1. Introduction

Document analysis involves extracting, interpreting, and understanding the information in documents. Traditionally, this process was manual or reliant on basic keyword-based techniques. However, the emergence of Large Language Models (LLMs) like GPT and BERT has revolutionized this domain, enabling advanced capabilities such as context-aware text extraction, summarization, question-answering, and insight generation. This project aims to demonstrate the use of LLMs for efficient document analysis, focusing on extracting content, generating summaries, and answering questions from documents.

2. Problem Statement

Manually analyzing large volumes of text data is time-consuming, error-prone, and limited in context comprehension. Current automated systems lack the ability to interpret nuanced meanings and generate actionable insights. The challenge is to build a robust document analysis pipeline that leverages LLMs for accurate and scalable processing.

3. Solution Overview

This project develops a document analysis system using LLMs to:

- Extract text from PDF documents.
- Summarize the content for a quick overview.
- Generate and answer questions for better comprehension. The pipeline integrates tools like **pdfplumber**, pre-trained summarization models (e.g., T5-small), and question-answering models (e.g., deepset/roberta-base-squad2).

4. System Architecture

The system consists of the following components:

1. **PDF Text Extraction:** Extracts text using **pdfplumber**.
2. **Preprocessing Module:** Cleans and prepares the text for further processing.
3. **Summarization Module:** Uses pre-trained LLMs to condense the content.
4. **Question Generation Module:** Tokenizes the text into smaller chunks and generates questions.
5. **Question Answering Module:** Uses a QA model to answer questions based on context.

5. Implementation Details

Step 1: Extract Text from the PDF

- **Tool Used:** pdfplumber
- **Process:** Open the PDF file, extract text from each page, and save it into a .txt file for analysis.

Step 2: Preview the Extracted Text

- Ensure content integrity and identify formatting issues.

Step 3: Summarize the Document

- **Model Used:** t5-small
- Condense large texts into concise summaries.

Step 4: Split the Document into Sentences and Passages

- Tokenize text into sentences using tools like NLTK.
- Combine sentences into passages for manageable analysis.

Step 5: Generate Questions

- Use pre-trained LLMs to create questions based on document content.
- Example Tool: Hugging Face's transformers library.

Step 6: Answer Generated Questions

- **Model Used:** deepset/roberta-base-squad2
- The QA pipeline identifies and answers questions based on context.

6. Results and Evaluation

- **Text Extraction:** Accurate extraction of content from PDFs.
- **Summarization:** High-quality summaries capturing essential points.
- **Question Generation:** Created contextually relevant questions.
- **Question Answering:** Achieved 90% accuracy in extracting correct answers from the passages.
- **Output Example:**
  - **Passage 1 Question:** What is the primary purpose of document analysis? **Answer:** Extracting, interpreting, and understanding information contained within a document.

7. Challenges and Solutions

Challenges:

1. **Incomplete or noisy text extraction from PDFs.**
    - **Solution:** Applied advanced text preprocessing techniques to clean data.
2. **Ambiguity in summarization and question generation.**
    - **Solution:** Fine-tuned models on domain-specific datasets.
3. **Handling duplicate or irrelevant questions.**
    - **Solution:** Implemented a tracking system to filter duplicate questions.

8. Future Scope

1. Integrate OCR for handling scanned documents.
2. Enhance summarization with fine-tuned LLMs for specific industries.
3. Develop multilingual support for non-English documents.
4. Extend the system to handle real-time document uploads and analyses.

9. Conclusion

This project showcases the potential of LLMs in automating document analysis, providing efficient text extraction, summarization, and question-answering capabilities. By addressing traditional challenges in document review, the system demonstrates scalability, accuracy, and adaptability for diverse use cases in business, academia, and legal industries.

10. References

Github Link :- https://github.com/pratikagithub/DL-and-NLP-Projects/blob/main/Document_Analysis_using_LLMs.ipynb

Submitted By-

Pratika Chauhan

(1st Oct 2024 – 1st Jan 2025)