History of Massachusetts Q&A Bot – Project Report

1. Introduction

The "History of Massachusetts - A RAG System Q&A Bot" is a Retrieval-Augmented Generation (RAG) application designed to provide users with accurate answers to questions about Massachusetts' history. The system uses advanced AI techniques, integrating a vector store and OpenAI's GPT-3.5-turbo, to offer concise and contextually relevant responses.

2. Objectives

- To create a user-friendly Q&A system focused on Massachusetts' history.
- To utilize RAG methods for efficient data retrieval and contextual answering.
- To demonstrate the power of AI in domain-specific knowledge applications.
- To assist students, researchers, and history enthusiasts in exploring historical data efficiently.

3. Data Collection and Preprocessing

3.1 Sources of Data

Wikipedia: Pages about Massachusetts' history were scraped using the Wikipedia API.

3.2 Data Preparation

- Text was retrieved from relevant Wikipedia pages.

- Content was split into structured sections using recursive processing.
- The processed data was stored in CSV and Markdown formats.

4. System Architecture

4.1 Core Components

1. Vector Database: FAISS was used for similarity-based retrieval of document embeddings.
2. LLM: OpenAI's GPT-3.5-turbo powered the language understanding and generation tasks.
3. Frontend: Streamlit-based user interface for querying and displaying results.

4.2 RAG Workflow

1. Input queries are processed to retrieve relevant context from the vector store.
2. Contextual compression filters unnecessary data to refine the retrieved documents.
3. The language model generates precise answers using the retrieved context.

5. Implementation

5.1 Tools and Technologies

- Scraping: Wikipedia API for structured and clean data extraction.
- Embedding Generation: HuggingFace's all-MiniLM-L6-v2 model.
- Database Management: FAISS for vector storage and similarity search.
- Model Fine-Tuning: No explicit fine-tuning; used GPT-3.5-turbo for inference.

- Deployment: Streamlit for hosting and interaction.

5.2 Key Features

- **Contextual Compression**: Reduces noise in retrieved documents using LLMChainExtractor.
- **Adaptive Retrieval**: Retrieves top 5 most relevant chunks for accurate responses.
- **Interactive UI**: Allows users to input questions and receive concise answers.

6. Results

- Successfully retrieved and answered historical queries about Massachusetts.
- Demonstrated high accuracy in retrieving relevant data using FAISS.
- Reduced response time for queries through efficient embeddings and compression.
- Supported a variety of question types, from specific events to broader historical themes.

7. Challenges

- Handling ambiguous or overly broad queries.
- Ensuring data consistency during preprocessing and chunking.
- Optimizing the system for large-scale retrieval without significant compute resources.

8. Future Scope

- Expanding the knowledge base to include other U.S. states and historical topics.

- Adding multimodal capabilities like maps or images for enriched answers.
- Incorporating conversational abilities to allow follow-up questions.

9. Key Features and Outputs

- **Performance Metrics:**
    - High relevance in retrievals based on embeddings.
    - Accuracy in answering domain-specific queries.

- **System Outputs:**
    - Concise and contextually grounded answers.
    - Markdown-based data storage for easy updating and expansion.