Million Technical AI Engineer Test: Automating Real Estate Document Analysis

The Million Technical AI Engineer Test project offers an efficient solution for extracting, processing, and answering questions based on real estate documents in PDF format. By utilizing the FLAN-T5 model, the system can generate direct answers to specific questions by quoting relevant content from the document. This approach automates the process of extracting key information, making it particularly useful in industries like real estate, where handling large volumes of documents is common.

The solution begins with text extraction, where the PyMuPDF library is used to extract content from the provided PDF files. This ensures that the text is ready for processing. After extraction, the next step is text preprocessing, which ensures the text is clean and suitable for input into the FLAN-T5 model. The preprocessing step removes unnecessary spaces, corrects formatting issues, and prepares the extracted text for better model interpretation.

At the core of the project is Question Answering (QA) using the FLAN-T5 model. This state-of-the-art language model, based on the Transformer architecture, processes the clean text and generates responses to predefined questions. The answers are directly quoted from the document, allowing users to retrieve specific information without manually searching through large documents. Additionally, the system includes functionality to automatically download the pre-trained FLAN-T5 model from Hugging Face, ensuring that users always have the most up-to-date version without the need for manual updates.

The project's workflow is designed for simplicity, making it accessible to users even with limited technical experience. After cloning the repository and installing the required dependencies, users can easily run the project. The system will automatically download the model and process the PDF to generate answers for predefined questions. Users can customize the PDF URL and the list of questions according to their needs, allowing for flexibility and personalization of the results.

The system outputs the answers directly to the console, along with the relevant text quotes from the document. This provides users with easy access to the information they need while ensuring the accuracy and context of the responses.

The project is highly customizable and user-friendly. Users can change the URL of the PDF to process different documents and edit the questions to extract specific information. The model also has a token limit of 2048 tokens, so for larger documents, users may need to truncate or split the text for proper processing. Once the model is downloaded, it is stored locally, preventing the need for repeated downloads in future runs, which improves efficiency.

This project is open-source and licensed under the MIT License, allowing other developers to contribute, modify, and distribute the code. Contributions are welcome, and users can easily fork the repository to make changes or improvements.

In summary, the Million Technical AI Engineer Test project is a powerful tool for automating the extraction of key information from real estate documents. By leveraging the FLAN-T5 model, the system provides fast and accurate answers to specific questions based on the document's content. This solution streamlines the process of document analysis, making it an invaluable asset for industries dealing with large amounts of textual data