining the time context can offer valuable perspectives for analyzing text retrieval data. We can scrutinize responses during each period and compare them to gain insights into knowledge evolution.

The graph in Figure 8 displays the time distribution of seven query answers regarding the coronavirus outbreak in the database. It is important to note that the timeline is constantly updated with new articles being added. The graph is a stacked bar chart.

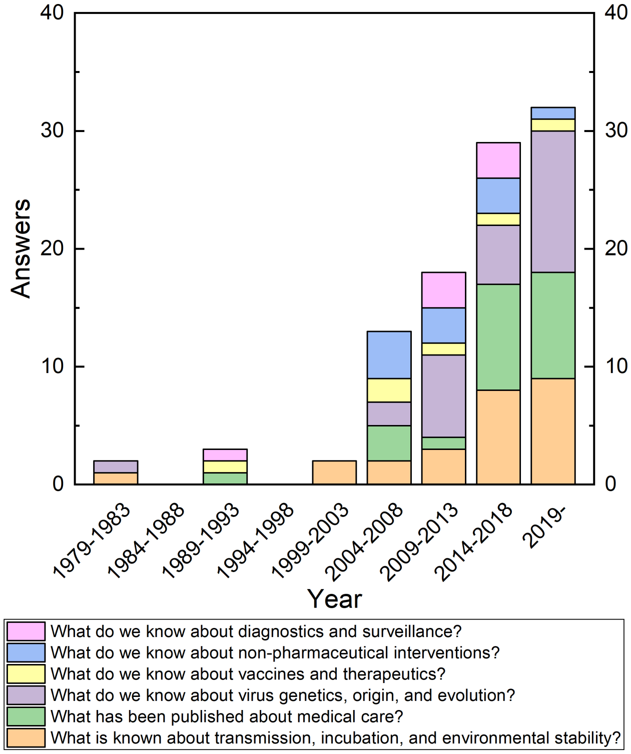


Fig. 8. Time Distributed Answers by Query Question

TABLE II
TIMELINE RESULTS FOR THE QUERY "CAN ANIMALS TRANSMIT 2019-NCOV?"

Query results publish time	***Extracted answer result from the context***
2009-12-03	"infections that transmit from animals to humans"
2013-01-20	"rapidly transmit between animals"
2013-02-07	"recombinant viruses engineered to transmit across synapses"
2016-04-30	"they represent a potential reservoir of viruses that transmit from wildlife to humans or domestic animals"
2020-02-11	"the virus should have been identified from animals sold at the market"
2020-02-20	"could be of bat origin but involve other potential intermediate hosts"
2020-03-14	"may be wild animals"

VIII. CONCLUSION AND FUTURE WORK

Our AI Web Service offers a visual and intuitive way to find complex connections in searches using probabilistic models, neural networks, and transformers. Get the relevant information you need, faster.

At its core, performing searches requires understanding the language. We could use the same word with different meanings without problems. But for computers, ambiguity is the main difficulty. This is because natural languages are designed to make our communication efficient and were not designed for computers.

Automatic retrieval of high-quality information from text refers to some combination of relevance, novelty, and interest. Typically, this includes text categorization, concept/entity extraction, production of granular taxonomies, and entity relationship modeling. These strategies aim to lead the Search to the syntactic-semantic levels of information retrieval, reducing the loss of fundamental meanings about the scope where the words are inserted and affected by the query intent. Indeed, an intelligent search engine based on natural language processing must have better and better applications.

ACKNOWLEDGEMENT

This work could not have been done without the support of Capes, Intel, and Petrobras. The authors also would like to thank the collaboration of the entire Applied Computational Intelligence Laboratory (ICA) team.

REFERENCES

[1] Z. Kaddari, J. Berrich, N. Rahmoun, S. Belouali, T. Bouchentouf, Inkad covid-19 intellisearch: a multilingual search engine for answering questions about covid-19 in real-time from the scientific literature, in: 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS), 2021, pp. 1–6. doi:10.1109/ICDS53782.2021.9626759.

[2] M. Pérez-Montoro, L. Codina, Chapter 5 - the essentials of search engine optimization, in: M. Pérez-Montoro, L. Codina (Eds.), Navigation Design and SEO for Content-Intensive Websites, Chandos Publishing, 2017, pp. 109 – 124.

[3] A. H. Lashkari, F. Mahdavi, V. Ghomi, A boolean model in information retrieval for search engines, in: 2009 International Conference on Information Management and Engineering, 2009, pp. 385–389.

[4] W.-S. Hong, S.-J. Chen, L.-H. Wang, S.-M. Chen, A new approach for fuzzy information retrieval based on weighted power-mean averaging operators, Computers Mathematics with Applications 53 (12) (2007) 1800 – 1819.

[5] M. Melucci, Vector-Space Model, Springer US, Boston, MA, 2009, pp. 3259–3263.

[6] M. N. Hoque, R. Islam, M. S. Karim, Information retrieval system in bangla document ranking using latent semantic indexing, in: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1–5.

[7] M. Sanderson, Christopher d. manning, prabhakar raghavan, hinrich schütze, introduction to information retrieval, cambridge university press. 2008. isbn-13 978-0-521-86571-5, xxi 482 pages., Natural Language Engineering 16 (1) (2010) 100–103.

[8] L. Park, K. Ramamohanarao, The sensitivity of latent dirichlet allocation for information retrieval, Vol. 5782, 2009, pp. 176–188.

[9] T. A. Nakamura, P. H. Calais, D. de Castro Reis, A. P. Lemos, An anatomy for neural search engines, Information Sciences 480 (2019) 339 – 353.

[10] J. Turian, L.-A. Ratinov, Y. Bengio, Word representations: A simple and general method for semi-supervised learning, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 384–394.

[11] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space (2013).

[12] M. Pagliardini, P. Gupta, M. Jaggi, Unsupervised learning of sentence embeddings using compositional n-gram features, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 528–540.

[13] K. Blagec, H. Xu, A. Agibetov, M. Samwald, Neural sentence embedding models for semantic similarity estimation in the biomedical domain, BMC Bioinformatics 20 (12 2019).

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 5998–6008.

[16] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, R. Salakhutdinov, Transformer-XL: Attentive language models beyond a fixed-length context, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2978–2988.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[18] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 5753–5763.

[19] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (null) (2003) 993–1022.

[20] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14, JMLR.org, 2014, p. II–1188–II–1196.

[21] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, M. Wattenberg, Embedding projector: Interactive visualization and interpretation of embeddings, ArXiv abs/1611.05469 (2016).

[22] R. Rehurek, P. Sojka, Software framework for topic modelling with large corpora, in: Proceedings of the LREC 2010 Workshop on new challenges for NLP frameworks, 2010, pp. 45–50.