Document Search Engine using Text Analysis hosted over the cloud

MSc Research Project
MSc Cloud Computing

Vishwas Mudalahippe Shankarappa
Student ID: 21205825

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vishwas Mudalahippe Shankarappa
Student ID:	21205825
Programme:	Cloud Computing
Year:	2023
Module:	MSc Research Project
Supervisor:	Rejwanul Haque
Submission Due Date:	14/08/2023
Project Title:	Document Search Engine using Text Analysis hosted over the cloud.
Word Count:	7331
Page Count:	20

I hereby certify that the information contained in this (my submission) is informationpertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at therear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placedinto the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Document Search Engine using Text Analysis hosted over the cloud

Abstract

A document search engine is a programme that allows users to access a document database and perform searches inside that database. Modern algorithms are used to create an index of a database's content, making it easier for users to find the specific data they need. Businesses often use document search engines to help employees more quickly and accurately locate the information they need among the plethora of company manuals, reports, and other written documents. A business needs both local and remote access to its data, regardless of whether it is stored locally or on the cloud. Document search engines, which allow for the reliable and efficient filing and retrieval of massive volumes of data, are, thus, an indispensable resource for today's businesses.

Keywords: Document Semantic Search, Cloud Management, Text Analysis, Business Information Retrieval Systems, Electronic Library

1. Introduction

The importance of efficiently obtaining information that is both accurate and relevant has grown in today's modern society, which is marked by a profusion of information. The discipline of computer science known as "information retrieval" (or "IR") focuses on the study, development, and use of methods and systems that efficiently locate and provide user-requested information. Just a few examples of such tools include internet search engines, document management systems, and digital libraries. The proliferation of digital technology has led to a meteoric rise in the amount of information available in a broad range of file types. One of the most significant variables determining the effectiveness of document retrieval systems is the process of transforming paper documents into digital format. There is a wide variety of content that has to be digitised, including books and paper documents. Additional types of materials abound. It is standard routine to employ optical character recognition (OCR) technology when scanning or collecting information from documents with the intention of transforming its content. Once a batch of papers has been digitised, the documents in that batch may be stored in an electronic archive and made searchable using a variety of keywords and queries. In the realm of information retrieval, ranking documents is a crucial activity since it is utilised in so many kinds of Internet-based software and services.

1.1 Motivation

The sheer volume of manuals, reports, and other documents that businesses create can make it challenging to manually manage and search through. Using document management software is one approach to fixing this issue. Employees might potentially save time and effort by using a document search engine that can automatically index and search through these documents. This might lead to greater productivity by making information discovery faster and more precise. Accurate and efficient information retrieval is growing in significance for a variety of reasons, including the need to make educated judgements, the pursuit of knowledge acquisition, the search for a competitive advantage, and the need for personal convenience. The necessity to deal with too much information, the difficulty of dealing with too much information, and the difficulty of coping with too much information are all additional factors. As the digital world expands at a dizzying rate, efficient information retrieval systems are becoming more and more important in helping individuals and businesses get the specific data they need when they need it. One of the main sources of innovation in many different sectors is the constant progress of information retrieval. The efficiency, effectiveness, and convenience of retrieval systems are all areas that this development hopes to enhance.

1.2 Business Problem

As the volume of data generated by businesses grows, so does the complexity of managing its storage, retrieval, and application. Employees without access to a document search engine may waste considerable time doing inefficient, potentially error-prone manual searches for the information they need. This might lead to lower output, higher costs, and more difficulty in adhering to regulations. Finding documents with the use of large-scale language models like BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pre-trained Transformer). As a consequence of the LLMS's capacity to enhance the precision and relevance of data retrieval, it has made great strides in the realm of document research. LLMS models may easily acquire a comprehensive knowledge of language since they can consume information from many textual sources during the pre-training phase. Training participants acquire the ability to distinguish statistical patterns, word pairs, and contextual hints within the dataset utilised for training. They store a massive amount of linguistic information that might be useful in applications like document retrieval. Improved search engine rankings are a side effect of LLMS models' deep familiarity with a language's semantics. To find relevant texts that do not have an exact keyword match, LLMS models need to grasp the context in which words and phrases are employed. Users can find useful results that may have been missed using more traditional keyword-based search methods. When users take these steps, they improve their chances of discovering such resources.

1.3 Research Questions

The following research questions need to be addressed based on the business problem mentioned above:

RQ1: How effective is the Cloud-Based Document Search Engine using Text-Analysis in identifying relevant documents compared to traditional document search methods?

RQ2: How does the search mechanism perform based on different algorithms?

RQ3: How to make the search query faster?

RQ4: What is the impact of different cloud architecture systems on the similarity search based on the execution time and space storage?

2. Related Work

2.1 Document Extraction related research

In (Balaji, P., Merker, H., & Gupta, 2023), a comprehensive pipeline for zero-shot text categorization is described. This pipeline may categorise text based on a user-defined collection of features without the requirement for specialist datasets. The fairness of the evaluation is checked by an evaluator, and the pipeline also includes a document parser and a text classification model called EntailClass. Sentence order is maintained, and the sentences are linked to the chapter headings in this way. The pipeline is completely automated, so it can be easily integrated into any existing processes. Results showed that the pipeline outperformed state-of-the-art algorithms on a test set of supervisory documents, with an accuracy of 87.2%.

Blueprint is a declarative domain-specific language that use fuzzy constraints to build document structure and extract field-value mappings (Mishchenko, A., Danco, D., Jindal, A., & Blue, A., 2022). Researchers Mishchenko, D., D. Danco, A. Jindal, and A. Blue created Blueprint. However, many users found it difficult to write extraction algorithms despite its usefulness in corporate settings. Studio, an IDE created specifically to address this issue, is now available. Studio paves the way for data-driven, code-free programme synthesis in addition to traditional coding languages. Comparable accuracy to deep learning-based methods was achieved, but with far better interpretability and debuggability. It was also shown that the same amount of labour could be accomplished, but with the added bonus of a reduced time to create the system.

An strategy for keyword extraction is described using a semantic hierarchical network model (Zhang, T., Lee, B., Zhu, Q., Han, X., and Chen, 2023). Both the external context and the internal structure of the keywords are considered by this approach. The approach begins with a hierarchical extraction of feature words to build a semantic graph, which is then applied to the text to uncover the document's keyword collection. The results show that the method excels above state-of-the-art alternatives in terms of precision, recall, and F-measure, and that it accurately displays the hierarchical relationships between words inside the semantic network.

Current technical developments in text summary are discussed in this paper (Arya, C., Diwakar, M., Singh, P., Singh V., Kadry, S., and Kim, 2023) with a focus on news summarising. Correctly summarising several news items calls for a synthesis technique, since it is necessary to extract, compare, and score terms. The presented method employs models and sentence ratings to identify and prioritise the most pertinent sentences from the original articles, resulting in more accurate and comprehensive summaries than can be obtained using conventional approaches. Five English-language news websites covering the same topic or event were analysed for this study.

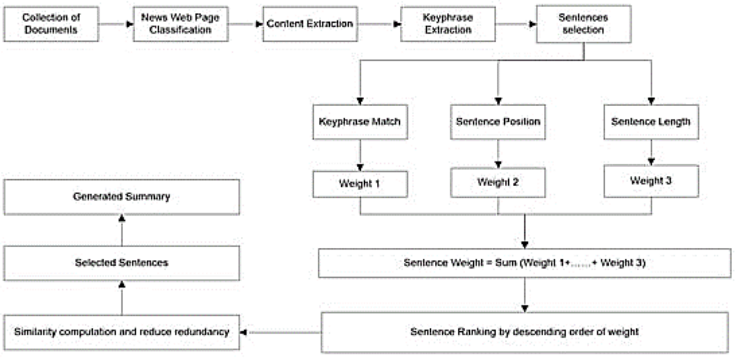


Figure 1: New Summarization using the keyword and document extraction (**Arya, C., et.al., 2023**)

Meuschke, N. et al., 2023 argue that information extraction from academic PDF publications is crucial for a range of use cases; yet, choosing the best effective solution may be challenging due to a lack of published performance assessments. This study provides a thorough evaluation technique, evaluating 10 freely available tools for extracting data from documents, citations, and tables. The research relies on DocBank, a dataset that covers a wide range of topics. While GROBID, CERMINE, and Scientific Parse all do a fine job at metadata and reference extraction, Adobe Extract is far and away the best option when it comes to table extraction. However, enhancements and combinations of tools are necessary to achieve good extraction quality for the bulk of the content sections. Both the data and the code may be accessed by researchers in this field through open channels.

In order to increase the output of transformer models, Banerjee, S. et al., 2023 propose a technique for producing disaster-related news headlines. Using a DNN-based phrase extractor to pull out key sentences and a transformer-based abstractive summarizer to put it all together, this method uses the "extract then abstract" methodology. An extract-then-abstract method describes this strategy. The DNN is trained to generate two binary labels, one for whether or not there is any information related to the aftermath of the disaster, and another for whether or not the headlines closely reflect the ground reality. It has been shown that employing the proposed method results in higher ROUGE scores for the headlines than when using the raw data from the original media. The research shows that this method may be useful in creating newsworthy headlines by capitalising on disaster-related content.

Guo, J., Kok, S., and Bing, L., 2023 outline a procedure called "document-level relation extraction" (DocRE), which includes guessing the nature of relationships between entity pairs in a text. This can be challenging since it entails picking out the few good connections among the many unfavourable ones. Working with datasets that have several annotation errors makes this problem far more challenging to resolve. To improve the discriminability and robustness of DocRE models, the authors of this study devised a loss function and a novel negative label sampling mechanism. In addition, in order to evaluate the efficacy of their sampling method, they introduced two additional data regimes. The results showed that their method outperformed competing techniques on a wide range of datasets, including those with imperfect annotated labels.

Sun, Q., et al. 2023 discusses document-level Relation Extraction (RE), which aims to identify semantic linkages between entities in a document. The current methods have issues due to the prevalence of negative occurrences relative to good ones, as well as the need for static graphs governed by heuristic rules. In this research, we make advantage of dynamic graph attention networks to provide a novel two-stage framework we call TDGAT. The first stage uses binary classification to identify entity pair associations, while the second stage disentangles finer ties and builds on prior knowledge to produce graphs for enhanced prediction. The dynamic graph approach analyses the interplay between several bits of linked data. Experimental results on DocRED have shown that TDGAT is superior to other models in its ability to identify connections between words.

While much work has gone into defining and evaluating methods for parsing English phrases into Abstraction Meaning Representation (AMR) graphs, the task of parsing whole texts into a unified graph representation has received comparatively less attention. Here, we provide a simple method for generating a unified graph representation by employing a super-sentential level of AMR coreference annotation developed in earlier work. Both the loss of information and the incoherence that might follow from merging too much can be avoided using this technique. Then, we'll take a look at the state of the art in document-level AMR parsers and explain the adjustments made to the Smatch measure to make it suitable for comparing graphs across documents. In this article, we will go through how to incorporate the best available AMR parser and coreference resolution tools into a pipeline. This technique provides a solid groundwork for the follow-up study.

The DOCAMR standard modifies the AMR file format so that it may be used in the world of documents. The MULTI-AMR entity and event coreference annotations have been adequately merged into a single graph in order to reflect the intended meaning of this document. All semantic information is kept safe, and the system does its best to follow AMR guidelines even at the phrase level.

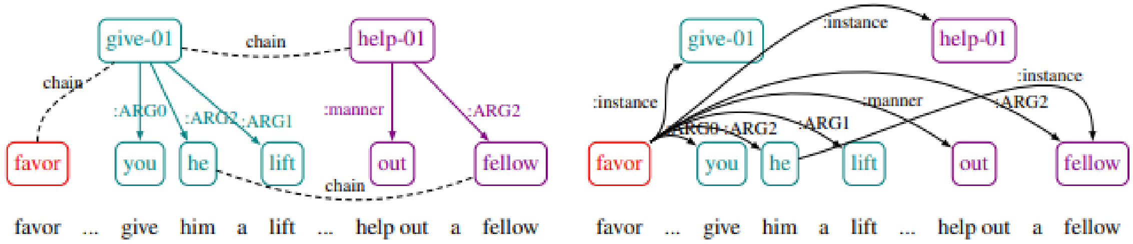


Figure 2: an example of a merging technique that results in either data loss or data distortion. At the sentence level, nodes with the same hue have the same AMR. Dashed lines represent the links in the identification chain. element not serving as a predicate. As a result of the kindness of the other coreference nodes, Favour has been able to acquire several ARG nodes. There is now no connection between the give-01 and help-01 predicate nodes that were merged in and the parameters that were assigned to them.

When an identity chain is generated, a new coref-entity node is appended to the graph. The coref entity connection links this new node to the rest of the chain. Figure 2 clearly demonstrates this. We offer two major revisions to the nodes that make up the coref-entity. The first step is to construct a single entity from the identified entities. Any pronominal nodes that are part of a chain should be deleted next. The method in which coreferent named objects and pronouns are handled in AMR sentence annotations accounts for both of these cases.

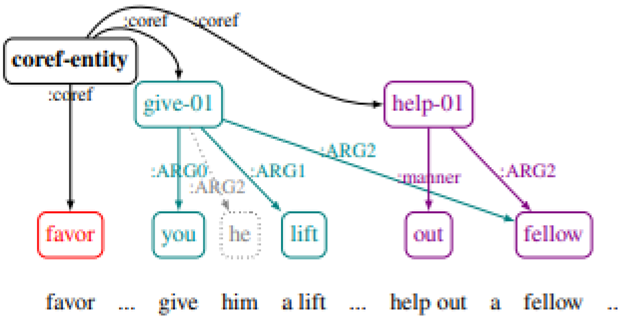


Figure 3: DOCAMR implementation on Figure 3 example.

DOCAMR was implemented, as shown in Figure 4, using the data in Figure 3. By relocating the linkages to the non-pronominal node, the identification chain between "he" and "friend" gets rid of the pronoun (shown in dotted grey) entirely. Since it may be disastrous to combine the identity chains of 'favour,' 'give-01,' and 'help-01,' a 'coref-entity' node has been added, and all other nodes are linked to it by the ':coref' edge. They ensure that the intended meaning is not lost while preventing any node in the chain from receiving preferential treatment.

DOCAMR is the format for document-level AMR graphs, which are coreference-annotated graphs. The DOCAMR rollout is still in the planning stages. DOCAMR can effectively remove superfluous data from graphs representing phrase representations while retaining all necessary data. The implementation of Smatch was enhanced to make use of sentence provenance when comparing two graphs at the document level in order to find node alignments. Results for a document-level parsing pipeline were provided at the end of the study, and they may provide a solid foundation for future studies in this field.

2.2 Artificial Intelligence used in query search

Ding, M.A., and Goldfarb, A., 2023 examine the quantitative marketing literature on AI from a macroeconomic perspective. Using the five levels of prediction, decision, tool, strategy, and society, the authors classify 96 academic research papers on AI in marketing. They look at every facet of a task, from the research subject to the artificial intelligence model used to the overall decision type and each individual research article. The authors draw attention to the dearth of marketing studies that address strategic and societal issues, and provide suggestions for future research. They find that there is a dearth of marketing publications that address both strategy and society.

Mahmood, M., Al-Kubaisy, W.J., & Al-Khateeb, B., 2023 all agree that the semantic gap between human perception and machine language makes multimedia information retrieval (MIR) a challenging task. The focus of this research is on improving a model for the retrieval of multimedia material that has already been created. Some examples of these methods are the Discrete Wavelet Transform (DWT), the Vector Space Model (VSM), the Latent Semantic Index (LSI), and Neural Networks. With the help of training and evaluation, we can create a multimedia retrieval system that can handle a wide range of multi-modal datasets.

The challenges faced by conventional search engines are investigated in this study (SAHITHYA, A., and KUMAR, S.V., 2023), along with a machine learning-based solution proposed to improve search results relevance. The proposed search engine plans to employ machine learning strategies to improve the accuracy of the results.

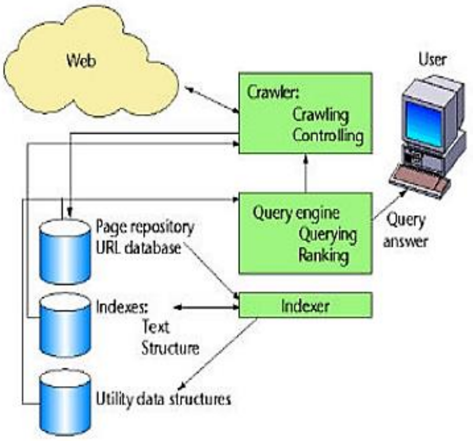


Figure 4: An end-to-end document search and document retrieval (**SAHITHYA, A. and KUMAR, S.V., 2023**)

2.3 Cloud Platforms used for hosting document search software

This article (Adrakatti, A.F., & Mulla, K.R., 2023) examines the proliferation of online courses and lectures in higher education. Video-sharing sites host the vast bulk of these recordings, however students may find it difficult to navigate these sites due to the lack of robust metadata search capabilities. While many academic institutions and libraries do record and archive lectures for later use, the great majority of them are not accessible to the public. When recorded lectures are indexed with an open-source application designed for full-text search inside video footage, students may more easily find and obtain the information they need. The writers created the programme to address this issue.

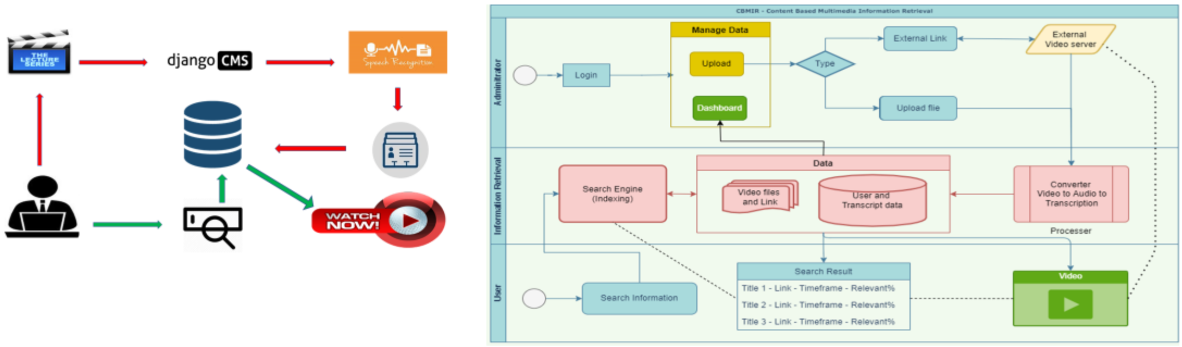


Figure 5: The Django frame along with the AWS cloud framework for the video-based information retrieval (**Adrakatti, A.F. and Mulla, K.R., 2023**)

Priya V.D. and Chakkaravarthy S.S., 2023 developed a programme called CBMIR (Content-Based Multimedia Information Retrieval) to address this issue. By measuring how long the speech is in the video and obtaining data based on it, CBMIR takes a practical solution to the problem. The application incorporates voice recognition technology and automated text indexing to facilitate the search process. Videos of lectures presented in English by Indian universities, research centres, and non-profits are the only content allowed on the platform.

By imitating legitimate systems and applications on the network, honey traps help identify fraudulent packets and confuse attackers (Sharma, Y., Bhatt, C., Batra, A., and Chauhan, 2023). To do this, a simulation of the live network's conditions must be set up. Some have suggested using containerized honeypots as a practical and long-lasting solution to this issue. Research honeypot detectors have also been used to study malicious hacking attempts. Insights about the actions and motivations of potentially dangerous users may be gleaned from the data provided by the monitored process.

2.4 Summary of the research done

Prediction, decision, tool, strategy, and social repercussions are the five broad areas where AI has an effect on marketing. Technology like vector space models, linear support vector machines, conditional random fields, spherical harmonics, and neural networks are all contributing to the development of multimedia information retrieval (MIR). Search engines powered by machine learning are among the multimedia formats that may be retrieved by MIR. Instructing AI with FAIR data is helpful. Search engine and recommendation system retrieval systems both benefit from query rewriting. Due to data constraints, zero-shot text categorization was developed. Data extraction from documents is facilitated by Blueprint and Studio. Keyword extraction and news summarising using a semantic hierarchical network have shown encouraging results. Honeypots and the blockchain both improve the safety of data storage. Integrating AI into healthcare improves patient data management. This all-encompassing analysis emphasises the many applications of AI and the revolutionary changes it can bring about in several fields.

2.5 Research Contribution

Diverse approaches can be used to improve document search in the future. Transfer learning allows zero-shot learning to compensate for insufficient data. Keyword extraction using a semantic hierarchical graph has the potential to be more precise. Techniques such as vector space models (VSM), latent semantic indexing (LSI), and neural networks are used in multimedia information retrieval (MIR). Search engine optimisation, multilingual IR, and multimedia asset retrieval can all benefit from query reformulation. Honey traps help with packet detection, while blockchain improves the administration of medical records. Containers and microservices are used in cloud-native architectures to facilitate scaling. The quality of searches may be improved with the use of cloud-based ML services like NLP and picture recognition. The efficiency of document-analysis teams can be increased with the use of cloud-based storage and collaboration tools. These novel approaches have the potential to improve the reliability and productivity of document retrieval systems.

3. Research Methodology & Design Specification

3.1 Parsing the pdfs

In order to make the content in PDF files searchable and usable, a procedure known as "PDF parsing" is performed. If search engines are to properly index PDFs and return useful search results, they must possess this capacity. Search engines function better when they are able to extract information such as content, titles, headers, and more. Inadequate parsing might lead to disappointing search results for end users. Beyond SEO, PDF parsing has real-world applications in fields including business, data mining, and analytics. Organisations may get valuable intelligence from PDFs, which can subsequently be utilised to inform strategy and decision-making. Financial data, customer responses, and market tendencies all provide sources from which these insights might be drawn. Content management systems utilise PDF processing to classify and organise documents for easier retrieval. It also plays a significant role in maintaining legal and regulatory compliance by efficiently assessing lengthy contracts and legal documents. The flexibility of PDF parsing makes it important for data-driven applications and information management across many industries. In today's business world, this is a boon to both efficiency and sound judgement.

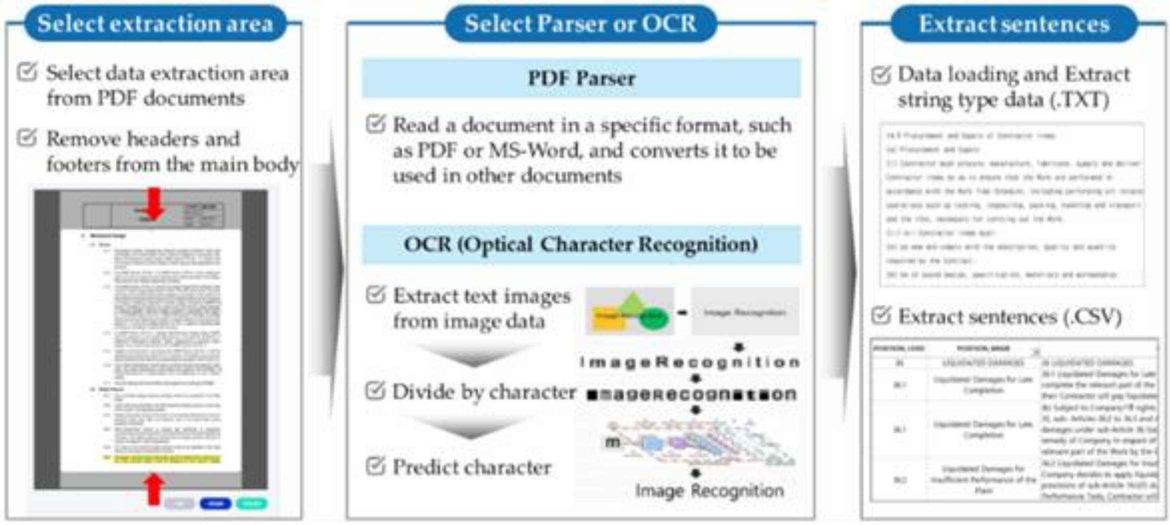
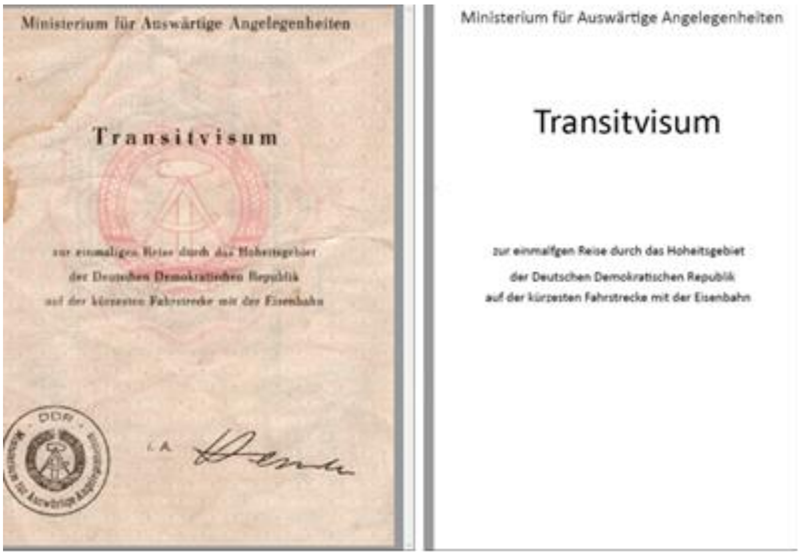

Select extraction area
☑ Select data extraction area from PDF documents
☑ Remove headers and footers from the main body
Select Parser or OCR
PDF Parser
☑ Read a document in a specific format, such as PDF or MS-Word, and converts it to be used in other documents
OCR (Optical Character Recognition)
☑ Extract text images from image data
☑ Divide by character
☑ Predict character
ImageRecognition
Image Recognition
Extract sentences
☑ Data loading and Extract string type data (.TXT)
☑ Extract sentences (.CSV)

Figure 6: Example of Parsing for extracting the information from a pdf (Choi, S.W. and Lee, E.B., 2022)

3.1.1 Parsing Scanned pdfs using OCRs

OCR, which stands for optical character recognition, has found a home in many fields beyond its original use. Businesses use OCR to streamline data analysis, computerise routine tasks, and improve the quality of their decisions. It expedites processes like invoicing, helps those who are visually impaired using tools like screen readers and Braille displays, and aids with regulatory compliance by digitising papers.



Ministerium für Auswärtige Angelegenheiten

Transitvisum

zur einmaligen Reise durch das Hoheitsgebiet
der Deutschen Demokratischen Republik
auf der kürzesten Fahrstrecke mit der Eisenbahn

DDR Ministerium für Auswärtige Angelegenheiten

i.A.

Ministerium für Auswärtige Angelegenheiten

Transitvisum

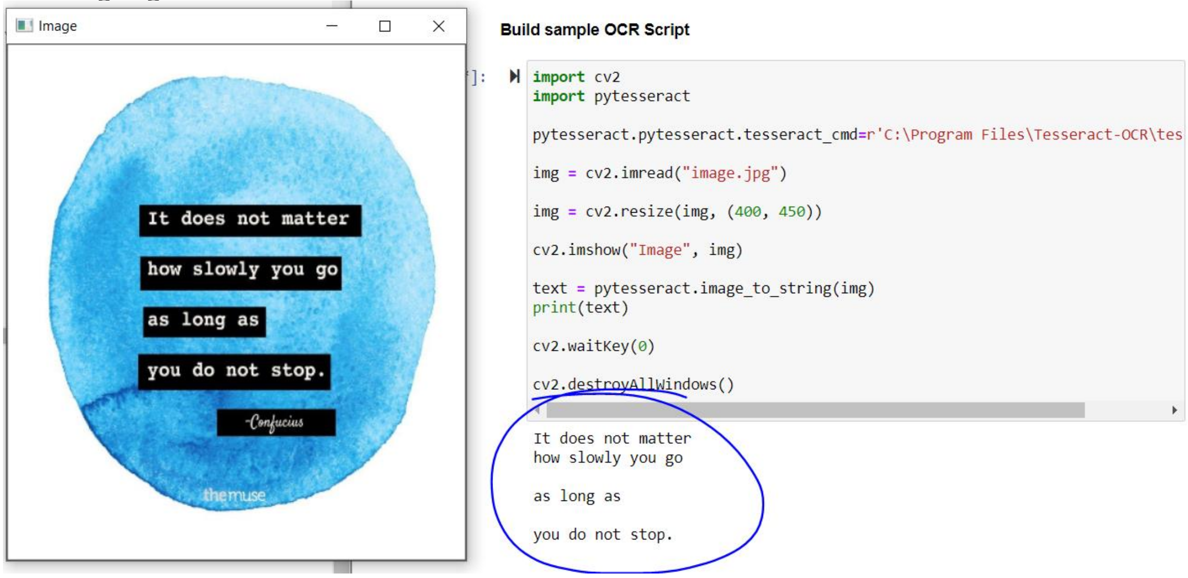
zur einmalfgen Reise durch das Hoheitsgebiet
der Deutschen Demokratischen Republik
auf der kürzesten Fahrstrecke mit der Eisenbahn

Figure 7: Example of original text in pdfs and the extracted text using OCR

In addition, OCR strives to preserve cultural artefacts and foster cross-cultural understanding.Current studies aim to refine language recognition, enhance accuracy via the application of deep learning, and provide solutions to layout-related problems. Possible future directions include investigating how well augmented reality and navigation technologies work together. The ultimate goal is to increase the accuracy, speed, and diversity of optical character recognition (OCR) in order to enhance data management, accessibility, and information processing.

3.1.2 Tesseract

Tesseract's gradual evolution into a robust OCR system emphasizes the complexity of the task. Techniques like Hidden Markov Models and Deep Belief Networks hold potential for improvement. Tesseract's history demonstrates evolutionary and revolutionary synergy in OCR advancement. It finds applications in various sectors like finance for invoice and receipt processing, banking for automated cheque handling, and healthcare for digitizing medical records. Tesseract OCR's reliability and effectiveness make it valuable for diverse industries, enabling digitization and efficient information extraction.



It does not matter
how slowly you go
as long as
you do not stop.

Build sample OCR Script

```
import cv2
import pytesseract

pytesseract.pytesseract.tesseract_cmd=r'C:\Program Files\Tesseract-OCR\tes

img = cv2.imread("image.jpg")

img = cv2.resize(img, (400, 450))

cv2.imshow("Image", img)

text = pytesseract.image_to_string(img)
print(text)

cv2.waitKey(0)

cv2.destroyAllWindows()
```

```
It does not matter
how slowly you go

as long as

you do not stop.
```

Figure 8: Example of tesseract OCR (Source: Analytics vidya)

3.1.3 PDF Handling Libraries

"PDF handling" is shorthand for "Portable Document Format" (PDF) file administration and manipulation. The Portable Document Format (PDF) is widely used because it faithfully represents the original document's layout, fonts, pictures, and formatting when viewed on any computer, regardless of operating system. Processing PDFs consists of a wide range of operations, including opening and reading the files, extracting text and graphics, and merging and separating pages. To further facilitate collaboration and navigation, PDFs can be edited to contain interactive elements, bookmarks, and notes. Converting PDFs to other formats, and vice versa, is another crucial feature. Developers can easily add PDF support to their apps utilising a wide range of libraries and tools that come along with APIs. This streamlines processes and makes life easier for consumers. PDF handling is crucial for the efficient management of digital documents and their wide-ranging applications in a variety of industries, including education, publishing, finance, and administrative work.

3.2 Vectorization

In order to transform textual input into numerical representations, the Natural Language Processing (NLP) technique known as vectorization is required. This method is what allows us to do this. Unstructured and complex linguistic data, such as sentences, paragraphs, or entire papers, is beyond the capabilities of machine learning algorithms. However, this information is useful for natural language processing. In order for computers to understand and analyse the fundamental patterns and structures contained in human language, a procedure called vectorization must be performed. This elementary technique is crucial in many applications of Natural Language Processing (NLP), including sentiment analysis, machine translation, document clustering, and more. In this study, we take a look at five vectorization strategies used in NLP and provide our findings.

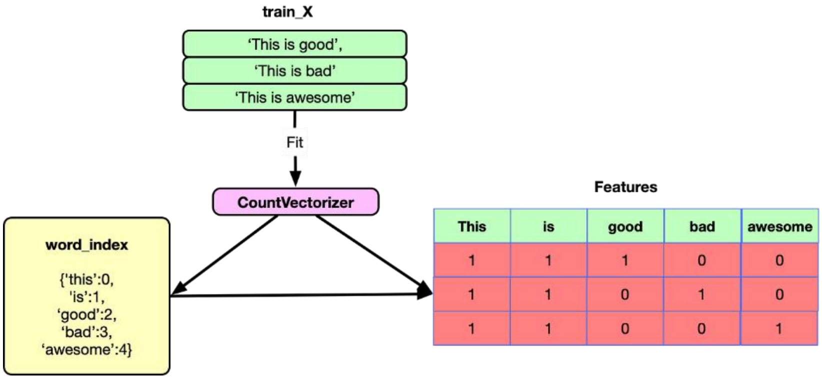
Each of these approaches provides a different means of representing textual data as vectors, so expanding the scope of linguistic research and study.

3.2.1 Count Vectorizer

One of the most crucial text preparation strategies in NLP is called CountVectorizer. With this method, text is converted into a numerical form suitable for use in machine learning.

It creates a document-term matrix, where the rows represent documents and the columns represent unique words or tokens, with the frequency of each word shown in the columns.

Once the vocabulary has been generated and tokenized, the system then tracks how often each token appears in source documents. It's simple and efficient, but it can lead to the domination of often used keywords because it doesn't take context into account. It is commonly used with methods like term frequency-inverse document frequency (TF-IDF) and word embeddings to address this problem. Despite its shortcomings, CountVectorizer is used for a wide range of purposes, such as text classification, sentiment analysis, and information retrieval. It has also been essential in the growth of advanced text analysis study.



This	is	good	bad	awesome
1	1	1	0	0
1	1	0	1	0
1	1	0	0	1

Figure 9: Countvectorizer working example

3.2.2 TF-IDF

The TF of a text may be found by dividing the number of occurrences of a phrase by the total number of words in the text. The Inverse Document Frequency (IDF) index compares the total number of documents to the fraction of those that include a word to determine the phrase's importance across all documents. By combining the power of TF and IDF, the TF-IDF algorithm gives more weight to more words. Search engines utilise the TF-IDF algorithm to prioritise material by determining which terms are most relevant and distinctive. Document annotation, sentiment analysis, and keyword extraction are all aided by the TF-IDF algorithm. However, it fails to communicate the intricate interconnections between the concepts. Two methods that attempt to address this concern are word embeddings and deep learning. The TF-IDF technique is helpful in Natural Language Processing and text analysis even when newer methods are developed.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

3.2.3 Universal Sentence Encoder (USE)

Encoding words with high-dimensional numerical vectors termed embeddings, the Universal Sentence Encoder (USE) is a powerful natural language processing paradigm. The embeddings accurately represent the sentences' semantic content. Its deep neural network architecture and extensive pre-training on text data make it suitable for a range of language processing applications, including document similarity analysis, classification, and clustering. The USE's adaptability allows it to handle text in several formats and languages, expanding the scope of its use. Document retrieval and semantic search are only two examples of the tasks that may be handled by storing semantic relationships. We must input text into the model in order to create embeddings when integrating the USE. The resulting embeddings can be used as features in subsequent procedures like sentiment analysis and classification.

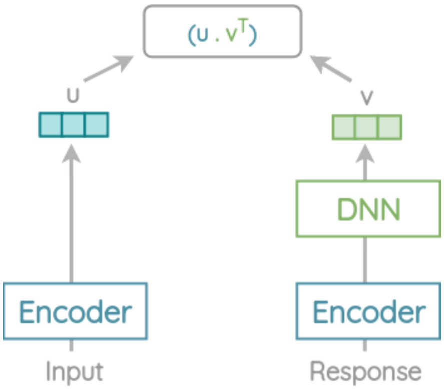


Figure 10: Working of USE

3.2.4 Sentence Transformer

Sentence Transformer 3.0 is a state-of-the-art natural language processing model for generating embeddings, or dense vector representations, of sentences. Built on the transformer architecture, it excels in understanding the meaning behind words and phrases. Token embeddings are pooled using methods like mean and max pooling to generate sentence-level vectors of a given size. This model's power lies in its capacity to complete a variety of tasks, including semantic similarity calculations, information retrieval, data clustering, and the creation of recommendation systems. It removes linguistic barriers by accommodating several languages and producing embeddings for a wide range of languages. Sentence Transformer 3.0 is a game-changer for NLP since it streamlines precise semantic search, enhances recommendation systems, and lowers the barrier to entry for cross-lingual apps. It's a revolutionary method that, with just 200 words, lets machines analyse sentences in depth, thereby boosting a wide range of natural language processing applications thanks to its robust embeddings and multilingual capabilities.

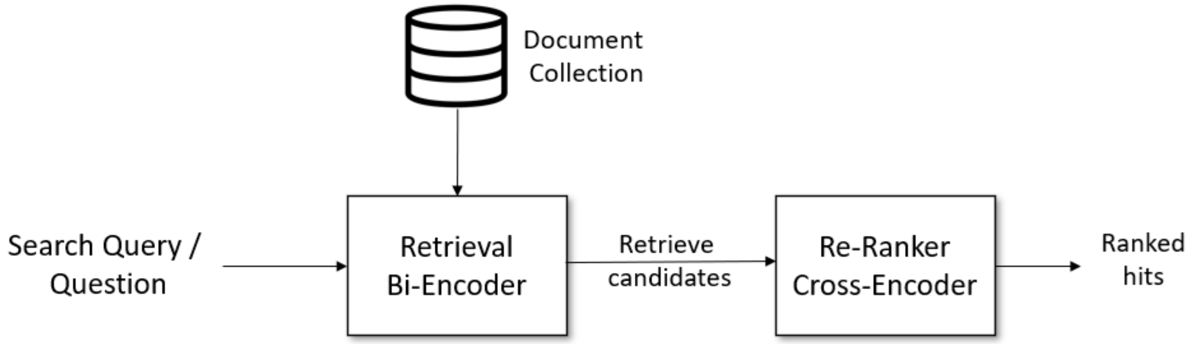


Figure 11: Sentence Transformer network

3.3 Similarity Ranking

The dot product (or scalar product) is a basic mathematical operation for finding the sum of the products resulting from the same components of two vectors. The dot product is also known as the vector product. From a geometric perspective, the operation of projecting one vector onto another is the subject at hand. The main goal here is to figure out how closely two vectors are aligned with one another. When the dot product is large, it means that the vectors are more similar to one another than when it is small. To determine how well two vectors coincide in terms of the frequency and occurrence of tokens (words), the dot product operation can be employed as a measure. To do this, the "dot product" technique is used. The idea provided previously can also be seen as a measurement of the extent to which the vectors exhibit similar properties. When the token frequency patterns of two vectors are similar, we can expect the dot product of those vectors to have a large value. This will be an indication of a higher degree of similarity between the vectors. Conversely, a smaller value for the dot product indicates less overlap and separation between the vectors. In order to determine a similarity score for each passage and query combination, a dot product operation must be performed between the vectors representing the passage and the query. This process is implemented in the similarity score calculation. The ratings that have been explained above are a means of gauging how closely the offered paragraph and the inquiry match up. Therefore, we can assess how well the assigned texts answer the research questions.

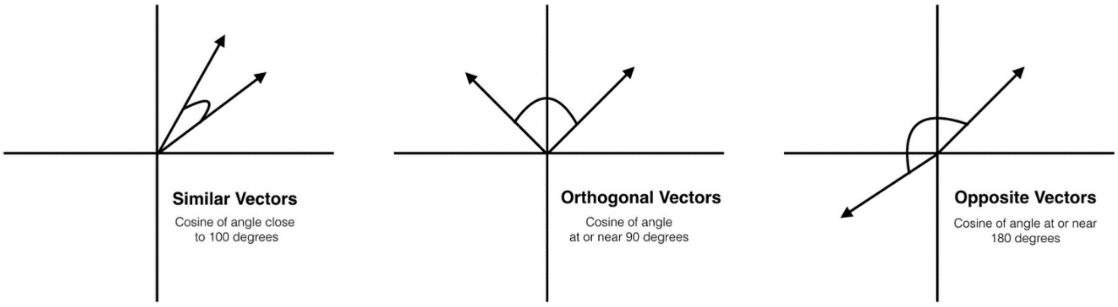

Similar Vectors
Cosine of angle close to 100 degrees
Orthogonal Vectors
Cosine of angle at or near 90 degrees
Opposite Vectors
Cosine of angle at or near 180 degrees

Figure 12: Example of similarity in the vectors

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

4. Implementation

Following is the proposed network frame that can be used for the analysis,

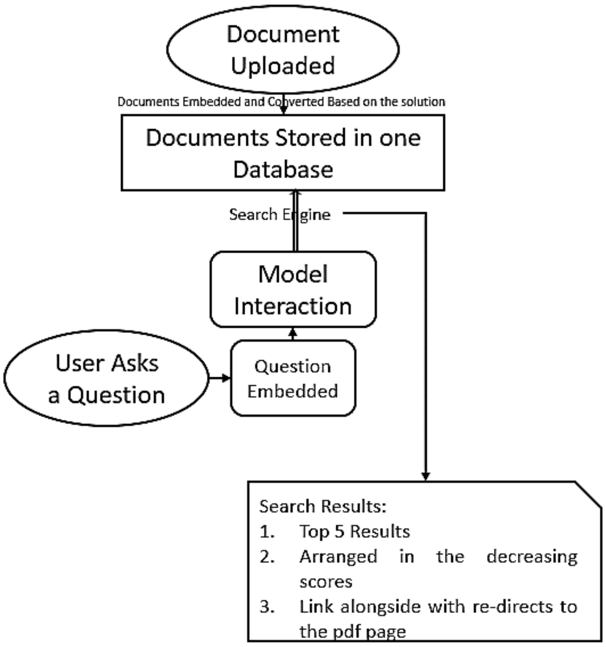

Document Uploaded
Documents Embedded and Converted Based on the solution
Documents Stored in one Database
Search Engine
Model Interaction
User Asks a Question
Question Embedded
Search Results:
1. Top 5 Results
2. Arranged in the decreasing scores
3. Link alongside with re-directs to the pdf page

Figure 13: Architecture for the Document Querying System

The flow of this system is as shown below,

Step 1 – Database Creation

a. All the Documents (in PDF needs to be converted and Embedded) – Document Extractor (Using py2PDF or PDFMiner)
b. PDFs in Italian Language/English – Transformers models like BERT
c. No APIs for the Language Translation – if any different languages are present.
d. The Database of all the documents needs to be created for Same

Step 2 – Question Asked

a. Questions can be in any Language (Preferably in Italian)
b. Search Question needs to be only One. It shouldn't be an append of two separate questions since the intent can vary.
   [Solution – Universal Sentence Encoder, Word Embedding, Extra Features Extraction and Usage of BERT, Other Recommender Filtering Techniques]

Step 3 – Question Searched

a. All the top (5/10/) answers based on similarity to be shown – [Similarity Measures – Cosine, Jaccard and L1 and L2 Norms]

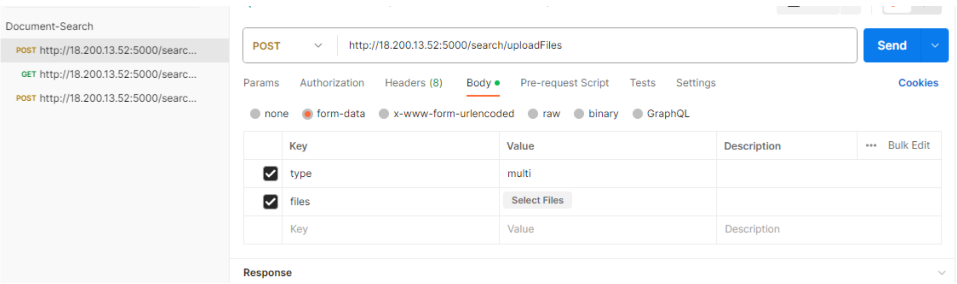
4.1 Cloud Framework

A Python Flask-based web application is deployed on an EC2 instance which allows users to upload PDF files, extracts text content from the uploaded files, and performs similarity-based search to find the most relevant matches for user queries. It leverages the power of TF-IDF vectorization for efficient and accurate text comparison.

Endpoint /search/uploadFiles (POST): This endpoint allows users to upload multiple PDF files. The uploaded files are saved to the server's file system under the UPLOAD_FOLDER (public/docs). The application checks if the uploaded files already exist in the database; if not, it saves their information (filename and creation date) to the database. The text content is then extracted from these PDF files and stored in the database, associating each extracted paragraph with the corresponding file and page numbers.

Endpoint /search/sync (GET): This endpoint is used to asynchronously extract text content from PDF files. It checks for new PDF files in the specified UPLOAD_FOLDER (public/docs). If any new files are found, they are saved to the database with their filenames and creation dates. The text content is then extracted from these PDF files and stored in the database,similar to the /search/uploadFiles endpoint. The text extraction is performed in the background using threading to improve application performance.

Endpoint /search/predict (POST): This endpoint receives a user's search question in JSON format. It performs a similarity search using TF-IDF (Term Frequency-Inverse Document Frequency) vectors. The application calculates the similarity between the user's query and the text content of PDF files stored in the database. The search returns the top 10 matches based on similarity scores, i.e., the most relevant paragraphs that match the user's search query.


Document-Search
POST http://18.200.13.52:5000/searc...
GET http://18.200.13.52:5000/searc...
POST http://18.200.13.52:5000/searc...
Send
Params
Authorization
Headers (8)
Body
Pre-request Script
Tests
Settings
Cookies
○ none ● form-data ○ x-www-form-urlencoded ○ raw ○ binary ○ GraphQL
Bulk Edit
Response

POST http://18.200.13.52:5000/search/uploadFiles

	Key	Value	Description
☑	type	multi	
☑	files	Select Files	
	Key	Value	Description

Figure 14: The APIs are accessed through Postman

5. Evaluation

Whether the desired result has been in the 10 recommended results or not. If Yes == 1 else, No == 0

$$CR = \frac{C}{A}$$

CR – The correct rate;

C – The number of sample recognized correctly;

A – The number of all sample;

Ranking Document Accuracy= rating from 1 to 10 with 1 being the highest.

Result in rank 1 = 1

Result in rank 2 = (10-1)/10 = 0.9

.

.

.

Result in rank 10 = 0.1

$$Final\ Selection\ Score\ =\ document\ accuracy\ *\ ranking\ score$$

More the FSS better the model

5.1 Case 1 – Question: Com'è la performance con case insensitive?

Correct Answer – Naturalmente il cliente deve essere avvisato che in caso di Case Insensitive le performance subiranno un lieve peggioramento.

Answer Number	COUNT VECTORIZER	USE	SENTENCE TRANSFORMER	TF-IDF
1	Mentre COMPED continua a dare errori ma su questo non possiamo intervenire.	Naturalmente il cliente deve essere avvisato che in caso di Case Insensitive le performance subiranno un lieve peggioramento.	Naturalmente il cliente deve essere avvisato che in caso di Case Insensitive le performance subiranno un lieve peggioramento.	Mentre COMPED continua a dare errori ma su questo non possiamo intervenire.
2	Tale flag permetterA la gestione Case Sensitive o Case Insensitive a discezione dell'utente.	Tale flag permetterA la gestione Case Sensitive o Case Insensitive a discezione dell'utente.	Tale flag permetterA la gestione Case Sensitive o Case Insensitive a discezione dell'utente.	Naturalmente il cliente deve essere avvisato che in caso di Case Insensitive le performance subiranno un lieve peggioramento.
3	Oracle - gestione globale Case Insensitive	Oracle - gestione globale Case Insensitive	Oracle - gestione globale Case Insensitive	Oracle - gestione globale Case Insensitive
4	SCHEDA DI LAVORO N. SL/4097 Riferimento interno 20973	Descrizione ll problema era in Entity. Aggiunto IGNORENULL sulla proprieta JOBID	Mentre COMPED continua a dare errori ma su questo non possiamo intervenire.	Tale flag permetterA la gestione Case Sensitive o Case Insensitive a discezione dell'utente.
5	Se sono in insert devo vedere il campo NOTIFICA, se sono in update devo vedere il campo NOTIFICAREV	24/01/2011 10.04.01: DM_FAXOUT_SENDFAX --> 127.0.0.1 -> \n Proprieta: JOBID erro re La dimensione della stringa non é valida	Tecnicamente cambia la query eseguita a seconda dell'impostazione.	Se sono in insert devo vedere il campo NOTIFICA, se sono in update devo vedere il campo NOTIFICAREV

6	WCF	Fascicolo senza il web visible attivo	Descrizione	Select * from DM_TABELLE Where TABELLA = 'STATO' and VALORE = 'StatoDiArchiviazione ' Si deve estrarre lo stato.
7	DATA CREAZIONE 24/01/2011 VERSIONE	Il comportamento é analogo su tutti i provider.	WRITER DI ARXIVAR: passa al WCF una archiviazione con lista destinatari lunga 1 ma con unico	DATA CREAZIONE 24/01/2011 VERSIONE
8	4.6.13 RIFERIMENTO ABLETECH\r.zanardini - Nessuno TITOLO DM_INOLTRO	Il comportamento é analogo su tutti i provider.	Come gia spiegato via mail io ho corretto il BUG sia sulla Branch 4.6.12 che sulla online version. Peccato che non Si riesca a testare sulla 4.6.12.	Fascicolo senza il web visible attivo
9	A fronte di una archiviazione con la query	Il comportamento é analogo su tutti i provider.	Descrizione	Tale flag permetterA la gestione Case Sensitive o Case Insensitive a discezione dell'utente.
10	Select * from DM_TABELLE Where TABELLA = 'STATO' and VALORE = 'StatoDiArchiviazione ' Si deve estrarre lo stato.	Assicurarsi che quando si chiude il pannello dell'anteprima che word vada giu dai processi (ovviamente se non c'é altro aperto)	Descrizione	Fascicolo senza il web visible attivo

Calculation: Count Vectorizer doesn't has any correct answer hence the score is 0*1 = 0
USE: 1*1 = 1
Sentence Transformer: 1*1 = 1
TF-IDF: 1*0.9 = 0.9

5.2 Case 2 – Question: Qual è il problema se si esporta un pdf da Word senza anteprima?

Correct Answer – successivamente se tento l'esportazione in pdf senza anteprima, viene mostrato comunque il messaggio di errore di prima
Calculation: Count Vectorizer: 1*(1-6)/10 = 1*0.4 = 0.4
USE: 1*(1-3)/10 = 0.7
Sentence Transformer: 1*(1-3)/10 = 0.7
TF-IDF: 1*(1-4)/10 = 0.6

5.3 Case 3 – Question: A chi sono estesi i permessi della rubrica?

Correct Answer: i permessi sulla rubrica devono essere estesi anche ai contatti
Calculation: Count Vectorizer: 1*(1-8)/10 = 1*0.2 = 0.2
USE: 1*(1-5)/10 = 0.5
Sentence Transformer: 1*(1-0)/10 = 1

TF-IDF: 1*(1-0)/10 = 1

5.4 Final Analysis

Table : Overall 50 questions combined performance

Model Name	Correct	Wrong	Accuracy	Score
Sentence Encoder - Hugging Face	49	1	98%	98%
TF-IDF	47	3	94%	82%
Count Vectorizer	35	15	70%	42%
Universal Sentence Encoder	45	5	90%	66%

FastText	48	2	96%	71%

In the above table, the Sentence Encoder achieved the highest accuracy at 98%, with only one wrong prediction. TF-IDF had an accuracy of 94%, while CountVec reached 70% accuracy. The Universal Sentence Encoder (USE) achieved 90% accuracy, and FastText performed well at 96% accuracy. Each model's correct, wrong predictions, accuracy percentage, and a score were recorded and analyzed. The Sentence Encoder excelled, while the other models showed varying levels of accuracy and effectiveness in the given task.

6. Conclusion and Future Work

This study has provided a comprehensive evaluation of several methods for retrieving documents. Among these methods are the use of improved sentence encoders, the TF-IDF algorithm, and the count vectorizer for semantic search. Semantic search has been shown to be more effective than traditional methods at locating and retrieving the most pertinent parts of documents in response to user queries. Semantic search has proven this to be the case. By effectively capturing the meaning of language in its context of usage, semantic search, which is powered by sentence encoders, has significantly increased the accuracy of our document search engine and the quality of the user experience it provides. The results of this study show how important it is for modern information retrieval systems to make use of natural language processing techniques. Semantic search, made feasible by contemporary sentence encoders, has been shown to close the gap between the user's intent and the text's substance. Understanding the semantics of language has led to a more nuanced and user-centric approach to document search. In summary, this study's results show that semantic search powered by sophisticated language encoders can drastically improve document retrieval. As we move farther into the age of sophisticated natural language comprehension, there will be several opportunities to develop and optimise document search engines in order to better satisfy the evolving needs of its users.

Several promising new lines of inquiry that could lead to even more significant discoveries have been uncovered by this investigation. By better understanding the user's purpose, search precision can be increased with techniques like sentiment analysis and context modelling. The improvement of interpretable semantic search approaches, such as attention visualisation, will increase user trust by providing users with greater insight into the relevance of search results. If the search engine could be enhanced to incorporate text, graphics, and audio in cross-model searches, it might provide users with a more immersive retrieval experience. It may be necessary to look into transfer learning procedures as a means to increase language support across multilingual semantic search. By leveraging parallel processing and hardware acceleration to increase the real-time efficiency of semantic search, response times could be reduced. Search results may be more customised to the user's needs and constantly improved if active learning was used to incorporate user preferences and incorporate user interactions. All of these possibilities lead to an exciting future of sophisticated document retrieval systems that meet the needs of a wide range of users.

References

Balaji, P., Merker, H. and Gupta, A., 2023. EntailClass: A Classification Approach to EntailSum and End-to-End Document Extraction, Identification, and Evaluation. *JUCS: Journal of Universal Computer Science*, *29*(1).

Mishchenko, A., Danco, D., Jindal, A. and Blue, A., 2022. Blueprint: a constraint-solving approach for document extraction. *Proceedings of the VLDB Endowment*, *15*(12), pp.3459-3471.

Zhang, T., Lee, B., Zhu, Q., Han, X. and Chen, K., 2023. Document keyword extraction based on semantic hierarchical graph model. *Scientometrics*, *128*(5), pp.2623-2647.

Arya, C., Diwakar, M., Singh, P., Singh, V., Kadry, S. and Kim, J., 2023. Multi-Document News Web Page Summarization Using Content Extraction and Lexical Chain Based Key Phrase Extraction. *Mathematics*, *11*(8), p.1762.

Meuschke, N., Jagdale, A., Spinde, T., Mitrović, J. and Gipp, B., 2023, March. A Benchmark of PDF Information Extraction Tools Using a Multi-task and Multi-domain Evaluation Framework for Academic Documents. In *International Conference on Information* (pp. 383-405). Cham: Springer Nature Switzerland.

Banerjee, S., Mukherjee, S., Bandyopadhyay, S. and Pakray, P., 2023. An extract-then-abstract based method to generate disaster-news headlines using a DNN extractor followed by a transformer abstractor. *Information Processing & Management*, *60*(3), p.103291.

Guo, J., Kok, S. and Bing, L., 2023. Towards Integration of Discriminability and Robustness for Document-Level Relation Extraction. *arXiv preprint arXiv:2304.00824*.

Sun, Q., Zhang, K., Huang, K., Xu, T., Li, X. and Liu, Y., 2023. Document-level relation extraction with two-stage dynamic graph attention networks. *Knowledge-Based Systems*, *267*, p.110428.

Naseem, T., Blodgett, A., Kumaravel, S., O'Gorman, T., Lee, Y.S., Flanigan, J., Astudillo, R.F., Florian, R., Roukos, S. and Schneider, N., 2021. DOCAMR: Multi-Sentence AMR Representation and Evaluation. *arXiv preprint arXiv:2112.08513*.

Ding, M.A. and Goldfarb, A., 2023. The economics of artificial intelligence: A marketing perspective. *Artificial Intelligence in Marketing*, *20*, pp.13-76.

Mahmood, M., Al-Kubaisy, W.J. and Al-Khateeb, B., 2023. Multimedia information retrieval using artificial neural network. *IAES International Journal of Artificial Intelligence*, *12*(1), p.146.

SAHITHYA, A. and KUMAR, S.V., 2023. BUILDING SEARCH ENGINE USING MACHINE LEARNING TECHNIQUES. *Journal of Engineering Sciences*, *14*(01).

Adrakatti, A.F. and Mulla, K.R., 2023. A Fast and Full-Text Search Engine for Educational Lecture Archives. *Code4Lib Journal*, (55).

Priya, V.D. and Chakkaravarthy, S.S., 2023. Containerized cloud-based honeypot deception for tracking attackers. *Scientific Reports*, *13*(1), p.1437.