Full-Time Data Scientist Job Assignment: Multilingual PDF RAG System

Background

Our organization needs to develop a Retrieval-Augmented Generation (RAG) system capable of processing multilingual PDFs, extracting information, and providing summaries and answers to questions based on the content. The system should handle various languages including Hindi, English, Bengali, and Chinese, and be able to process both scanned and digital PDFs.

Input

- PDFs in multiple languages - English, Bengali, Chinese, etc
  (sample pdfs are shared here)
- Mix of scanned and digitally created PDFs

Problem Statement

Develop a RAG pipeline for summarizing content and answering questions based on the input PDFs. The system should be scalable to handle large amounts of data (up to 1TB) and provide accurate, relevant responses.

To-Do List

1. Implement text extraction techniques, including OCR for scanned documents and standard extraction for digital PDFs. The sample pdfs are given with this assignment. Please use them for reference. Adding more pdf samples is recommended, since we also want to see the scalability of the application.
2. Create an advanced RAG system with the following features:
    a. Chat memory functionality
    b. Query Decomposition
    c. Optimized chunking algorithms
    d. Hybrid search combining keyword and semantic search techniques
    e. Integration with high-performance vector databases capable of handling large-scale data.
    f. Selection and implementation of appropriate LLM and Embedding models
    g. Implement reranking algorithms for improved result relevance
    h. Develop metadata filtering capabilities

Evaluation

1. Query Relevance-
   The extent to which search results match the user's query, ensuring that the information provided aligns with user intent.
2. Retrieval test-
   Whether the chunks returned matched with the context of the query.
3. Latency-
   The time it takes for a search engine to return results after a query is submitted, with lower latency contributing to a better user experience.
4. Fluency-
   The clarity and coherence of how information is presented, including readability and organization, facilitating easy navigation and understanding by users.
5. Size of models:
   Much can be achieved using small embedding models and small LLMs nowadays. Even LLMs of parameter size of 2B have high context lengths. We are interested in seeing how reliable a RAG pipeline you can build with as small models as possible.

Deliverables

- Fully functional RAG pipeline meeting all specified requirements (github repo)
- Detailed technical documentation of the system architecture and components (Not more than 3 pages/ 5 slides)
- User guide for operating and maintaining the system
- Performance and evaluation reports demonstrating system capabilities
- Presentation summarizing the project, challenges faced, and future improvement areas

Note: Please do not get overwhelmed by the assignment. Submit whatever you can in 72 hours of receipt. We are interested in knowing how you think, plan and execute data science problems, your implementation speed and your ability to come up with innovative solutions in less time.