www.cya.unam.mx/index.php/cya

Contaduría y Administración
Accounting & Management

Contaduría y Administración 71 (2), 2026, e552

Methodology for the management of scientific literature in Spanish through information retrieval approaches using natural language processing

Metodología para la gestión de la literatura científica en español mediante enfoques de recuperación de información utilizando el procesamiento del lenguaje natural

Josué Padilla Cuevas, Alma D. Cuevas-Rasgado, José A. Reyes-Ortiz*, Maricela Bravo

Universidad Autónoma Metropolitana, México
Universidad Autónoma del Estado de México, México

Received November 05, 2023; accepted April 9, 2024
Available online February 17, 2026

Abstract

The scientific papers are written in natural language, with a significant proportion being in Spanish, and have no structure processable by computers, which results in tedious and time-consuming manual analysis. Thus, managing scientific texts in Spanish is a challenge that requires advanced computational methods. Therefore, this paper presents a novel methodology that includes three Information Retrieval (IR) approaches based on Natural Language Processing (NLP). The main aim is the information management from scientific documents in Spanish. The IR approaches implemented in the methodology are based on textual, probabilistic, and semantic similarity to retrieve documents regarding a question. The proposed methodology is applied to the scientific Spanish literature generated during the COVID-19 pandemic. An evaluation process based on 100 queries over 249,474 scientific documents to accurate the recoverability of relevant documents was carried out. The results show that the probabilistic approach implemented in the methodology achieved an 85% f-measure, supported by the Latent Dirichlet

*Corresponding author.
E-mail address: jaro@azc.uam.mx (José A. Reyes-Ortiz).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

https://doi.org/10.22201/fca.24488410e.2026.5282

Allocation (LDA) topic discovery algorithm. Finally, the proposed methodology is considered domain-independent to retrieve documents in Spanish.

JEL Code: C63, C67, I12, L86
Keywords: information management; natural language processing in spanish texts; text-similarity, semantic and probabilistic approaches; recovering scientific documents

Resumen

Los artículos científicos están escritos en lenguaje natural, una proporción significativa en español, y carecen de una estructura procesable por las computadoras, lo que provoca un análisis manual tedioso, laborioso y lento. De este modo, la gestión de textos científicos en español es un reto que requiere métodos computacionales avanzados. Por ello, este trabajo presenta una metodología novedosa que incluye tres enfoques de Recuperación de Información (RI) basados en el Procesamiento del Lenguaje Natural (PLN). El objetivo principal es la gestión de información a partir de documentos científicos en español. Los enfoques de RI implementados en la metodología se basan en la similitud textual, probabilística y semántica para recuperar documentos en relación con una pregunta. La metodología propuesta es aplicada a la literatura científica en español generada durante la pandemia de COVID-19. Además, se lleva acabo un proceso de evaluación basado en 100 consultas sobre 249.474 documentos científicos para precisar la recuperabilidad de los documentos relevantes. Los resultados muestran que el enfoque probabilístico implementado en la metodología alcanza una medida F del 85%, apoyado en el algoritmo de descubrimiento de tópicos Latent Dirichlet Allocation (LDA). Por último, la metodología propuesta se considera independiente del dominio para recuperar documentos en español.

Código JEL: C63, C67, I12, L86
Palabras clave: gestión de información; procesamiento de lenguaje natural en textos en español; enfoques de similitud textual, semántica y probabilística; recuperación de artículos científicos

Introduction

Scientific article repositories such as PubMed, IBECS, SCIELO and LILACS provide an excellent source of knowledge for the research community, which needs to manage accurate information quickly and efficiently. The PubMed database is estimated to contain more than 19 million documents and searching for information on a specific topic in this large volume of text can be very time-consuming. This paper proposes a methodology for managing scientific literature in Spanish through Information Retrieval approaches using Natural Language Processing. Three IR models have been implemented based on probabilities, semantics and vector spaces. PLN techniques have been used to reduce the information search space to retrieve documents relevant to a query in a minimum time and efficiently. The proposed methodology is applied to scientific documents in Spanish generated during the COVID-19 pandemic period, showing promising accuracy.

Formal description

Given a set of scientific documents in Spanish D = {d1, d2, …, dn}, and a set of user queries Q = {q1, q2, …, qm}, the objective of an IR methodology is to return an ordered list of documents such that the documents placed at the top are more relevant to a given q Q than those placed at the bottom. The relevant documents are what the user is looking for and can be considered to contain the expected information to the query.

As a solution to the stated issue, the present paper outlines a methodology that implements, evaluates and compares three approaches based on IR, NLP and Topic Discovery, with the aim of managing scientific literature in Spanish.

Baeza-Yates & Ribeiro-Neto (1999) defines IR as representing, storing, organizing, and accessing information items and Korfhage (1997) mentions that it is the location and presentation to a user of relevant documents to an information need expressed as a question. IR can be said to provide users with a ranked list of documents that meet their information needs. On the other hand, NLP provides well-known techniques for syntactic, semantic, and morphological analysis of scientific literature to reduce its dimensionality and find relevant information in real time.

The rest of this paper is organized as follows. Section 2 presents the related work, analyzing algorithms, approaches, and techniques for IR. Section 3 exhibits the proposed methodology based on IR models and NLP techniques to recover scientific documents in Spanish. Section 4 presents the methodology applied over documents generated during the COVID-19 pandemic. Finally, Section 5 provides conclusions and future work of this paper.

Related work

This section presents research advances in retrieving relevant documents from large amounts of digital information. It describes work based on datasets in different domains and languages, and discusses the methods, algorithms, approaches, and techniques used.

COVIDASK is a system that combines biomedical text mining and IR techniques to provide documents for natural language queries. Its evaluation is performed with a dataset manually created by COVID-19 experts and based on information from various sources such as: the Center for Disease Control and Prevention (CDC) and World Health Organization (WHO) (Jinhyuk et al., 2020). A SemBioNLQA semantic system for answering biomedical questions in natural language. The task IR is implemented with lexical-syntactic patterns and a probabilistic level text analysis using the BM25 algorithm. The evaluation

consists of 30 questions: 10 fact-type, 11 list-type, and 9 yes/no questions. They achieve a 60% precision (Sarrouti & El Alaoui, 2020).

In Burak Ozyurt et al. (2020), the Bio-AnswerFinder system is presented that locates relevant documents in biomedical texts. IR is performed iteratively by querying in English and using improved keywords from a traditional search engine. The system was evaluated by experts in the biomedical field and achieved 71% precision.

AskHERMES provides relevant documents corresponding to clinical question searches based on a dataset obtained from abstracts of scientific articles from MEDLINE, PubMed, eMedicine, and Wikipedia texts. It integrates the latest version of the BM25 model because of its simplicity, interpretation, and computational speed. This system was evaluated on a scale of 1 to 5 based on 60 questions asked by three experts and obtained the following results: Ease of Use 4, Quality of Answers 2.5, and Overall Performance 3 (Cao et al., 2011).

CAiRE-COVID is a COVID-19 text information retriever system. It won one of the ten tasks in the Kaggle COVID-19 Open Research Dataset Challenge. It is IR engine uses classification algorithms such as Bag-of-Words and BM25 (Su et al., 2020). In MiPACQ, the IR is implemented with the Lucene index, which uses the vector space model with normalized TF-IDF parameters created from documents of scientific articles from Medpedia (Cairns et al., 2011).

The IR task in Xu et al.(2018) is performed with documents from the biomedical domain. Two strategies are proposed: the first being document labeling that combines several traditional models such as BM25 and vector space. In addition, a second approach is used, based on machine learning methods, since there is a wealth of specific terminologies in the biomedical domain. In A. Taan et al. (2021) a CrossLingual Information Retrieval (CLIR) system is proposed that provides an answer in a language other than the query language, in this case between Arabic and English. It uses Quran as a dataset and performs a query expansion technique to improve the performance impact of the methods of IR (Boolean, Vector Space, and BM25).

Articles Otegi et al. (2022), Zhang et al. (2020) and Esteva et al. (2021) use the Allen Institute open research dataset, CORD-19. An Information Retrieval System (IRS) response to sanitary experts queries on COVID -19 is presented in Otegi et al. (2022). The approach used for IR is based on language models, feedback, and Bayesian networks. Zhang et al. (2020) shows the Covidex information search engine that provides a robust infrastructure based on English language keywords and machine learning techniques. The system was evaluated in the TREC-COVID challenge and received the highest score. Finally, CO-Search, a multi-level semantic search system, is designed to handle coronavirus queries. It includes a keyword retriever that accepts an input query and returns an ordered list of relevant documents.

It uses a deep learning approach along with models BM25 and TF-IDF. The TREC competition dataset was used to evaluate the effectiveness of the search system (Esteva et al., 2021).

In Badenes Olmedo et al. (2021), an investigation of the influence of text length on IR tasks is shown using a Probabilistic Topic Model (PTM). The goal is to reduce the dimensions of the vectors in IR models to simplify the operations without affecting the system performance. The paper Berger & Lafferty (2017) shows a new probabilistic IR approach based on statistical machine translation methods. The main idea is to transform a document into an ideal query. The TREC competition dataset is used to evaluate this work.

The medical information retrieval system Wang et al. (2017) considers semantic associations and applies a heuristic approach to query feedback to improve retrieval in structured knowledge bases. The authors use the PubMed digital database for Information Retrieval. The performance of the BM25 model was compared with the Boolean model and the vector space model in Asnath Tinega et al. (2018). The results showed that the documents best ranked as relevant were determined using the OkapiBM25 probability-based ranking algorithm.

COSE is a search engine for a corpus of scientific publications. It contains an algorithmic system that returns relevant documents to COVID -19 queries based on TF-IDF and a Transformers model, achieving 71.45% precision (Raza, 2022).

CovBERT proposes to perform IR using a deep learning-based text classifier. It uses the pre-trained Bert language model and performs fine tuning using the COV-Dat-20 dataset consisting of 4304 articles divided into four categories: COVID-19, Virology, Public Health and Mental Health. CovBERT achieved an accuracy of 94% in the text classification task for IR (Khadhraoui et al., 2022).

In Mahmood et al. (2019), multimedia information retrieval (image, audio, video, and text) is used. Document retrieval is performed using vector space models, LSI, and a neural network approach. On the other hand, in Gong et al. (2021) visual semantic embedding (VSE) networks are used in the image-text retrieval task on a Flickr30K dataset. The proposed UNITER network achieved 61.5% recall for the task of retrieving the relevant text descriptions in the images. While the traditional networks VSRN, SCAN and VSE++ achieved 50.3%, 47.1% and 29.4% recall respectively for the same task. Finally, in the paper Rodriguez & Carver (2019), a comparison is made to evaluate the performance of vector space IR, probabilistic, and LSI algorithms. The results show that the LSI technique has low precision and recall compared to the probabilistic and vector models.

Table 1 presents a synthesis of the works described in this section. It shows that vector space, probabilistic BM25, and semantic LSI models are currently widely used for information retrieval tasks involving large volumes of documents of documents in different languages and domains. On the other hand, most work focuses on the study of documents in English, but Spanish is one of the five most widely

spoken languages in the world (Gordon, 2005). Therefore, there is a need for the management and retrieval of scientific texts in Spanish.

This paper presents a methodology based on IR approaches: textual similarity, probabilistic and semantic, as well as text processing variants and a topic discovery algorithm for adequate and optimized information retrieval, management, and use, whether at a personal or organizational level.

Table 1
Related works comparison.

Reference	Language	Dataset/Source	Technique/approach	Evaluation
(Sarrouti & El Alaoui, 2020)	English	Biomedical documents PubMed	Lexical-syntactic patterns BM25 algorithm	Precision 60%
(Burak Ozyurt et al., 2020)	English	Biomedical abstracts PubMed	Improved keywords from a traditional search engine	Precision 71%
(Cao et al., 2011)	English	Scientific articles from MEDLINE	BM25 model	Precision 60%
(Su et al., 2020)	English	CORD-19 Open Research Dataset	Anserini with Lucene index, BM25 model	Precision 10% Recall 22% F-measure 13%
(Cairns et al., 2011)	English	Medpedia	Lucene index TF-IDF	Precision 84%
(Xu et al., 2018)	English	Biomedical documents	BM25 model machine learning methods	NA
(A. Taan et al., 2021)	Arabic	Quran	Boolean, Vector Space, and BM25	Precision 72%
(Otegi et al., 2022)	English	CORD-19	Language models, feedback, and Bayesian networks	Recall 85%
(Berger & Lafferty, 2017)	English	TREC competition dataset	Probabilistic approach based on statistical machine translation	Precision 71%
(Wang et al., 2017)	English	Medical Knowledge bases	Heuristic approach	Precision 41%

(Asnath Tinega et al., 2018)	English	PubMed digital	BM25 model	Precision 62.2% Recall 78.5%
(Raza, 2022)	English	Scientific publications	TF-IDF Transformers model	precision 71.45%
(Khadhraoui et al., 2022)	English	COVID-19 Research Dataset	Deep learning-based text classifier	Accuracy 94%
(Mahmood et al., 2019)	English	image, audio, video (Flickr30K dataset)	vector space models, LSI, and a neural network	Recall 61.5%

Source: Author's own.

Proposed methodology

This section outlines a methodology for managing scientific literature in any field of research in Spanish, through Information Retrieval approaches using Natural Language Processing. The methodology is organized by phases that are detailed described below, starting with the search and selection of suitable documents in Spanish from which to extract information. Subsequently, it is necessary to carry out the processing of the texts in order to clean up the set documents and to prepare them for better management and thus reduce processing time and computing resources. After cleaning the document text, document retrieval is realized to transform them into vectors as input to the algorithms; they categorize the retrieved document according to their relevance. Finally, an automatic evaluation based on queries is carried out. Figure 1 depicts the general methodology for recovering Spanish scientific documents related to an input question.

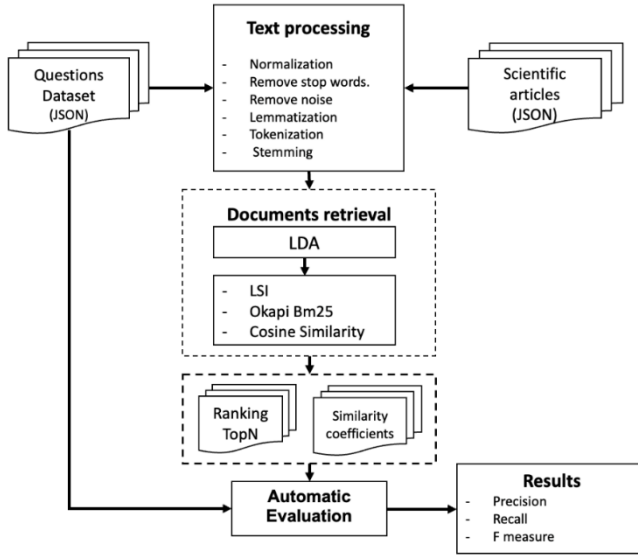

Figure 1. Proposed methodology to retrieve Scientific documents in Spanish.
Source: Author's own.

Search and selection of the dataset

In this phase, a dataset is selected from repositories of scientific documents in unstructured format containing a large amount of information in any area of interest, such as social sciences, arts and humanities, life sciences and biomedicine, physical sciences, or technology.

Text processing

Text processing aims to clean up documents and queries to manage information and thus reduce processing time and computing resources. The tasks of tokenization, conversion to lowercase, and removal of noise and stop words are performed. Lemmatization and stemming techniques are also applied in this phase.

- Tokenization. Divides documents into basic units or words to reduce complexity and simplify further processing.
- Noise removal. Removes any fragments, words, or characters that are not relevant to the context of the data and are commonly referred to as noise that affects algorithm performance.

- Elimination of stop words. Deletion of words that are commonly used in the language and have low semantic content.
- Conversion to lowercase. This task is responsible for converting texts into lower-case letters to standardize them.
- Lematization. It converts nouns to their singular form in the masculine gender and verbs to the infinitive. The goal of this task is to standardize and unify the different inflections or conjugations, reducing the vocabulary of scientific documents. Morphological analysis and Spanish dictionary are used for this task.
- Stemming. The roots of words are determined to reduce the vocabulary of Spanish texts. The Snowball Stemmer is used for this task.

NLP processing configurations

In the text processing phase, different text configurations are used to test their impact on the retrieval of relevant documents. In particular, the following combinations will be examined:

a) Simple processing consisting of tokenization, lower case conversion, removing noise and stop words.
b) Processing and lemmatizing
c) Processing and stemming

The purpose of implementing lemmatization and stemming techniques in Scientific texts is to reduce the search space for retrieving relevant documents that meet the information needs of researchers. This is achieved by reducing morphological variants of word forms to common roots or lexemes, thereby improving the ability to search and respond to queries in real time.

Documents retrieval

In this phase, relevant documents are retrieved in relation to a particular question. A document is deemed relevant if it satisfies the information requirements of a user request, meaning it offers comprehensive, precise, and prompt responses to the inquiry made (van Rijsbergen, 1998). Additionally, a topic discovery algorithm is utilized to decrease the dimensionality of the texts. This process necessitates employing an algorithm to determine the similarity (or distance) between the set of documents and the question. This methodology implements three approaches to retrieve documents regarding a question: textual,

probabilistic, and semantic similarity. Both, a topic discovery process and the IR approaches are described as follow.

Topic discovery algorithm

The methodology incorporates innovative features by integrating IR techniques with a topical discovery algorithm. It is employed to narrow down the search area for pertinent data, as not all records in the dataset pertain to the sought-after subject. The Latent Dirichlet Allocation (LDA) algorithm is implemented for this task because according to the literature it is one of the most efficient algorithms. It is based on a simple assumption of interchangeability of words and topics in a document. It is a dimensionality reduction technique with probabilistic semantics (Blei et al., 2003). The algorithm has been implemented using the techniques described in the STTM paper Jipeng et al. (2022).

Text-based similarity approach

It was proposed and developed by Salton et al. (1975). It uses a matrix containing the vocabulary of a collection of documents and the query. The main idea is to calculate the similarity between the query vector and the document vectors. Different distance formulas are used for this task: the cosine similarity, Jaccard similarity and the Dice coefficient. The implementation of the cosine similarity metric in this phase is due to its proficiency in handling data of variable length, including a scientific article and a question. The approach Equation 1 considers a vector space consisting of N documents Di, each identified by one or more terms Cim that can be set between 0 and 1 depending on the degree of relevance (Wong et al., 1985). A document is a vector that relates to each occurrence of words within the text.

$$D_1 \rightarrow \vec{d} = (C_{i1}, C_{i2}, \cdots, C_{im}) \tag{1}$$

Equation 2 is used to calculate the weight value of each term in the document collection, which expresses the numerical measure of the frequency of words in the documents between the inverse frequency of the term in the collection, also known as TF-IDF.

$$\text{tfidf}_{i,j} = \text{tf}_{i,j} \times \log \frac{N}{\text{df}_i} \tag{2}$$

Equation 3 is utilized to calculate the cosine angle similarity. This normalizes the vectors with respect to their length, if the cosine of the angle is 1, it means that the query and the document are equal (the vectors are parallel), otherwise it means that the vectors are orthogonal and there is no relationship between them.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3}$$

Probabilistic similarity approach

It is based on the principle of sorting by probabilities, where documents are classified as relevant or irrelevant based on a query (Ross, 2014). It assumes that if there is an ideal answer to the query posed, it will be found in the relevant documents. The ideal answer is initially unknown, but it is known to be in a probability combination between 0 and 1. For this approach, the initial default values are 0.5 and are called "Maximum Uncertainty", which means that the document with the answer can be in the relevant or irrelevant set.

This phase implements the Okapi probabilistic model, which considers the relative information of the term frequency with respect to the documents and the query from a probabilistic point of view using the 2-Poisson distribution (Trotman et al., 2014). Given a query Q with keywords q1, ..., qn, the BM25 score of a document D is defined as follows (see Equation 4).

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \tag{4}$$

Where f(qi, D) is the frequency of the term qi in document D, |D| is the number of words contained in document D, avgdl is the average number of words present in the document collection, k1 and b are parameters that have default values of 1.2 and 0.75, respectively. Finally, IDF (qi) is the inverse frequency of the query term qi and is calculated using Equation 5.

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \tag{5}$$

Where N is the number of documents in the corpus and n(qi) is the number of documents containing the query term qi (Lee, 2007).

Semantic similarity approach

In this phase, is implement a semantic approach to information retrieval (IR), utilizing the latent semantic indexing (LSI) method. which is based on indexing terms, placing them in a semantic context to avoid phenomena such as synonymy or polysemy. It allows retrieving information about the meaning of a document, overcoming the problems of lexicon matching. It assumes that there is an underlying or latent structure in the use of a word that is partially hidden by the variability in the choice of terms (Deerwester et al., 1990). LSI is a statistical method that estimates a latent structure and uses singular value decomposition, which segments a large matrix of term-document association data and allows the construction of a "semantic space" (Singular Value Decomposition SVD) (Forsythe et al., 1977). Equation 6 shows the SVD, where M is the word frequency matrix, K is the orthogonal matrix, S is the diagonal matrix, and D is the transposed matrix. The SVD provides an orthogonal matrix of the word vector, a transpose representing the document vector, and a diagonal representing the singular value of the word-document correspondence.

$$M = KSD^t \tag{6}$$

Evaluation

The presented automatic evaluation is based on the Precision, Recall and F-measure metrics. The evaluation process uses a set of questions in Spanish, a collection of documents and a file with the similarity coefficients generated from the previous phase. Different configurations of experiments are also used to obtain the best results, combining IR approaches, topic discovery process, NLP text processing tasks and the number of top documents to be retrieved.

Evaluation metrics

In this phase the performance of information retrieval needs to be evaluated, the level of performance needs to be measured against some scale, some parameters to measure performance are effectiveness and

efficiency. Metrics are used to provide a measure of the degree of retrievability of scientific texts. The retrievability of documents does not provide an exact answer to the given query and must be ranked according to their degree of relevance to the query (Cyril & Michael, 1966). The evaluation of this ranking is achieved by means of metrics such as Precision, Recall and F-measure, which are well-known in the literature and widely used in the IR task to enable accurate and efficient information management.

Precision is defined as the fraction of retrieved documents that is relevant and provides a measure to avoid irrelevant documents or commonly called noise; Equation 7 is used.

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} \tag{7}$$

Recall is the fraction of relevant documents that were retrieved and is a measure for the retrieval of relevant documents, or commonly referred to as silence (see Equation 8). For this metric, a term has been added to determine if the relevant document is in the top ranked list.

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} \tag{8}$$

The F-measure (Equation 9) is a harmonic measure combining the metrics of precision and recall.

$$\text{F measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

Experimental setup

To evaluate the retrievability of documents, the following experiments and configurations are performed based on a query:

1. Similarity-LDA. The textual similarity approach was configured with processed, processed-lemmatization, processed-stemming. Also, top candidates or best-positioned documents are tested in 3, 5, 10, 15 and 20 with LDA.
2. Probabilistic-LDA. The probabilistic similarity approach was configured with processed, processed-lemmatization, processed-stemming. Also, top candidates or best-positioned documents are tested in 3, 5, 10, 15 and 20 with LDA.

3. Semantic-LDA. The semantic similarity approach was configured with processed, processed-lemmatization, processed-stemming. Also, top candidates or best-positioned documents are tested in 3, 5, 10, 15 and 20 with LDA.

Case study methodology in COVID-19 Spanish documents

The proposed methodology can be applied to any research area for effective management of scientific documents in Spanish. A digital repository with a large volume of documents is required to find timely, accurate, complete, and relevant information in response to user requests. This article presents a case study on the COVID-19 pandemic. During this period, researchers' information needs increased due to the large number of scientific articles published in a short time. The study aims to provide insights into the challenges faced by researchers in managing the overwhelming amount of information available.

For the search phase of the dataset, the BioASQ archive of scientific literature in Spanish was selected. This dataset consists of 249,474 articles from the medical domain, which have been human validated and constructed using publications from different scientific repositories such as IBECS and LILACS. The scientific texts are formatted in JSON and comprise the following tags: id, title, year, abstractText, db, and decsCodes. A sample of the BioASQ dataset is illustrated in Figure 2.


```
{"articles": [
{"id": "ibc-ET6-1764",
"title": "La Quercetina como un potencial nutracéutico contra la enfermedad por coronavirus 2019 (COVID-19)",
"abstractText": "INTRODUCCIÓN: La enfermedad del coronavirus 2019 (COVID-19) es una enfermedad viral que afecta
a varios órganosy sistemas. Los tratamientos preventivos o profilácticos son especialmente útiles en enfermedades
infecciosas emergentes como COVID-19 porque reducen la necesidad de hospitalización y el gasto en salud pública.
Aunque el efecto preventivo del SARS-CoV-2 de varios agentes terapéuticos (e.g., hidroxicloroquina/cloroquina,
remdesivir,lopinavir y ritonavir) se ha evaluado ampliamente, ninguno de ellos ha demostrado una gran eficacia
clínica. MÉTODO: Por lo tanto, aquí nuestro objetivo es abordar y discutir los estudios publicados recientemente
sobre el potencial quimioprofiláctico de la quercetina contra el SARS-CoV-2. METODOLOGÍA: Se realizó una búsqueda
de la literaturaen bases como PubMed/MEDLINE,Scielo, Scorpus, Web of Science, Cochrane Library y Clinical Trials.gov.
Se incluyeron y evaluaron críticamente estudios que abordan la quercetina contra el SARS-CoV-2 u otros tipos
decoronavirus. RESULTADOS: Algunos estudios han demostrado que la quercetina, un flavonoide aprobado por la FDA
que se utiliza como agente antioxidante y antiinflamatorio, inhibe la entrada del coronavirus (SARS-CoV) en la célula
huésped.Además, un estudio in silico mostró que la quercetina es un potente inhibidor de la proteasa principal del
SARSCoV-2 (Mpro),lo que sugiere que este flavonoide también es activo contra COVID-19. CONCLUSIONES: Debido a que la
quercetina podría prevenir y disminuir la duración de las infecciones por SARSCoV-2, es plausible suponer que el uso
profiláctico de este flavonoide produce varios beneficios clínicos. Pero, estas pruebas preliminares deben ser
confirmadas mediante ensayos in vitro y, posteriormente, en un ensayo clínico aleatorizado",
"journal": "Ars pharm",
"year": 2021,
"db": "IBECS",
"decsCodes": ["D006801","D018352","D015203","D019587","D065129","D000998","D011480","D011794"]
```


Figure 2. Scientific text from the BioASQ dataset.
Source: Author's own.

After selecting the dataset, the text processing phase is initiated. Special characters and punctuation marks, such as "?", "¿", "¡", "!", "/", ".", ";", ":", "*", "Ç", "$", and "&", are removed. Additionally, common Spanish empty words like prepositions and determiners, including "es" (is), "yo"

(I), "tú" (you), "de" (of), "en" (in), "para" (for), "por" (by), "y" (and), "o" (or), "esto" (this), "esos" (those), and "todo" (all), are eliminated.

On the other hand, the documents have been lemmatized and the words have been preserved in their canonical form. For example: vacunó (vaccinated), vacunará (will vaccinate), vacunaste (vaccinate) and vacunado (vaccinated), their lemma is vacunar (vaccinate). For stemming, word roots were identified, e.g.: vacunó (vaccinated), vacunará (will vaccinate), vacunaste (vaccinate), and vacunado (vaccinated) is vacun.

The LDA topic discovery algorithm was implemented to reduce the search space. After filtering, the number of documents was reduced by 60% and the top 10 subject words in the COVID-19 period are as follows: “coronavirus”, “SARS-CoV2”, “pandemia” (pandemic), “vacuna” (vaccine), “covid”, “paciente” (patient), “infección” (infection), “virus” y “tratamiento” (treatment). The word cloud for processing scientific texts in Spanish is shown in Figure 3.

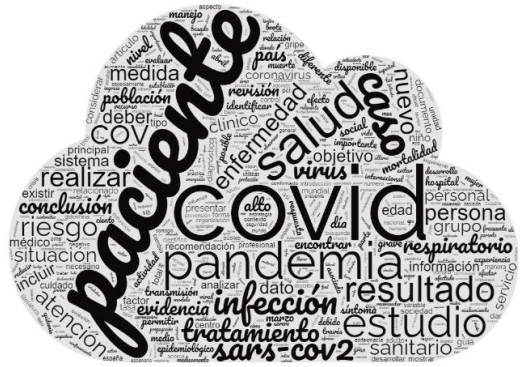

Figure 3 Scientific text word cloud
Source: Author’s own.

After text processing and the reduction of the dimensionality, numerical vectors were obtained for each text and query, using the tf-idf measure. Table 2 shows the vector values from the BioASQ dataset. The query used was “¿Qué quimioprofiláctico se puede utilizar contra el SARS-CoV2?" (Which chemoprophylactic can be used against SARS-CoV2?).

Table 2
TF-IDF values obtained.

	clinic	cloroquin	cobertur	cochran	cohort	coinfeccion	colaps
D1	0.10926	0	0.65232	0	0	0	0.10236
D2	0.0516	0.06148	0	0.083782	0	0	0.55257
D3	0.41314	0.05398	0	0	0	0.32162	0
D4	0	0	0.45987	0	0	0	0
D5	0	0	0	0	0.07104	0.23613	0
D6	0.16417	0.27772	0.37018	0	0	0	0.69844
D7	0.93866	0	0.04762	0.2365	0	0	0
D8	0	0	0	0	0	0	0
D9	0.055951	0	0	0	0.18365	0.05985	0.03961
D10	0.117591	0.26657	0.23652	0.3651	0	0	0
D11	0	0	0	0	0	0	0.04722
D12	0.042278	0	0	0	0.69852	0	0
D13	0.12577	0	0	0.59685	0	0.63685	0
D14	0.08989	0.12634	0.36589	0	0	0	0.03793
D15	0.110221	0.36454	0	0	0	0.10886	0

Source: Author's own.

The cosine angle similarity is then calculated. Figure 4 shows the similarity coefficients obtained with respect to the question.

```
[0.51798, 0.37671, 0.34602,  0.34303, 0.27589,
 0.26553, 0.24317, 0.24315,  0.23133, 0.22982,
 0.22890, 0.21926, 0.17992,  0.17507, 0.16103,
 0.15403, 0.15041, 0.13591,  0.13525, 0.13329]
```

Figure 4. 1.2.1 Cosine similarity coefficients.
Source: Author's own.

Moreover, Figure 5 presents the similarity coefficients using the BM-25 ranking function of the probabilistic approach.

```
[0.84636, 0.82860, 0.80777,  0.80308, 0.77788,
 0.77645, 0.77427, 0.75500,  0.75495, 0.75148,
 0.73185, 0.72687, 0.72449,  0.71406, 0.70642,
 0.69996, 0.69060, 0.68356,  0.67987, 0.67805]
```

Figure 5. Bm25 similarity coefficients.
Source: Author's own.

To calculate the similarity coefficients of the semantic approach, the frequency matrix of the texts and the question presented in Table 5 is applied, together with the implementation of the LSI method.

Table 3
Frequency matrix of terms.

Terms	D1	D2	D3	D4	D5	D6	D7	D8	D9
cloroquina	0	0	1	0	0	1	0	0	0
virus	1	0	1	1	1	0	1	1	0
vacuna	0	1	0	0	1	1	1	1	1
paciente	0	1	0	0	0	1	0	0	0
enfermedad	1	1	0	1	1	1	1	1	1
China	0	0	1	0	0	0	0	0	1
covid	1	0	1	1	0	1	1	1	0
contagio	0	1	0	0	0	0	0	1	1
Sars-Cov	1	0	1	0	1	1	1	0	0
Wuhan	0	0	1	0	0	0	0	0	1
moderna	0	1	0	0	1	1	1	0	0
muerte	1	1	1	0	1	0	0	0	1
tratamiento	1	1	1	1	1	1	1	0	0
fiebre	1	0	1	0	1	0	1	0	0
casos	0	0	1	0	0	0	0	0	0

Source: Author's own.

Figure 6 illustrates the similarity coefficients when querying the reduced dimensional SVD matrix of the LSI method.

```
[0.89885, 0.85362, 0.81604,  0.71477, 0.53148,
 0.51873, 0.49105, 0.42096,  0.29095, 0.35269,
 0.31985, 0.26645, 0.20226,  0.18050, 0.16392,
 0.11632, 0.11552, 0.10947,  0.01560, 0.00236]
```

Figure 6. LSI similarity coefficients.
Source: Author's own.

For the evaluation phase, a dataset consisting of a number of scientific texts of 249,474, a word count of 45,322,119 and a vocabulary of 373,851 was used. In addition, 100 questions in JSON format in Spanish were required in the pandemic period with the following tags: "id_Question", "id_Documento" and "texto_Pregunta". An example is shown in Figure 7.

```
{"PreguntasCovid": [
    {"id_Pregunta": 1,
    "TextoPregunta": "¿Qué quimioprofiláctico se puede utilizar contra el SARS-CoV-2? ",
    "id_Documento": "ibc-ET6-1764"},

    {"id_Pregunta": 2,
    "TextoPregunta": "¿Qué atención requieren los pacientes con FRA?",
    "id_Documento": "ibc-194912"},

    {"id_Pregunta": 3,
    "TextoPregunta": "¿Qué manifestaciones oculares se pueden presentar con el COVID-19?",
    "id_Documento": "ibc-192560"},

    {"id_Pregunta": 4,
    "TextoPregunta": "¿Existen estudios de glucocorticoides y COVID-19?",
    "id_Documento": "ibc-196134"}
]
```


Figure 7. Spanish questions.
Source: Author's own.

Evaluation and results

The results of applying the methodology to a case study of Spanish scientific documents generated during the COVID-19 pandemic are presented below. Table 4 shows the results of the textual similarity LDA experiment on the reduced dataset. The results are sorted in descending order by the value of the F-measure. The best ranking configuration for this approach was to apply processing with stemming and retrieve the top 20 documents. A precision of 87%, a recall of 69% and an F-score of 77% were achieved.

Table 4
Textual Similarity-LDA experiment Results.

Experimental setup	Top N	Precision	Recall	F-measure
processed-stemming	20	0.87	0.6957	0.7731
processed	20	0.83	0.7007	0.7599
processed-stemming	15	0.82	0.6952	0.7525
processed	15	0.83	0.6773	0.7459
processed-stemming	10	0.82	0.6722	0.7388
processed-lemmatization	20	0.83	0.6438	0.7252
processed-lemmatization	15	0.81	0.6337	0.7111
processed	10	0.76	0.6616	0.7074
processed-lemmatization	10	0.78	0.6194	0.6905
processed-stemming	5	0.75	0.6324	0.6862
processed	5	0.71	0.629	0.667
processed-lemmatization	5	0.72	0.5897	0.6484
processed-stemming	3	0.59	0.5264	0.5564
processed	3	0.57	0.5329	0.5508
processed-lemmatization	3	0.58	0.5039	0.5393

Source: Author's own.

The results of the application of the probabilistic LDA experiment of the proposed methodology are shown in Table 5. The best result was obtained with 89% precision, 82% recall and an F-measure of 85% using processing, stemming and retrieval of the top 20 documents.

Table 5
Probabilistic-LDA experiment Results.

Experimental setup	Top N	Precision	Recall	F-measure
processed-stemming	20	0.89	0.827	0.857
processed	20	0.87	0.814	0.841
processed-stemming	15	0.86	0.807	0.832
processed	15	0.85	0.803	0.826
processed-stemming	10	0.85	0.801	0.825
processed-lemmatization	20	0.85	0.792	0.82
processed	10	0.84	0.794	0.816
processed-lemmatization	15	0.83	0.78	0.804
processed-stemming	5	0.82	0.78	0.799
processed-lemmatization	10	0.81	0.766	0.787
processed-lemmatization	5	0.79	0.751	0.77
processed	5	0.78	0.751	0.765
processed-stemming	3	0.75	0.726	0.737
processed-lemmatization	3	0.73	0.703	0.716
processed	3	0.72	0.702	0.711

Source: Author's own.

Table 6 shows the results of the Semantic-LDA experiment. The best values were achieved with 88% precision, 78% recall and 83% F-measure, using processed-stemming and the top-20 documents.

Table 6
Semantic-LDA experiment results.

Experimental setup	Top N	Precision	Recall	F-measure
processed-stemming	20	0.88	0.786	0.83
processed-stemming	15	0.86	0.784	0.82
processed-lemmatization	20	0.85	0.744	0.793
processed-stemming	10	0.82	0.766	0.792
processed-lemmatization	15	0.81	0.741	0.774
processed	20	0.83	0.7	0.76
processed	10	0.77	0.727	0.748
processed	15	0.8	0.698	0.746
processed-stemming	5	0.74	0.713	0.726
processed	10	0.75	0.683	0.715
processed-lemmatization	5	0.71	0.686	0.698
processed-stemming	3	0.65	0.638	0.644
processed	5	0.63	0.606	0.618
processed-lemmatization	3	0.62	0.611	0.615
processed	3	0.56	0.548	0.554

Source: Author's own.

Finally, Table 7 shows the 3 best F-score values in descending order for each of the Top N. The probabilistic-similarity (OkapiBM25 algorithm) with the configuration processed-stemming is the best positioned with a value of 83%.

Table 7
Final results. The 3 better of every top are shown.

Approach	Method	Top N	Precision	Recall	F-measure
Probabilistic	processed-stemming	20	0.89	0.827	0.857
Probabilistic	processed	20	0.87	0.814	0.841
Semantic	processed-stemming	20	0.88	0.786	0.83
Probabilistic	processed-stemming	15	0.86	0.807	0.832
Probabilistic	processed	15	0.85	0.803	0.825
Semantic	processed-stemming	15	0.86	0.783	0.819
Probabilistic	processed-stemming	10	0.85	0.801	0.825
Probabilistic	processed	10	0.84	0.794	0.816
Semantic	processed-stemming	10	0.82	0.765	0.791
Probabilistic	processed-stemming	5	0.82	0.78	0.799
Probabilistic	processed-lemmatization	5	0.79	0.751	0.77
Probabilistic	processed	5	0.78	0.751	0.765
Probabilistic	processed-stemming	3	0.75	0.726	0.737
Probabilistic	processed-lemmatization	3	0.73	0.702	0.716
Probabilistic	processed	3	0.72	0.702	0.712

Source: Author's own.

Figure 8 shows the overall results, ordered by F-score. The lowest scores for each approach implemented in the methodology, were obtained when retrieving the 3 main documents. Textual similarity achieved an F-score of 55%, LSI 64% and probabilistic 73%. On the other hand, the highest value obtained in the retrieving for the top 20 documents was for the textual-similarity of 77%, the semantic of 83% and the probabilistic of 85%.

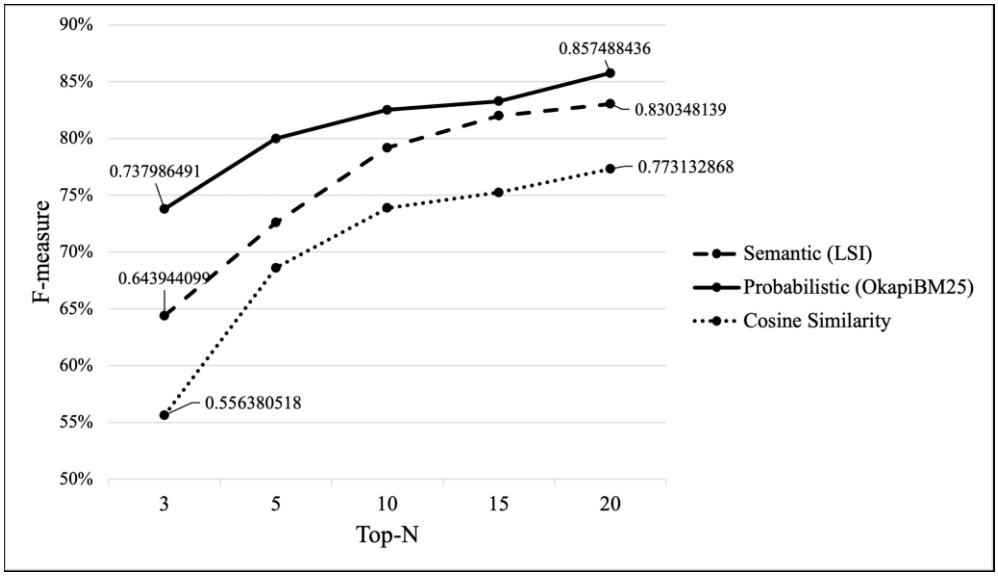

Figure 8. Overall results of retrieving relevant documents from the scientific literature using similarity approaches based on PLN techniques.
Source: Author's own.

Discussion

This section presents a comprehensive analysis of the results of the paper, aimed at understanding its relevance in the discipline of Information Retrieval. The proposed methodology was tested using 249,474 scientific documents in Spanish, and different experimental combinations were used, including three PLN-based IR approaches and a topic discovery algorithm.

The results show that the configuration, which includes the implementation of the LDA algorithm, PLN techniques such as text processing, word root extraction using stemming, and combination with the probabilistic BM25 approach, achieved a precision of 89%, a recall of 82% and an F-measure of 85%. These values are higher than those reported by Cao et al. (2011) and Asnath Tinega et al. (2018), who also used the probabilistic Okapi BM25 ranking function but combined it with lexico-syntactic patterns, inverted Lucene index, or machine learning methods. Although scientific papers from the same Pubmed data source were used, this work demonstrated a significant improvement by applying keyword filtering through the topic discovery algorithm and reducing the search space using text processing. Consequently, our proposed methodology outperforms a heuristic approach used in Wang et al. (2017) and compares favorably with Bayesian networks and language models proposed by Otegi et al. (2022).

Finally, the methodology has yielded promising results in retrieving scientific texts in Spanish, as compared to the research conducted in English by Khadhraoui et al. (2022), which employs a deep learning-based text classifier and requires more computational resources.

The contributions of this paper are novel in that the interaction between NLP techniques, textual similarity, probabilistic and semantic approaches, and the topic discovery algorithm increases the precision of document retrieval from large digital repositories and facilitates the correct management of information. Throughout the IR process, the practical and theoretical implications of the results are highlighted, as well as the limitations of the study that need to be considered, such as the processing time of a large number of digital documents and the implications of using more computing resources.

The methodology described can be replicated by following the phases outlined. Although the use case tested documents from the medical domain on COVID-19, the methodology can be implemented for documents from other domains, such as administrative, accounting, and economic. Documents can be extracted from UNAM institutional repositories, specifically from the area of Accounting and Administration from 2010 to 2024, and the importance of the types of observations in the results of audits carried out by bodies such as the superior audits of the states in Mexico can be analyzed using the number of complaints made by the supreme audit institutions or the amounts recovered for damages to employers determined by the higher auditing organizations in Mexico. These topics can also be extracted from the electronic portals of the Chief Audit Office of Mexico (Auditoría Superior de la Federación in spanish) and the digital libraries of various academic institutions. The process of document retrieval would be the same as described in the methodological section of this article.

The methodology presented in this article can be applied to any field of knowledge due to the increasing digitisation of information and the need to access relevant information.

Conclusions

This paper has presented a methodology for managing scientific literature in Spanish employing Information Retrieval approaches and Natural Language Processing. The methodology based on IR approaches retrieves relevant documents regarding an input question in natural language. Several domains, such as financial auditing, public administration, or government legislation can be encompassed by the proposed methodology. An evaluation process using several experiments combining the LDA with three proposed approaches for document retrieval: cosine similarity, probabilistic, and semantics was carried out.

The main contribution of this work is the creation of a methodology that involves the analysis, implementation, and combination of the three IR approaches with variants of text processing using cosine similarity, probabilistic document search and context-based semantics.

Experimental setups for each approach using NLP techniques and the LDA algorithm; an evaluation was carried out to compare the results and obtain the best setup. The creation and management of a dataset of questions in Spanish and the acquisition of the scientific text dataset were carried out.

The LDA algorithm was used to manage and reduce the number of documents to be analysed, omitting documents unrelated to the search topic. An application of the proposed methodology was performed, and it notes that the best result is achieved using the probabilistic-similarity (Okapi BM25 algorithm) with LDA and NLP techniques, such as processing and stemming texts, achieving an F-score of 85% as the highest result. The OkapiBM25 algorithm based on probabilistic can be used to retrieve Scientific documents that are relevant for an input question over large amounts of text with efficient results.

Furthermore, it is essential to mention that, according to the experiments carried out in this paper, the stemming technique outperformed lemmatization in the management and retrieval of documents from the scientific literature. We rely on that it is due to the nature of the stemming task, which involves the root of the word instead of the canonical form, as in the lemmatization task.

The corpus used in the application of the methodology of this article can be variable. Therefore, the proposed approaches to document retrieval can obtain relevant information in any domain from a new question without changing the methodology phases.

The approaches proposed in this paper have an advantage for a question-answering system since ranking the relevant documents. It reduces the time involved in the answers search in all collections in Spanish.

As future work, it is proposed to use deep learning algorithms to retrieve relevant documents from sets of Spanish texts. Also, it is recommended to try the enrichment of the space vector model with linguistic and WordNet-based features. Finally, a semantic annotation using MedLexSP will be applied to process the documents and questions to improve the results.

References

Cairns, B. L., Nielsen, R. D., Masanz, J. J., Martin, J. H., Palmer, M. S., Ward, W. H., & Savova, G. K. (2011). The MiPACQ clinical question answering system. AMIA Annual Symposium Proceedings Archive, 171-180. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC3243235/. Consultado: 05/11/2023

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., & Yu, H. (2011, April). AskHERMES: An online question answering system for complex clinical questions. Journal of biomedical informatics, XLIV, 277-288. https://doi.org/10.1016/j.jbi.2011.01.004

Lee, L. (2007). IDF revisited: a simple new derivation within the Robertson-Spärck Jones probabilistic model. Proceedings of the 30th 412 annual international ACM SIGIR conference on Research and development in information retrieval, 751-752. https://doi.org/10.1145/1277741.1277891

Cyril, C., & Michael, K. (1966). Factors Determining the Performance of Indexing Systems (1° ed.). ASLIB.

A. Taan, A., Rehman Khan, S. U., Raza, A., Hanif, A. M., & Anwar, H. (2021, November). Comparative Analysis of Information Retrieval Models on Quran Dataset in Cross-Language Information Retrieval Systems. IEEE Access, 9, 169056-169067. https://doi.org/10.1109/access.2021.3126168

Asnath Tinega, G., Mwangi, W., & Rimiru, R. (2018). Text Mining in Digital Libraries using OKAPI BM25 Model. International Journal of Computer Applications Technology and Research, 7(10), 398-406. https://doi.org/10.7753/ijcatr0710.1003

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3, 993-1022. Disponible en: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf. Consultado: 05/11/2023

Badenes Olmedo, C., Lozano Alvarez, B., & Corcho, O. (2021). Impact of Text Length for Information Retrieval Tasks based on Probabilistic Topics. Procesamiento del Lenguaje Natural, 67, 27-36. https://doi.org/10.26342/2021-67-2

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (1° ed.). New York: ACM press.

Berger, A., & Lafferty, J. (2017, July). Information retrieval as statistical translation. ACM SIGIR Forum, 51(2), 219-226. https://doi.org/10.1145/312624.312681

Burak Ozyurt, I., Bandrowski, A., & Jeffrey, S. (2020, January 10). Bio-AnswerFinder: a system to find answers to questions from biomedical texts. Database, MMXX, 1-12. https://doi.org/10.1093/database/baz137

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990, September). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., & Socher, R. (2021, April). COVID-19 information retrieval with deep-learning based semantic search, question answering, and

abstractive summarization. NPJ digital medicine, 4(1), 1-9. https://doi.org/10.1038/s41746-021-00437-0

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1977). Computer Methods for Mathematical Computations (1° ed.). Englewood Cliffs, NJ, USA: Prentice Hall.

Gong, Y., Cosma, G., & Fang, H. (2021, July 26). On the Limitations of Visual-Semantic Embedding Networks for Image-to-Text Information Retrieval. Journal Imaging, 7(8), 1-15. https://doi.org/10.3390/jimaging7080125

Gordon, R. G. (2005). Ethnologue: Languages of the world (15rd ed.). SIL International.

Jinhyuk, L., Sean, S. Y., Minbyul, J., Mujeen, S., WonJin, Y., Yonghwa, C., Miyoung, K., & Jaewoo, K. (2020, December). Answering Questions on COVID-19 in Real-Time. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, 1-10. https://doi.org/10.18653/v1/2020.nlpcovid19-2.1

Jipeng, Q., Zhenyu, Q., Yun, L., Yunhao, Y., & Xindong, W. (2022, March). Short text topic modeling techniques, applications, and performance: a survey. IEEE Transactions on Knowledge and Data Engineering, 34, 1427-1445. https://doi.org/10.1109/tkde.2020.2992485

Khadhraoui, M., Bellaaj, H., Ben Ammar, M., Hamam, H., & Jmaiel, M. (2022, March 11). Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study. Applied Sciences, 12(6), 1-19. https://doi.org/10.3390/app12062891

Korfhage, R. R. (1997). Information Storage and Retrieval (1° ed.). New York: Wiley Computer Publishing.

Mahmood, M., AL-kubaisy, W. J., & Al-Khateeb, B. (2019, June). Using artificial neural network for multimedia information retrieval. Journal of Southwest Jiaotong University, 54(3), 1-9. https://doi.org/10.35741/issn.0258-2724.54.3.19

Otegi, A., San Vicente, I., Saralegi, X., Peñas, A., Lozano, B., & Agirre, E. (2022, March 15). Information retrieval and question answering: A case study on COVID-19 scientific literature. Knowledge-based systems, 240, 1-12. https://doi.org/10.1016/j.knosys.2021.108072

Raza, S. (2022, November). A COVID-19 Search Engine (CO-SE) with Transformer-based architecture. Healthcare Analytics, 2, 1-14. https://doi.org/10.1016/j.health.2022.100068

Rodriguez, D. V., & Carver, D. L. (2019). Comparison of information retrieval techniques for traceability link recovery. IEEE 2nd International Conference on Information and Computer Technologies (ICICT), 186-193. https://doi.org/10.1109/infoct.2019.8710919

Ross, S. M. (2014). Introduction to probability models (11rd ed.). CA, San Diego, USA: Academic Press.

Salton, G., Wong, A., & Yang, C. S. (1975, November). A vector space model for information retrieval. Communications of the ACM, 18(11), 613-620.

Sarrouti, M., & El Alaoui, S. O. (2020, January). SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. Artificial Intelligence in Medicine, CII, 101767. https://doi.org/10.1016/j.artmed.2019.101767

Su, D., Xu, Y., Yu, T., Siddique, F. B., Barezi, E. J., & Fung, P. (2020). CAiRE-COVID: a question answering and multidocument summarization 375 system for COVID-19 research. EMNLP2020 NLP-COVID Workshop, 1-8. https://doi.org/10.18653/v1/2020.nlpcovid19-2.14

Trotman, A., Puurula, A., & Burgess, B. (2014, November 26). Improvements to BM25 and language models examined. ADCS '14: Proceedings of the 2014 Australasian Document Computing Symposium, 58-65. https://doi.org/10.1145/2682862.2682863

van Rijsbergen, C. J. (1998). A Non-Classical Logic for Information Retrieval. In F. Crestani, M. Lalmas, & C. J. van Rijsbergen, Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information (pp. 3-13). Boston, MA, USA: Springer US. https://doi.org/10.1007/978-1-4615-5617-6_1

Wang, H., Zhang, Q., & Yuan, J. (2017). Semantically Enhanced Medical Information Retrieval System: A Tensor Factorization Based Approach. IEEE Access, 5, 7584-7593. https://doi.org/10.1109/access.2017.2698142

Wong, S. K., Ziarko, W., & Wong, P. C. (1985, June). Generalized vector spaces model in information retrieval. SIGIR '85: Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, 18-25. https://doi.org/10.1145/253495.253506

Xu, B., Lin, H., Lin, Y., Ma, Y., Yang, L., Wang, J., & Yang, Z. (2018). Improve Biomedical Information Retrieval Using Modified Learning to Rank Methods. IEEE/ACM Trans Comput Biol Bioinform, 15(6), 1797-1809. https://doi.org/10.1109/tcbb.2016.2578337

Zhang, E., Gupta, N., Tang, R., Han, X., Pradeep, R., Lu, K., Zhang, Y., Nogueira, R., Cho, K., Fang, H., & Lin, J. (2020, November). Covidex: Neural ranking models and keyword search 385 infrastructure for the covid-19 open research dataset. Proceedings of the First Workshop on Scholarly Document Processing, 31-41. https://doi.org/10.18653/v1/2020.sdp-1.5