Gemma Model Document Q&A System

Group Members:

1.Bharat Singh Mehra: 2361095

2.Gunjan Gusain: 2361177

3.Megha Kharayat: 2361341

4.Karan Upadhyay: 2361255

Affiliation: Graphic Era Hill University, Bhimtal

Abstract:

The Gemma Model Document Q&A System is an advanced document-based question-answering (Q&A) framework that utilizes Large Language Models (LLMs), vector embeddings, and semantic search to facilitate precise information retrieval. This system processes PDF documents, extracts meaningful text, converts it into vector representations using FAISS, and employs Llama 3-8B via the GROQ API for generating accurate and context-aware responses. By leveraging retrieval-augmented generation (RAG), it enhances search accuracy, making it highly effective for applications in research, legal documentation, business intelligence, and knowledge discovery.

designed to efficiently retrieve and generate responses from a collection of PDF documents. The system integrates Large Language Models (LLMs), vector embeddings, and semantic search techniques to ensure precise and context-aware answers. Utilizing LangChain, FAISS, and Streamlit, it processes documents by extracting relevant textual data, embedding it for similarity searches, and generating responses using Llama 3-8B via the GROQ API. This system is particularly beneficial for research, legal analysis, and enterprise document management, offering an intelligent approach to retrieval-augmented generation (RAG).

Keywords: NLP, Information Retrieval, Large Language Models, Vector Embeddings, FAISS.

1. Introduction

In an era where vast amounts of unstructured textual data are generated daily, the ability to retrieve and comprehend information efficiently is crucial. Traditional search engines rely on keyword-based retrieval, which often fails to capture contextual meaning. The Gemma Model Document Q&A System leverages semantic search and neural embeddings to offer a superior alternative. It enables users to query a set of documents and receive precise, contextually relevant responses powered by LLMs. This approach significantly enhances knowledge discovery in fields such as law, academia, and business intelligence.

2. System Architecture

2.1. Data Ingestion & Preprocessing

- Document Loading: The system processes PDF files and extracts text using LangChain's PyPDFDirectoryLoader.

- Text Splitting: Documents are divided into overlapping text chunks to improve retrieval efficiency.

- Chunk Optimization: Each chunk maintains coherence to ensure better response quality.

2.2. Vectorization & Storage

- Embeddings Generation: Text chunks are transformed into vector embeddings using Google Generative AI Embeddings.

- FAISS Indexing: The embeddings are stored in FAISS, an efficient vector similarity search library.

- Semantic Search: When a query is made, FAISS retrieves the most relevant document sections based on vector similarity.

2.3. Retrieval & Response Generation

- Retriever Model: The query is matched with the most relevant document chunks.

- Language Model (LLM): The Llama 3-8B model, via GROQ API, processes the retrieved content to generate an intelligent response.

- User Interface: Responses and supporting document excerpts are displayed in an interactive Streamlit UI.

3. Experimental Setup & Evaluation

- The system was evaluated across different document sets, including academic papers, legal texts, and industry reports. Performance was measured based on:

- **Retrieval Accuracy**: How well the system identifies relevant passages.

Response Coherence: The fluency and correctness of generated answers.

Efficiency: Query response time with increasing document volume.

Results indicate that vector-based semantic search outperforms traditional TF-IDF and BM25 retrieval methods, particularly in handling complex queries that require deep contextual understanding.

4. Applications & Use Cases

- Research & Academia: Quickly extract relevant information from research papers.

- Legal Analysis: Retrieve case laws, statutes, and legal precedents efficiently.

- Enterprise Knowledge Management: Improve access to corporate documentation and internal knowledge bases.

- Healthcare & Clinical Documentation: Enable fast retrieval of patient records and medical research insights.

5. Conclusion & Future Work

The Gemma Model Document Q&A System provides a scalable, high-performance solution for extracting meaningful insights from unstructured documents. By integrating semantic search with LLM-powered response generation, it achieves context-aware document retrieval.

Future enhancements include:

- Multimodal Processing: Supporting images, tables, and graphs alongside text.
- Improved Fine-Tuning: Enhancing LLM responses with domain-specific datasets.
- Real-Time Processing: Optimizing retrieval and response times for larger datasets.

This system represents a significant advancement in intelligent document retrieval, bridging the gap between unstructured data and knowledge discovery.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., et al. "Attention is all you need." Advances in Neural Information Processing Systems, 2017.

[2] Johnson, J., Douze, M., Jégou, H. "Billion-scale similarity search with GPUs." IEEE Transactions on Big Data, 2019.

[3] OpenAI. "GPT-4 Technical Report." ArXiv preprint, 2023.

[4] Google AI. "Advances in Generative AI Embeddings." Google Research Blog, 2023.

[5] FAISS: Facebook AI Similarity Search