ASSIGNMENT REPORT

MULTIPLE DOCUMENT CHAT Q/A INTERFACE

The Interface is an advanced application that combines ChromaDB vector stores, Sentence Transformer embeddings, and the Llama-7 large language model hosted on Replicate to offer a sophisticated question-and-answer experience across multiple documents. ChromaDB vector stores efficiently manage document representations, while Sentence Transformer embeddings enhance the semantic understanding of text. Llama-7, with its state-of-the-art natural language processing capabilities, interprets user queries accurately, providing comprehensive responses. The user-friendly Streamlit deployment ensures a seamless and visually appealing experience, making this interface a powerful tool for precise information retrieval from diverse textual sources.

DATASET

This dataset is a collection of text files of Amazon Web Services (AWS) case studies and blog articles related to Generative AI and Large Language Models. The dataset was obtained in Kaggle used to train RAG pipelines.
Dataset: A Case study dataset: This https://www.kaggle.com/datasets/harshsinghal/aws-case-studies-and-blogs?select=CloudCall+Invests+in+AWS+Skill+Builder+Pivots+to+a+SaaS+Model+_+CloudCall+Case+Study+_+AWS.txt.

RAG TECHNIQUES

Sentence Transformer was used for obtaining the embeddings of the inputted text. The transformer used is: all-MiniLM-L6-v2.

It maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search. This Sentence transformer is used also because it is incredibly fast. The models are based on transformer networks like BERT / RoBERTa / XLM-RoBERTa etc. and achieve state-of-the-art performance in

various tasks. Text is embedded in vector spaces such that similar text is close and can efficiently be found using cosine similarity.

The Retrieval Process is done as follows:

1) Loading: A variety of files such as PDF,DOC,DOCX,TXT,JPG,JPEG,PNG,CSV files and so on can be accepted and converted in to text format. These Text formats are then scraped and loaded into a text variable text.
2) Splitting: **CharacterTextSplitter** is used to break a long text into smaller chunks of 1000 characters each, with a 100-character overlap between chunks to help store Semantic information, creating a list of text chunks.
3) Embedding: An HuggingFace Embedding model the **all-MiniLM-L6-v2** is then used to create the embeddings out of the sentences of the text data

VECTOR DATABASE USED

Chroma vector database was used to store the embeddings of the inputted data as it supports real time updation of the vector database. **Chroma** is a lightweight in-memory DB, so it's ideal for prototyping.Chroma also retrieves metadata of its documents along with its embeddings, making it useful for displaying sources of its answers.

MODEL HALLUCINATION

Prompt engineering can be used in providing context and preventing generated answers from straying beyond the scope of allotted vector data in large language models (LLMs). In this way, prompt engineering can be used to prevent hallucination of LLMs. In the submitted assignment, a prompt is made to ensure the responses of the chatbot stays within the vector database provided.

MMR (Maximum Marginal Relevance) to obtain the best possible responses by increasing the diversity of the response. MMR tries to reduce the redundancy of results while at the same time maintaining query relevance of results for already ranked documents/phrases etc.

CONCLUSION

There are still further improvements that can be made to the project to improve the relevancy and accuracy of results whilst also improving resource allocation.

Methods like Self-Querying and compression can be used to further accuracy, while the LLAMA2 model can also be further fine tuned to accustom it to a specific domain. To help adapt the model to work with a large number of documents, methods like MapReduce could be implemented in the Chain to help improve the performance.

SOURCES

RAG Optimization

Retrieval with LLAMA2