DWIT COLLEGE

DEERWALK INSTITUTE OF TECHNOLOGY

ASKPDF – A PDF CONTENT QUERY SYSTEM

A PROJECT – II PROPOSAL REPORT

Submitted to

Faculties of Humanities and Social Sciences

DWIT College

Submitted by

Pranaya Shrestha

June 4, 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

PDF	Portable Document Format
RAG	Retrieval Augmented Generation

1. INTRODUCTION

In an increasingly digital world, PDF documents have become a standard for sharing and storing information across various domains, including academia, business, and personal use. Despite their widespread adoption, interacting with the content of PDFs remains a challenge, especially when it comes to extracting specific information or answering questions about the document's content. This project, "AskPDF – A PDF Content Query System," aims to bridge this gap by providing a robust platform that allows users to upload PDF documents and ask questions about their content. This system will leverage natural language processing (NLP) and machine learning techniques to understand the context of the PDF files and provide relevant and accurate answers to user queries.

2. PROBLEM STATEMENT

Many users face difficulties when trying to extract specific information from lengthy PDF documents. Traditional methods of searching within PDFs are often time-consuming and inefficient, especially for complex queries that require understanding context or interpreting data. This project addresses the need for a more intuitive and efficient way to interact with PDF content by developing a system that allows users to upload PDFs and ask questions about their content, receiving accurate and relevant answers. The primary objective is to improve the accessibility and usability of information stored in PDFs through advanced natural language processing and document analysis techniques.

3. OBJECTIVE

The objectives of the project are mentioned below.

- To develop an intelligent system that enables users to upload PDF documents and interact with them.

- To enhance the accuracy and relevance of responses by integrating Retrieval-Augmented Generation (RAG) techniques, ensuring comprehensive and precise information retrieval from PDFs.

4. METHODOLOGY

a. Requirement Identification

i. Study of the existing system.

To develop an effective PDF content query system, it's crucial to analyze and understand the capabilities and limitations of existing solutions. This involves evaluating current tools that offer PDF searching, extraction, and querying features to identify gaps and areas for improvement. Existing platforms such as Chat PDF[1], Monica[2], and Soda PDF[3] all perform well in their respective areas, offering robust PDF management and query functionalities. This project aims to develop a similar system with the addition of a text-to-speech function for the generated responses, enhancing accessibility and user interaction.

ii. Requirement collection

After collecting the requirements for the project by observing the existing projects, the requirements collected can be categorized as follows:

a. Functional Requirement

- The user shall be able to upload PDF documents.
- The system shall extract and process the text from the uploaded PDF.
- The user shall be able to ask questions related to the content of the PDF.
- The system shall generate accurate answers based on the PDF content.
- The user shall receive the answers in a user-friendly format.

b. Non-Functional Requirement

- The system must provide quick and accurate responses to user queries.
- The service must be reliable and handle multiple user sessions simultaneously.

- The user interface must be intuitive and easy to navigate.
- The system must ensure the security and privacy of uploaded documents.

b. Feasibility study

i. Technical Feasibility

The project is technically feasible with the availability of document-processing libraries like LangChain. The integration of these technologies can be achieved using Python for backend processing and React.js for the web application. Leveraging robust NLP libraries and tools, we can efficiently handle natural language queries and extract content from PDFs.

ii. Operational Feasibility

The application is also operationally feasible, as it will be designed to be user-friendly, and intuitive, and require little technological knowledge to operate.

iii. Economical Feasibility

All tools used in the development are open-source and freely available. So, this system is also economically feasible to develop.

c. High-level Design System

1.1.1 System Development Model.

The iterative model has been chosen as the system development model for this project. This model emphasizes a cyclic process of prototyping, development, testing, and refinement, allowing for continuous improvement and adaptation throughout the project lifecycle. By breaking down the development process into smaller, manageable iterations, the iterative model enables the developer to address user feedback and incorporate changes incrementally, reducing the risk of costly rework and ensuring that the final product meets user requirements effectively. Each iteration involves the implementation of specific features or enhancements, followed by rigorous testing and evaluation to identify and resolve any issues. This iterative approach fosters flexibility and agility, allowing the developer to respond quickly to changing requirements or emerging challenges while maintaining a focus on delivering high-quality results. Additionally, the iterative model promotes a dynamic and iterative development process that maximizes efficiency and effectiveness.

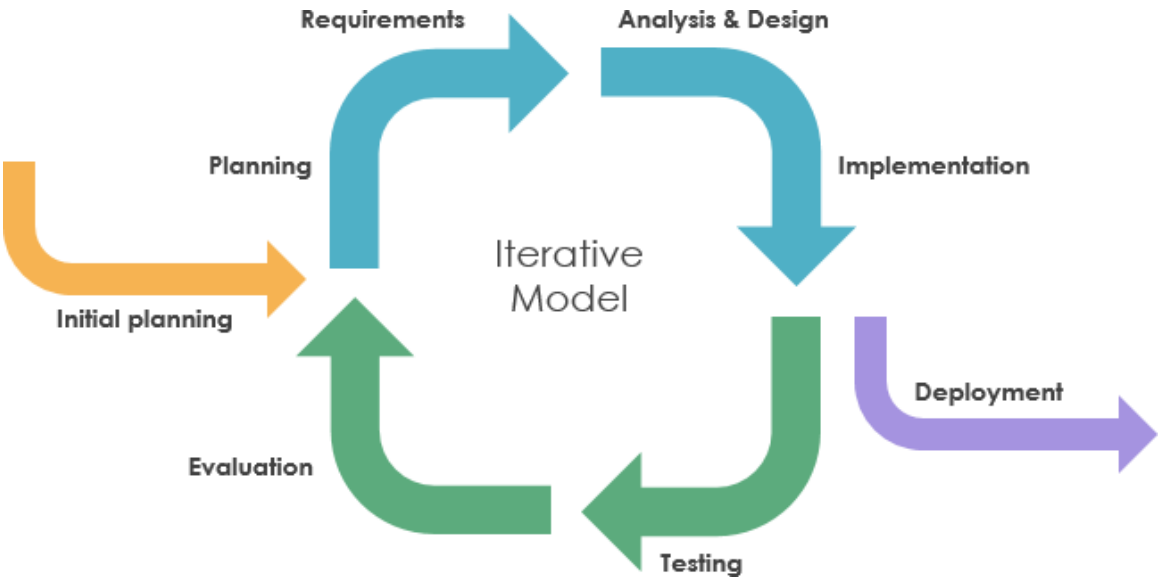

Figure 1: Iterative Model [4]

1.1.2 Flowchart

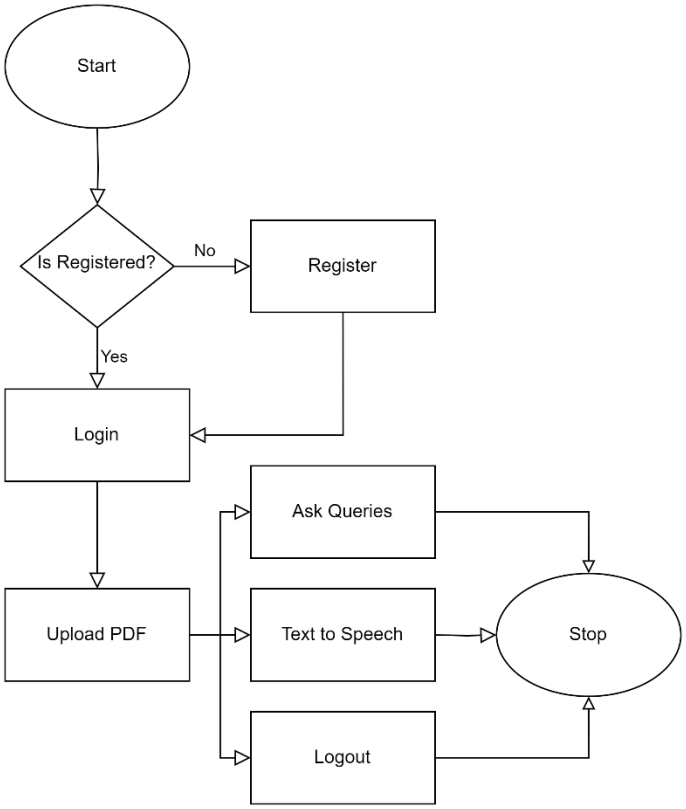

Figure 2: System Flow Chart of AskPDF

1.1.2 Methodology of purpose/ working system

The proposed system will integrate Retrieval-Augmented Generation (RAG) techniques as a central component of its methodology. RAG combines retrieval-based methods for content extraction with generative models for response generation, enabling the system to provide accurate and contextually relevant answers to user queries. By leveraging pre-trained language models and fine-tuning them on domain-specific data, the system will effectively understand the context of PDF documents and generate informative responses. Additionally, RAG facilitates the augmentation of generated responses with relevant content retrieved from the PDFs, ensuring

comprehensive and precise information retrieval. This hybrid approach enhances the system's ability to handle complex queries and improve the overall user experience by delivering more insightful and tailored responses.

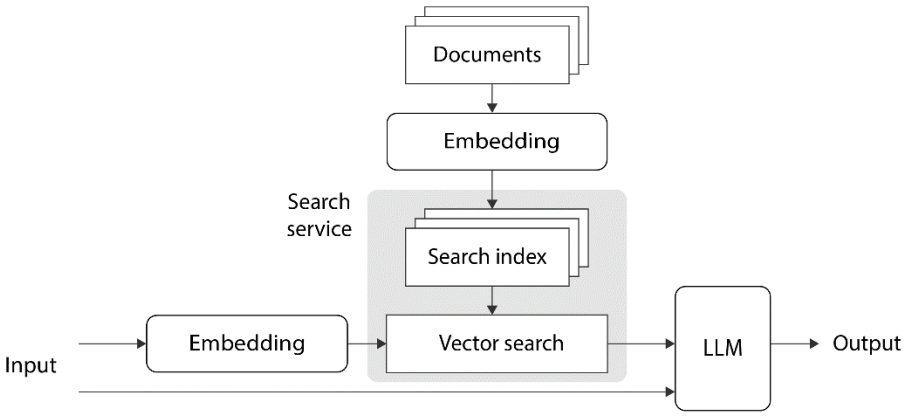

Figure 3: Retrieval Augmented Generation [5]

1.1.2.1 Use Case Diagram

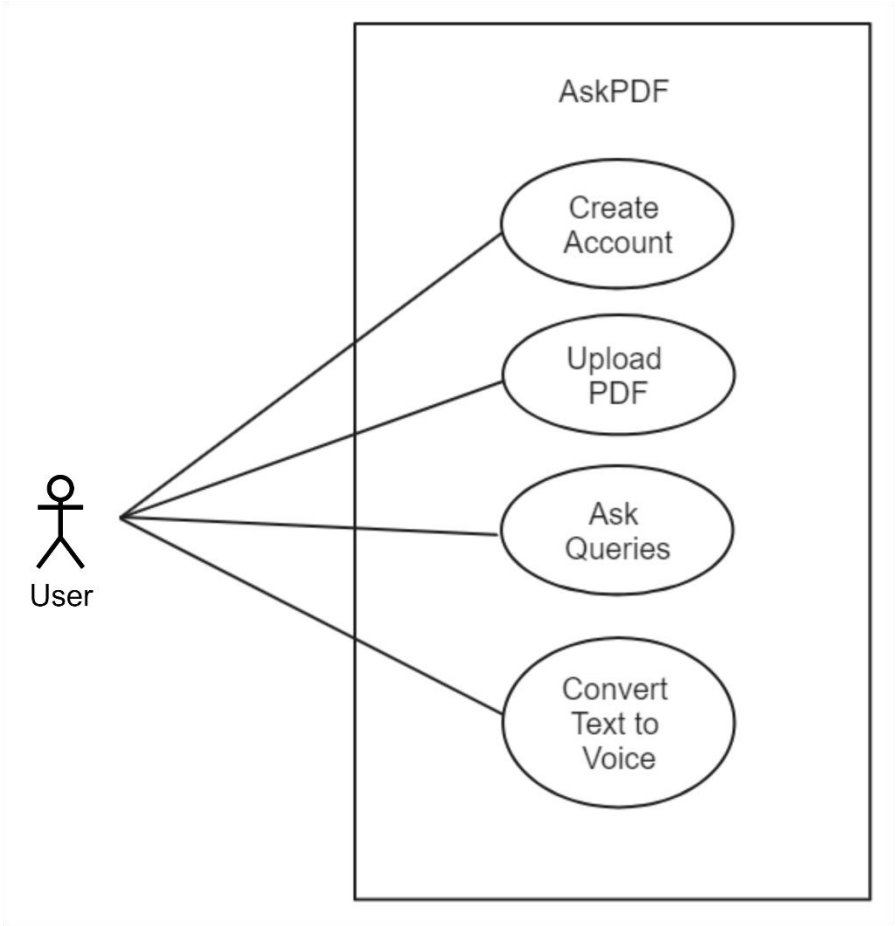

Figure 4: Use Case Diagram of AskPDF

5. GANTT CHART

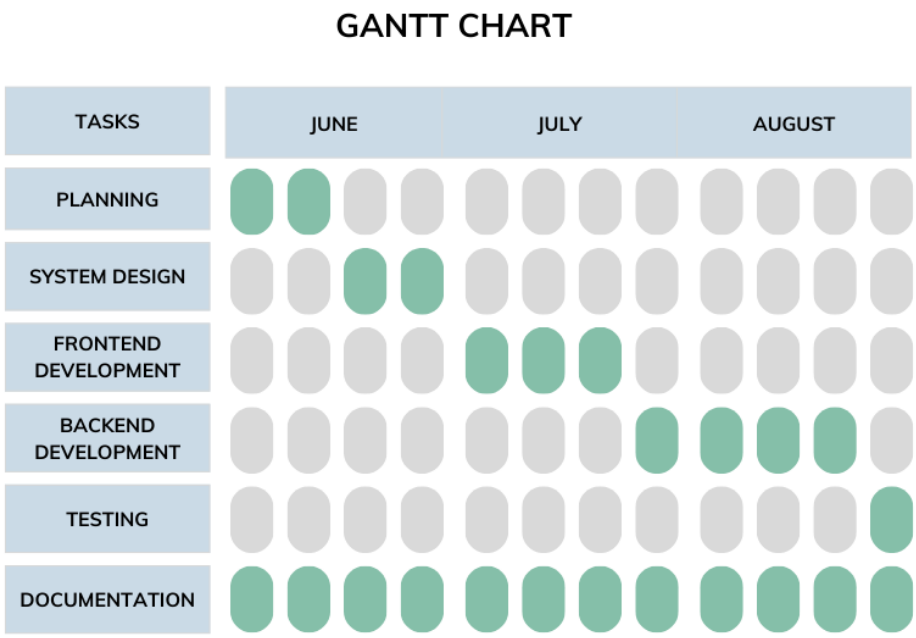

Figure 5: Gantt Chart

6. EXPECTED OUTCOME

The expected outcome of this project is a web application that enables its users to upload PDF documents of any size, ask questions about their contents, and receive accurate and contextually relevant answers.

7. REFERENCES

[1] “Chat pdf AI, a tool to read, summarize and explain a PDF document,” Soda PDF, https://www.sodapdf.com/chatpdf-ai/ (accessed Jun. 2, 2024).

[2] “CHATPDF: Make your PDF as a chatbot, ask ai everything,” ChatPDF | Make your PDF as a Chatbot, ask AI everything, https://monica.im/webapp/doc-chat (accessed Jun. 1, 2024).

[3] “Chat with any PDF,” ChatPDF, https://www.chatpdf.com/ (accessed Jun. 7, 2024).

[4] Jin, “Software development framework - iterative model,” Medium, https://medium.com/geekculture/software-development-framework-iterative-model-68584bfad773 (accessed Jun. 3, 2024).

[5] Bea Stollnitz, “Retrieval-augmented generation (RAG),” Bea Stollnitz, https://bea.stollnitz.com/blog/rag/ (accessed Jun. 2, 2024).