REPORT ON

PDF Answering AI

Name : DEEVYANSH DEWANGAN
Enrollment : 21112041
Branch : Chemical Engineering
Year : 4th year

Outline :

- Introduction
- Objectives
- Methodology
- Challenges and solutions
- Future work
- Conclusion

Introduction

The project "Answering AI" aims to develop a system that allows users to query a PDF document and receive accurate and contextually relevant responses. By leveraging natural language processing (NLP) techniques, the system extracts text from a PDF, processes it, and answers user queries based on the extracted content. This report details the development process, methodologies employed, and the final implementation of the Answering AI system.

Objectives

1. **Text Extraction**: Extract text content from PDF documents.
2. **Text Preprocessing**: Clean and preprocess the extracted text for analysis.
3. **Word Embedding**: Utilize word embedding techniques to convert text into numerical vectors.
4. **Question Answering**: Develop a mechanism to answer queries based on the text content.
5. **User Interface**: Create a user-friendly interface for interacting with the system.

Methodology

1. Text Extraction

The text extraction phase involves parsing the PDF to retrieve its text content. This is achieved using the `PyMuPDF` library, which provides robust tools for handling PDF files.

2. Text Preprocessing

Preprocessing involves converting the text to a format suitable for analysis. This includes tokenization, lowercasing, and removing punctuation and numbers.

3. Word Embedding

Word embedding translates text into numerical vectors, capturing semantic relationships between words. `gensim`'s Word2Vec model is employed for this purpose.

4. Question Answering

The core functionality involves answering user queries by identifying the most relevant part of the text. Cosine similarity between the query vector and context vectors determines the best match.

5. User Interface

A simple web interface using Flask allows users to interact with the system, ask questions, and view answers.

Challenges and Solutions

1. **Module Installation Issues**: Errors like `ModuleNotFoundError: No module named 'gensim'` were resolved by updating `pip`, `setuptools`, and `wheel`, and installing dependencies like `Cython` and `numpy`.
2. **Text Quality**: Ensuring the quality of extracted text involved cleaning and preprocessing steps to handle various text formats and remove noise.
3. **Performance**: Balancing accuracy and performance, especially for large PDFs, required optimizing the word embedding and query matching processes.

Future Work

1. **Advanced NLP Techniques**: Integrating more sophisticated NLP models like BERT or GPT for improved understanding and response accuracy.
2. **PDF Handling**: Enhancing the PDF extraction process to handle images, tables, and other non-text elements.

3. **User Interface**: Developing a more interactive and user-friendly interface, possibly with real-time query suggestions and improved visualization of answers.

Conclusion

The Answering AI project demonstrates the feasibility of creating an intelligent system that can understand and respond to queries based on the content of a PDF document. By leveraging NLP techniques, the system provides a useful tool for extracting knowledge from text-heavy documents and can be further enhanced with more advanced technologies and features.