Documentation: Llama 2 QA System

Michael Sargious

December 26, 2023

1 Overview

The Llama 2 QA System is a sophisticated application designed to perform advanced question-answering tasks based on academic papers retrieved from the arXiv API. This system leverages various Python libraries and technologies, including Streamlit for the web interface, LangChain for natural language processing, and FAISS for efficient similarity search in large databases. The application's primary goal is to provide users with concise, relevant answers derived from a curated set of academic papers.

2 System Components

2.1 Importing Libraries

Standard Libraries: re, os, json, urllib.request, xml.etree.ElementTree for regular expressions, operating system interactions, JSON operations, URL handling, and XML parsing respectively. External Libraries: streamlit, dotenv, langchain libraries for creating web applications, managing environment variables, and utilizing advanced language models and retrieval techniques.

2.2 Environment Setup

load_dotenv(): Loads environment variables (like API keys) from a .env file for secure access and configuration.

2.3 Core Functions

- fetch_papers: Purpose: Fetch and parse academic papers related to "llama" from the arXiv API.

  Process: Makes an HTTP request to the arXiv API. Parses the XML response to extract paper titles and summaries. Returns a list of paper details.

- save_list_to_json and load_list_from_json: Purpose: Persist and retrieve the list of papers as a JSON file.

  Functionality: These functions save a list of strings to a JSON file and load it back into the application, respectively.

- initialize_qa_chain: Purpose: Initialize the QA (Question Answering) chain using LangChain and FAISS.

  Process: Retrieves the OpenAI API key. Initializes the OpenAI Embeddings model for text analysis. Loads or creates a list of papers, saving it as a JSON file. Sets up FAISS index for efficient similarity search among papers. Configures a retriever with a similarity score threshold. Initializes the ChatOpenAI model for generating answers. Returns a configured QA chain ready for use.

2.4 Streamlit Web Interface

UI Elements A title for the application.

- A title for the application.
- A text input field for users to enter their queries.
- A button to trigger the QA process.

Functional Workflow On clicking 'Get Answer', the app initializes the QA chain. It fetches a response from the QA model based on the user query. The app displays the answer and lists the titles of source papers.

2.5 Displaying Results

The system shows the answer derived from the QA model. Titles of source papers used to generate the answer are displayed, providing transparency and reference material.

3 Conclusion

The Llama 2 QA System is an innovative solution combining the power of AI, natural language processing, and information retrieval to answer user queries effectively. By sourcing information from academic papers, it ensures that the answers are grounded in well-researched content, making it an invaluable tool for researchers, students, and anyone seeking knowledge in the field.