Using RAG, LLMs, and LangChain to Create a Query Application on Open Repositories

Demo Application on DSpace-7

Le Yang
University Libraries
University of Oregon
Eugene Oregon USA
yanglegd@gmail.com

Zhongda Zhang
University Libraries
University of Oregon
Eugene Oregon USA
zhongda.zhang@gmail.com

ABSTRACT

The project introduces an innovative methodology that leverages retrieval-augmented generation (RAG), large language models (LLMs), embeddings, and LangChain to create an advanced query application for open repositories, exemplified by DSpace-7 prototype at the University of Oregon Libraries. Using RAG techniques, the application seamlessly integrates relevant information from repository data, including database records and textual documents. LLMs, like GPT-4o, form the foundation for context-aware interactions, while embeddings enhance query comprehension and relevance. LangChain acts as a mediator, facilitating user interaction with repository systems. The goal is to improve user engagement, streamline information retrieval, and enhance the overall experience in open repository environments.

CCS CONCEPTS

• Human-centered computing • Information systems → Web applications • Information systems → Digital libraries and archives

KEYWORDS

Open Repositories, DSpace, Retrieval-Augmented Generation (RAG), Large Language Model (LLM), Embeddings, LangChain

1 OVERVIEW

AI has significantly impacted the advancement of cultural heritage institutions and the development of applications, affecting various aspects of workflows and practices [3, 4]. It has enhanced digital repositories' functionality, searchability, and effectiveness, making them more valuable tools for researchers, educators, and the general public [6, 8]. The utilization of AI-powered search algorithms, such as natural language processing (NLP) and machine learning, has significantly improved the precision and relevance of search results within repositories [1]. AI-driven chatbots and virtual assistants provide users with user-friendly and interactive interfaces to interact with information sources, respond to user queries, guide through the repository, and assist in complex searches [5]. AI plays a transformative role in the development of open repositories, speeding up innovation, enhancing discovery, and increasing the usability and accessibility of deposited resources [2, 7]

To continue advancing the AI-related development on open repositories, we started an experimental project, creating an AI-powered query application that enable users to use NLP to search open repositories (Figure 1). The main technologies we are using include RAG, LLMs, embedding, LangChain, and Streamlit. To demo the application, we developed a prototype on the University of Oregon's DSpace-7 institutional repository Scholars' Bank.

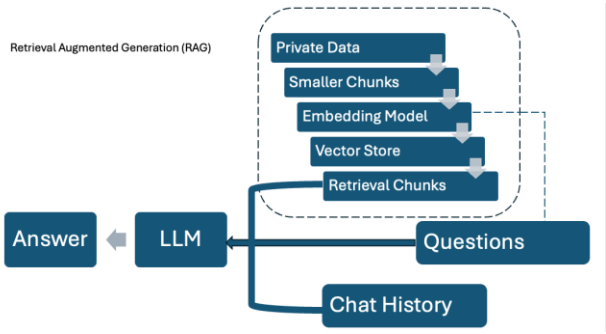

Figure 1: Overview of the Query Application

2 EMBEDDINGS & DATA PREPROCESSING

In ML and NLP, embeddings represent the semantic understanding of words, phrases, and documents in a continuous vector space. These embeddings serve as a bridge, converting data into a lower-dimensional space, enabling meaningful connections. While pre-training allows us to use pre-trained embedding models for a broad vocabulary, private or proprietary data sources in our demo repository necessitate developing custom embedding models. This approach maximizes the potential of our document data for ML applications.

The entire process of transforming files into embeddings involves multiple steps. We start with using LangChain document loader to ingest data, then utilized text splitters to divide large documents into chunks, making it easier to index and process the data. We then select an API endpoint, offered by OpenAI, to convert documents into vector data, stored and queried using ChromaDB, a vector database.

3 RAG & INTERACTION WITH PDFs

AI can engage in conversations with PDFs using a process combining NLP, LangChain, and LLM. As shown in Figure 2, when a question is provided, LangChain converts the query into vectors and searches the document vectors for the most pertinent matches. LangChain then transmits the user query and relevant text chunks to LLM so that it can generate a response based on the input provided. The process of identifying and transmitting relevant text information to the LLM is also called RAG.

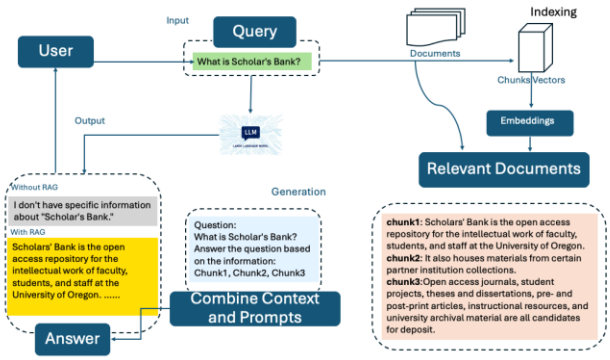

Figure 2: How Query Application Functions

4 QUERYING DATABASE: NLP & LLMs

LangChain facilitates SQL query creation and execution based on natural language prompts through SQL Chains and Agents. As a bridge between natural language and databases, LangChain leverages techniques like LLMs to understand user queries. These LLMs are trained on extensive text data to interpret user intent accurately. Prompt engineering guides LLMs in generating customized output by providing instructions or templates specific to the database structure and query type. LangChain then translates the natural language query into an equivalent SQL statement using the LLMs' output and the retrieved data structure. This seamless interaction allows users to query the database using its native language. Finally, LangChain employs LLMs again to format raw data into a user-friendly response.

5 STREAMLIT UI

Streamlit, an open-source app framework, simplifies data application creation. Its Python API allows easy UI development without knowledge of HTML, CSS, or JavaScript. Streamlit offers widgets for user interaction and real-time data visualization. It integrates with ML libraries, APIs, databases, and external services. In our Streamlit app (Figure 3), we firstly import Streamlit into the application, then add the Streamlit UI and design the layout by using the built-in options. We then used Python data structure to store chat history and user inputs. Lastly, we utilized an API to communicate with the AI model and display its response in the app interface.

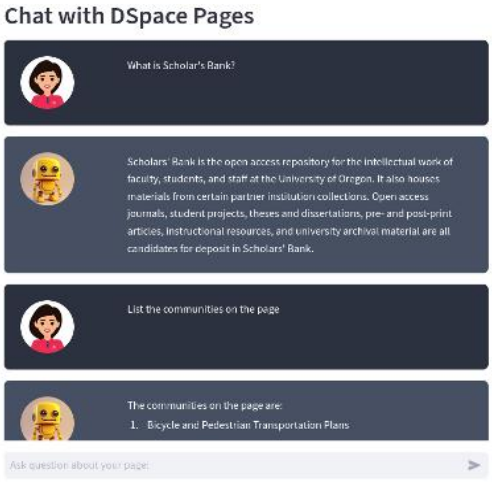

Figure 3: Streamlit UI

6 CONCLUSION

By utilizing advanced NLP techniques and conversational AI, the proposed approach offers a promising solution for effective discovery and interaction with repositories, as well as enhancing the accessibility and usability of the systems. Future development tries to focus on three improvements: expanding the range of supported file formats beyond PDFs, replacing paid AIs and LLMs with open models while ensuring the same level of performance, upscaling its capabilities to other open repositories and potentially the entire library search.

REFERENCES

[1] Dilawar Ali, Kenzo Milleville, Steven Verstockt, Nico Van de Weghe, Sally Chambers, and Julie M. Birkholz. 2023. Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections. *Journal of Documentation*, ahead of print (2023). DOI: https://doi.org/10.1108/JD-01-2022-0029.

[2] Kirsten Carter, Abby Gondek, William Underwood, Teddy Randby, and Richard Marciano. 2022. Using AI and ML to optimize information discovery in under-utilized, Holocaust-related records. *AI & SOCIETY* 37, 3 (2022), 837-858. DOI: https://doi.org/10.1007/s00146-021-01368-w.

[3] Ryan Cordell. 2020. *Machine learning + libraries: A report on the state of the field*. Retrieved from https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

[4] Amber Cushing and Giulia Osti. 2023. "So how do we balance all of these needs?": How the concept of AI technology impacts digital archival expertise. *Journal of Documentation* 79, 7 (2023), 12-29. https://doi.org/10.1108/JD-08-2022-0170.

[5] Aleksandra Fostikov. 2023. First impressions on using AI powered chatbots, tools and search engines: ChatGPT, perplexity and other–possibilities and usage problems. *Review of the NCD* 42, (2023), 12-21. http://rih.iib.ac.rs/1219/1/81%20A_Fostikov_First%20Impressiones_ENG%20SRB_NCD%2042%20%282023%29%201-21.pdf.

[6] Lise Jaillant and Arran Rees. 2023. Applying AI to digital archives: trust, collaboration and shared professional ethics. *Digital Scholarship in the Humanities* 38, 2 (2023), 571-585. https://doi.org/10.1093/llc/fqac073.

[7] Clemens Neudecker. 2022. Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. In Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022), Sept. 19th-23rd, 2022 (CEUR Workshop Proceedings, Vol. 3234). https://ceur-ws.org/Vol-3234/paper2.pdf.

[8] Lilin Yu, Ash Charlton, Wilfrid Askins, Melissa Terras, and Rosa Filgueira. 2023. Frances: cloud-based historical text mining with deep learning and parallel processing. In *2023 IEEE 19th International Conference on e-Science (e-Science)*, pp. 1-10. IEEE, 2023. https://doi.org/10.1109/e-Science58273.2023.10254798.