IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 8, August 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Semantic Vector Search: Intelligent Document Retrieval System

Mr. Shankar Prabhath U, Mr. Darshan, Dr. T. Subbaraj

Department of Masters of Computer Applications, Rajarajeshwari College of Engineering, Bangalore, Karnataka, India

ABSTRACT: The "Seman Vector Search: Intelligent Document Retrieval System" is an advanced informa on retrieval system designed to improve the accuracy and relevance of document searches. U lizing state-of-the-art embedding models, the system converts documents into high-dimensional vector representa ons, capturing their seman c meaning. These embeddings are stored in a specialized vector database, enabling efficient and precise similarity searches. When a user submits a query, it is transformed into an embedding, and the system retrieves the most relevant documents by finding the closest vectors in the DataBase This approach significantly improves search results by comprehending the context and meaning of both queries and documents, making it ideal for applica ons such as search engines, recommenda on systems, and various Natural Language processing tasks. The project aims to demonstrate the efficacy of embedding-based search techniques in delivering more pertinent and contextually accurate information.

KEYWORDS: Natural language interpretation.

I. INTRODUCTION

In the era of informa on overload, retrieving factual and pertinent documents efficiently is a cri cal challenge across various domains, including academia, enterprise, and the web. Tradi onal keyword-based search systems o en fall short in understanding the seman context of queries and documents, leading to subop mal search results. The "Seman Vector Search: Intelligent Document Retrieval System" addresses this challenge by utilizing cutting-edge Machine Learning method to enhance the accuracy and relevance of document retrieval. At the core of this system in the use of embeddings—dense vector representa ons of documents generated by state-of-the-art natural language interpretation (NLP) models. These embeddings capture the seman c meaning of documents, enabling the system to perform more accurate similarity searches compared to tradi onal keyword-based approaches. By storing these embeddings in a specialized vector database, the system can efficiently handle high-dimensional data and quickly retrieve the most relevant documents in response to user queries. When a user inputs a query, it is transformed into an embedding using in same NLP model that processed the documents. The query embedding is then use to search the vector database, retrieving documents whose embeddings are closest in response to the question in the highdimensional vector space. This process ensures that The papers that were recovered are not just relevant keywords but also in terms of their overall seman c content.

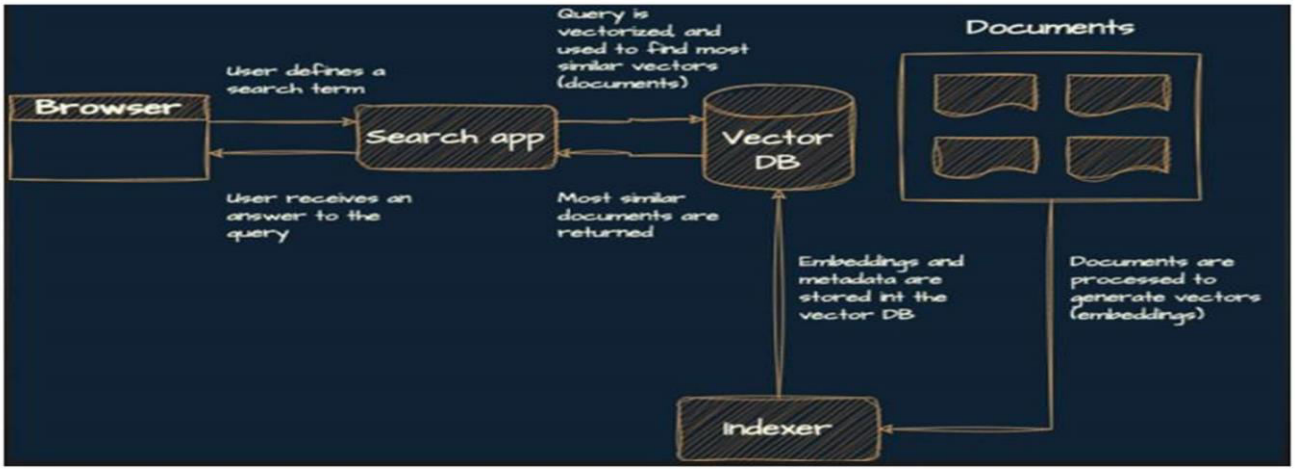
II. RELATED WORK

The query embedding is then use to search the vector database, retrieving documents whose embeddings are closest in response to the question in the highdimensional[1] vector space. This process ensures that The papers that were recovered are not just relevant keywords but also in terms of their overall seman c content[2].

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



III. METHODOLOGY

Data Collection and Preparation
Source Data: Large and diverse datasets are collected, including books, ar cles, websites, and other text sources.

Preprocessing: Data is cleaned and tokenized. This includes removing unwanted characters, handling special tokens, and spli ng the text into manageable pieces.

Data Collection and Preparation
Document Corpus: Collect a large corpus of documents for the domain of interest.

Preprocessing: Clean the data (remove HTML tags, special characters), tokenize the text, and create training samples (e.g., pairs of related documents).

Pre-training Phase
Model Setup: Ini alize the Mistral/Mixtral model with transformer layers.

Training Objec ve: Train the model on tasks like masked language modeling (MLM) and next sentence predic on (NSP) using the collected corpus.

Training: Use distributed training on GPUs to handle the large dataset and complex computa ons.

IV. EXPERIMENTAL RESULTS

Machine learning algorithms and natural language processing (nlp) are utilizedin semantic search, an advanced search technology, to grasp the meaning and context of search queries, providing more accurate and relevant search results. If we ask something related to science, mathematics, and other related questions it will be give output as answer for that query using mistral it going to trained model for that related query

V. CONCLUSION

Semantic vector search represents a significant advancement in the studded on intelligent systems for retrieving documents. Through utilizing The possibilities of natural language process and deep learning techniques (NLP), this method transcends traditional keyword-based searches to offer more nuanced and contextually relevant results.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Enhanced Relevance: Semantic vector search improves the relevance of search outcomes by comprehending the meaning and context behind user queries. This results in more precise and meaningful document retrieval, aligning with the user's intent rather than just matching keywords.

Contextual Understanding: Unlike conventional search systems that rely heavily on exact keyword matches, semantic vector search comprehends the difference between phrases and words. This enables it to obtain semantically relevant texts even without the exact search terms

REFERENCES

1. Chen, K., Corrado, G., Mikolov, T., & Dean, J. (2013). Distributed Representa ons of Word and Phrases and their Composi onality. Neural Informatics Advancement on Process Systems, 26, 3111-3119.
2. Vaswani, A., Jones, Lsu Gomez, A. N., Kaiser, Ł., Uszkoreit J Shazeer, N., Parmar, N., & Polosukhin, I. (2017). All your's needs is an en on. Neural Informa Advances on Processing System, 30, 5998-6008
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirec onal Transformers for Language Understanding. Proceeding of the 2019 Conference of the North American Chapter of the computer assoceation onal Linguis cs: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.
4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Mul task Learners. OpenAI Blog. Describes the GPT series of models, which are effective instruments for seman c vector search due to their contextual understanding capabili es.
5. Guo, J., Fan, Y., Ai, Q., & Cro , W. B. (2016). A Deep Relevance Matching Model for Ad Hoc Retrieval. The 25th ACM worldwide Conference on Information and Knowledge Management: Proceeding, 55-64.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details