Healthcare Guide Chatbot with RAG and Vector Database

Table of Contents

1. Introduction

This report details the development of a Healthcare Guide Chatbot that combines Large Language Models (LLM) with a vector database to provide personalized healthcare recommendations. The chatbot is built using Flowise and integrates various components to deliver accurate and relevant healthcare advice.

2. Domain Selection

Domain: Healthcare guide
Scope: The chatbot provides personalized health recommendations and advice based on user inputs, focusing on symptoms, diet plans, and management of health conditions.

3. Data Collection and Preprocessing

Data Collection: The healthcare data was collected from publicly available health guides, medical websites, and datasets. The dataset includes symptoms, diet plans, stress management techniques, and other healthcare-related information.

Preprocessing: The collected data was cleaned and preprocessed to ensure it was in a suitable format for indexing and retrieval. This included removing irrelevant information, handling missing data, and ensuring consistency in the text format.

4. Vector Database Implementation

Database Selection: Pinecone was chosen as the vector database for its efficiency in handling semantic similarity searches.

Indexing: The preprocessed healthcare data was indexed in Pinecone. This involved creating embeddings using OpenAI's text-embedding-ada-002 model and storing these embeddings in Pinecone for efficient retrieval.

5. Application Development

User Interface: The chatbot interface allows users to input natural language queries related to their healthcare needs.

Backend Integration: The backend utilizes LangChain to manage the flow of interactions. It combines LLM for query processing with the vector database for data retrieval.

Components Used:

- **Recursive Character Text Splitter**: Splits text into manageable chunks for processing.
- **CSV File**: Manages the healthcare data.
- **OpenAI Embeddings**: Creates embeddings for queries and context.
- **Pinecone Upsert Document**: Stores and indexes data in Pinecone.
- **Conversational Retrieval QA Chain**: Retrieves relevant information.
- **ChatOpenAI**: Main chat interface for user interaction.

6. Evaluation and Testing

The application was tested with various healthcare-related queries to ensure performance, accuracy, and usability. Example queries include:

- "What are the symptoms of diabetes?"
- "Suggest a diet plan for hypertension."
- "How to manage stress effectively?"

7. Challenges and Solutions

Data Quality: Ensuring the collected data was clean and relevant was a challenge. This was addressed by thorough preprocessing and validation.

Embedding Accuracy: Creating accurate embeddings for the healthcare data was crucial. Fine-tuning the embedding model and validating the embeddings helped improve accuracy.

User Interaction: Designing an intuitive user interface that handles natural language queries effectively required iterative testing and user feedback.

8. Conclusion

The Healthcare Guide Chatbot successfully combines LLM and vector database technologies to provide personalized healthcare recommendations. The system efficiently processes user queries and retrieves relevant information, demonstrating the potential of combining advanced AI techniques for practical applications.

9. References

- [Flowise Documentation](#)
- [OpenAI Documentation](#)
- [Pinecone Documentation](#)

Video Demonstration

A detailed video walkthrough of the application demonstrating its features and usage can be found [here](#).

Source Code and Documentation

The source code and documentation explaining the implementation details can be found in the [GitHub Repo](#).