TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PURWANCHAL CAMPUS
DHARAN,SUNSARI

A
MAJOR PROJECT PROPOSAL
ON

" AI-POWERED PERSONAL DOCUMENT ASSISTANT:SIMPLIFIED MANAGEMANT AND RETERIVIAL OF PERSONAL RECORDS "

Submitted by:

Kritika Thapa	[PUR077BCT041]
Prashant Bhattarai	[PUR077BCT059]
Roshan Chaudhary	[PUR077BCT069]
Saurab Baral	[PUR077BCT075]

Submitted to:
Department of Electronics and Computer Engineering
Purwanchal Campus, Dharan

July, 2024

DECLARATION

We hereby declare that the proposal report of the project entitled "**AI-POWERED PERSONAL DOCUMENT ASSISTANT:SIMPLIFIED MANAGEMANT AND RETERIVIAL OF PERSONAL RECORDS**." which is being submitted to the Department of Electronics and Computer Engineering, IOE, Purwanchal Campus, Dharan, in partial fulfilment of the requirements for the award of the Degree of Bachelor of Engineering in Computer Engineering is a bona fide report of the work carried out by us. The materials contained in this report have not been submitted to any university or institution for the award, and we are the only authors of the complete work. Sources other than those listed here have been used in this work.

Kritika Thapa	[PUR077BCT041]
Prashant Bhattarai	[PUR077BCT059]
Roshan Chaudhary	[PUR077BCT069]
Saurab Baral	[PUR077BCT075]

CERTIFICATE OF APPROVAL

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering, IOE, Purwanchal Campus, AI-POWERED PERSONAL DOCUMENT ASSISTANT:SIMPLIFIED MANAGEMANT AND RETERIVIAL OF PERSONAL RECORDS ” submitted by Kritika Thapa, Prashant Bhattarai, Roshan Choudhary and Sourab Baral in partial fulfillment for the award of a Bachelor’s Degree in Computer Engineering. The Project was carried out under special supervision and within the time frame prescribed by the syllabus. We found the students to be hardworking, skilled and ready to undertake any related work to their field of study and hence we recommend the award of partial fulfillment of Bachelor’s degree of Computer Engineering.

.....

Head of Department

Mr. Prabin Songraula

Department of Electronics and Computer Engineering, Purwanchal Campus

Acknowledgments

I express my sincere appreciation to the Head of the Department (HOD), Mr. Pravin Sangroula and the Deputy Head of the Department, Mr. Pukar Karki for their invaluable guidance, insightful suggestions,constructive feedback, and continuous support throughout the preparation of this minor project paper. I am deeply thankful to the esteemed teachers and faculty members whose valuable feedback, encouragement, and scholarly insights have enriched this final report. Their dedication to academic excellence and commitment to fostering knowledge have been a source of inspiration. Their collective guidance have immensely contributed to the refinement and improvement of this project report.

Abstract

The ”AI-Powered Personal Document Assistant” is an innovative system designed to simplify the management and retrieval of personal records using advanced artificial intelligence (AI) and natural language processing (NLP) technologies. Traditional methods of document management often lack efficiency and accessibility, while existing AI solutions may struggle with persistent data retention and retrieval.

This project addresses these challenges by offering a unified platform where users can securely upload various personal documents, including academic records, professional certificates, medical files, financial papers, and legal documents. Key features include Optical Character Recognition (OCR) for text extraction, intelligent categorization and tagging, secure storage, and a robust knowledge base that ensures accurate and prompt retrieval of information.

TABLE OF CONTENTS

List of Figures

1. Introduction

Managing personal documents such as academic records, professional certificates, medical files, financial papers, and legal documents can be a daunting and time-consuming task. Traditional methods of document management often fall short in providing quick and efficient retrieval, while even advanced AI solutions sometimes fail to retain and recall specific information over time.

Our major project, the ”AI-Powered Personal Document Assistant,” aims to revolutionize personal document management by leveraging advanced artificial intelligence and natural language processing (NLP) technologies. This innovative system allows users to upload a multitude of personal files and retrieve specific documents or information using intuitive natural language queries. By integrating features like Optical Character Recognition (OCR) for text extraction, intelligent categorization and tagging, secure storage, and a persistent knowledge base, the system ensures that all personal information is stored securely in one place and can be accessed effortlessly. The ”AI-Powered Personal Document Assistant” offers a comprehensive and user-friendly approach to document management, making it simpler, faster, and more efficient for users to organize and retrieve their important personal records.

2. Problem Statement

There is a paradigm shift from paper-based to paperless management in today's digital era. This poses a lot of challenges in managing the documents digitally. There is a need to ensure the accuracy, security, accessibility, and privacy of the documents. Existing methods of organizing documents fall short and encounter several difficulties, including physical document archiving, version control management, manual data entry, security breach risk, scalability, and ineffective automation. To overcome these challenges, an intelligent document management system (IDMS) is proposed to overcome these difficulties by automatically extracting and managing the documents. The proposed system is created by utilizing the capabilities of AI, ML, and NLP techniques.

Managing personal documents such as academic records, professional certificates, medical files, financial papers, and legal documents is a complex and time-consuming task for individuals. Traditional document management methods often fail to provide quick and efficient retrieval, and even advanced AI solutions can struggle to retain and recall specific information over time. This inefficiency leads to frustration, lost time, and potential risks related to misplaced or inaccessible important documents. There is a pressing need for an intelligent, user-friendly system that can effectively manage and retrieve personal documents with precision and ease.

2.1 Background

2.1.1 Natural Language Processing(NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language. It encompasses a variety of techniques and algorithms designed to process and analyze large amounts of natural language data in order to derive meaning and insights from text. NLP techniques often leverage machine learning and deep learning models to train algorithms on large datasets of human language. These models learn patterns and relationships within the language data, enabling tasks such as language translation, text summarization, and chatbot interactions.

Key aspects of NLP include:

- Tokenization: Breaking down text into smaller units, such as words or phrases (tokens), to facilitate further analysis.
- Part-of-Speech Tagging (POS): Assigning grammatical tags to words based on their role in a sentence (e.g., noun, verb, adjective).
- Named Entity Recognition (NER): Identifying and categorizing named entities (e.g., names of people, organizations, locations) in text.

2.1.2 Optical Character Recognition(OCR)

OCR, or Optical Character Recognition, is a technology that allows computers to convert different types of documents, such as scanned images or PDFs, into editable and searchable text data. It works by detecting text in an image, segmenting it into individual characters, and then recognizing those characters using pattern recognition algorithms. OCR is widely used for digitizing documents, extracting data, enabling text search within documents, and assisting in accessibility for visually impaired individuals by converting text to speech or Braille.

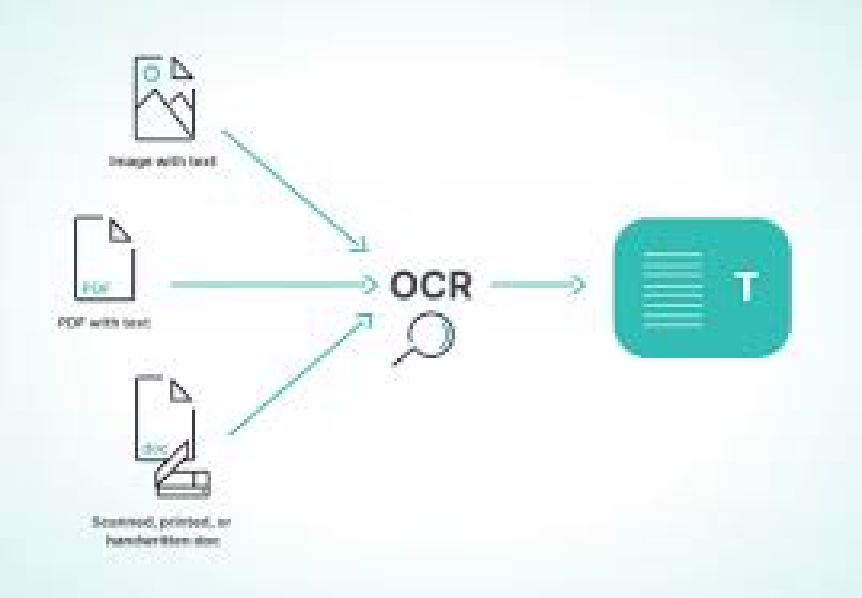

Figure 2.1: Optical Character Recognition

2.1.3 Backend Services

These services include authentication and authorization for secure user access, document management for storing and organizing files, a Natural Language Processing (NLP) service for interpreting user queries, an Optical Character Recognition (OCR) service for extracting text from images, and an integration layer for seamless communication with cloud storage. Together, these backend components ensure that users can upload, search, retrieve, and manage their personal documents effectively while maintaining data security and usability.

2.2 Problem Statements

This project aims to explore the integration of PDF, file management systems, and personalized chat assistants. The primary questions we seek to address are:

- What methods can effectively extract key information (metadata) from documents to enhance search and retrieval?
- What NLP techniques are suitable for interpreting user intent and context from queries re-

lated to document retrieval?

- How can the system ensure efficient and scalable document indexing, retrieval, and management as the volume of documents and users increases?
- What measures should be implemented to ensure the security and privacy of user documents and interactions within the system?

2.3 Objectives

The objectives of this project are as follows:

- Develop a system for uploading, categorizing, and storing various types of personal documents.
- Implement NLP techniques for intuitive natural language queries to retrieve specific documents or information.
- Integrate OCR technology to digitize text from scanned documents and images for easier indexing and retrieval.
- Ensure robust security measures, including encryption and secure access controls, for storing personal documents.

2.4 Scope

The project's scope encompasses the following:

- Advanced NLP Integration: Develop a system utilizing OCR for scanned documents, enabling seamless search across diverse document types including academic records, medical files, and legal documents.
- Intelligent Organization: Implement NLP-driven categorization and tagging to automate document sorting based on content, enhancing retrieval efficiency and user convenience.

- Security and Privacy: Ensure secure storage solutions with encryption and access controls to protect sensitive personal information, adhering to stringent data protection regulations.
- User-Centric Design: Design a user-friendly interface with intuitive navigation, responsive search capabilities, and personalized recommendations, supported by a persistent knowledge base that adapts to user preferences over time.

By focusing on these areas, our project aims to create a comprehensive and robust document management and retrieval system that leverages NLP to enhance accessibility, organization, and security of personal documents.

2.5 Technical Requirements

1. File Handling: Tool: Python with libraries such as Flask for backend support. Description: Allow users to upload documents in various formats (e.g., JPEG, PNG, PDF).
2. Storage Solution: Tool: Cloud-based storage (e.g., AWS S3, Google Cloud Storage,Firebase) or a local database. Description: Securely store uploaded documents.
3. Metadata Management: Tool: NodeJS, MySQL or MongoDB. Description: Store metadata about each document such as file type, upload date, and associated tags.
4. NLP Libraries: Tool: spaCy, NLTK, or transformer-based models (e.g., Hugging Face Transformers with BERT or GPT). Description: Understand and process user queries in natural language.
5. Optical Character Recognition (OCR): Tool: Tesseract OCR, Google Cloud Vision API. Description: Extract text from images and PDFs. Text Parsing:
6. Front-End Development:

   Tool: React.js, Angular. Description: Create a user-friendly interface for document upload and query input.
7. Back-End Development:

   Tool: Flask or Django for Python, Node.js for JavaScript. Description: Handle server-side logic, including file handling, query processing, and database interactions.

8. Integration:

   Tool: RESTful APIs Description: Ensure seamless communication between front-end and back-end components.

3. Literature Review

A.Silval, MM.Gomes., makes use of Natural Language User Interfaces (NLUI) to interact with users and provide information about the weather, maps, schedule, call, events, etc . NLUI allow Human-to-Machine (H2M), Human-Device Interaction (HDI) and Human-Computer Interaction (HCI), which involves the translation of human intention into devices' control commands through speech recognition play.[1] [2]

N. Zierau, C. Engel, M. Söllner, J. M. Leimeister., identifies several key challenges in building trust in SPAs, including the opaqueness and complexity of AI systems, security and privacy concerns, and embedded biases. [3]

X. Ling, M. Gao, D. Wang., presents an approach to intelligent document processing that integrates Robotic Process Automation (RPA) with machine learning and explore how RPA can automate repetitive tasks in document processing, while machine learning enhances the system's ability to understand and categorize documents accurately.[4]

4. Proposed Methodology

4.1 Model Diagram

Figure 4.1: Agile SD Model

In the ”AI-POWERED PERSONAL DOCUMENT ASSISTANT”, the Agile development model can be implemented in this project by organizing work into iterative cycles, or sprints, each focusing on a specific set of tasks such as data collection, OCR integration, data preparation, and storage. Regular stand-up meetings and sprint reviews will facilitate continuous feedback and collaboration among team members, ensuring that the system evolves based on user requirements and feedback. This iterative approach allows for rapid development, testing, and refinement of features, enabling

the project to adapt to changing needs and incorporate improvements quickly, ultimately enhancing the efficiency and effectiveness of the document management and retrieval system

4.2 System Design

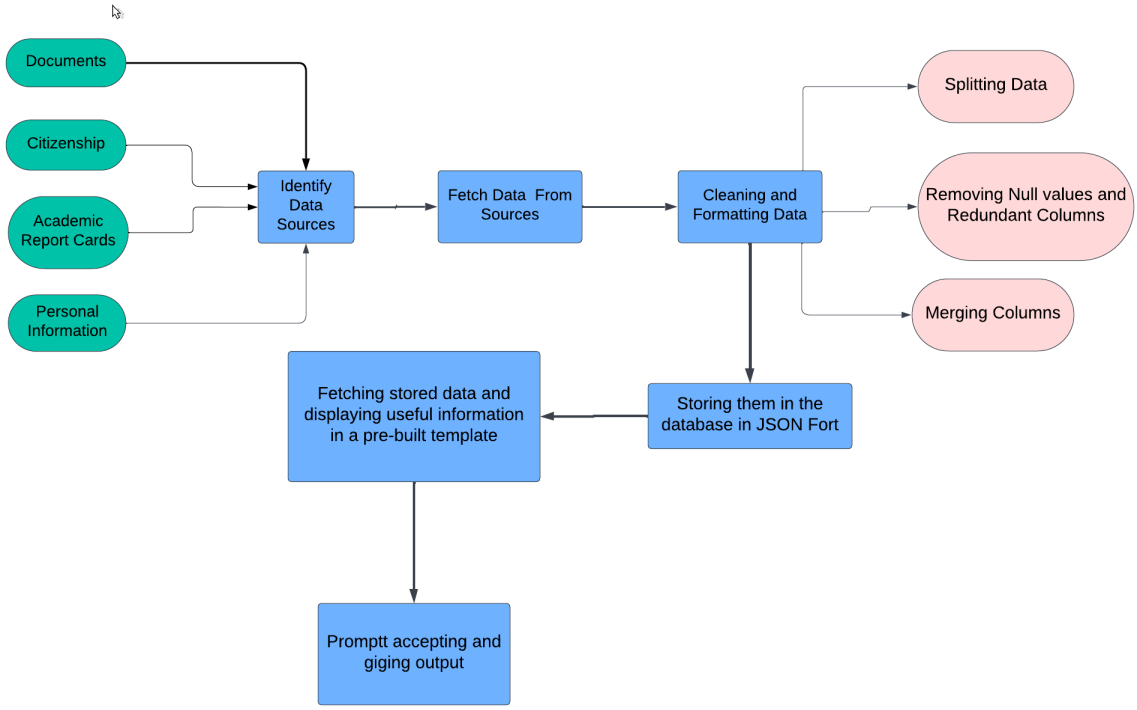

Figure 4.2: System Flow Diagram

This flowchart outlines the methodology for the Personal Document Management and Retrieval System using Natural Language Processing (NLP) within your project. Here's a step-by-step description based on the flowchart:

- Identify Data Sources: The process begins by identifying various data sources, including documents, citizenship records, academic report cards, and personal information.

- Fetch Data From Sources: Once the data sources are identified, data is fetched from these sources. This could involve accessing academic institutions, government databases, or personal repositories.

- Cleaning and Formatting Data:

  The fetched data undergoes a cleaning and formatting process, which includes: Splitting Data: Separating data into relevant categories or fields. Removing Null Values and Redundant Columns: Cleaning the data by removing unnecessary information and handling missing values. Merging Columns: Combining relevant data fields to create comprehensive records.

- Storing Data in JSON Format:

  After cleaning and formatting, the data is converted into key-value pairs and stored in a database in JSON format. This structured storage facilitates quick and efficient data retrieval.

- Fetching Stored Data:

  The system can fetch the stored data and display useful information in a pre-built template. This step involves retrieving the necessary data based on user queries or predefined criteria.

- Prompt Accepting and Giving Output:

  Finally, the system accepts user prompts and provides the output based on the retrieved and processed data. This output could be in the form of search results, document summaries, or other relevant information.

In summary, the flowchart represents a systematic approach to managing personal documents by identifying data sources, fetching and cleaning data, storing it efficiently, and providing user-friendly access to the information through NLP-driven processes.

4.3 Gantt Chart:

GANTT CHART

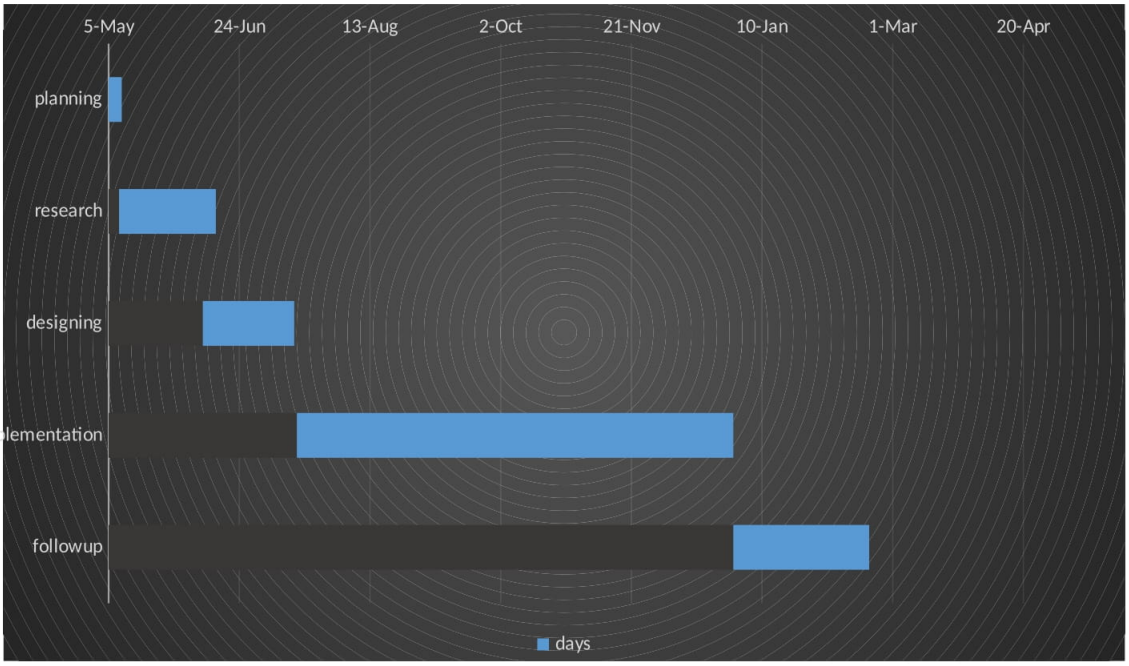

Figure 4.3: Gantt Chart

4.4 Applications

1. **Document Management**: Facilitate the organization, retrieval, and management of academic records and report cards for students, educators, and administrators, making it easier to access and review educational histories.

2. **Personal Records Organization**: Help individuals efficiently manage and retrieve personal documents such as citizenship papers, medical records, and financial documents, providing a centralized and secure repository for important information.

3. **Professional Certification Storage**: Enable professionals to store, categorize, and quickly retrieve certificates and credentials, streamlining the process of document submission for job applications or further education.

4. **Healthcare Record Keeping**: Assist patients and healthcare providers in managing and accessing medical files and history, ensuring that important health information is easily available when needed for consultations or emergencies.

4.5 Conclusion:

The Personal Document Management and Retrieval System using NLP effectively enhances the organization, secure storage, and intuitive retrieval of various personal documents. By employing advanced NLP and OCR, the system converts diverse document formats into searchable text, facilitating the management of academic records, citizenship papers, professional certificates, medical files, and legal documents. Key features like intelligent categorization, secure storage, and a user-friendly interface with an adaptive knowledge base significantly improve user experience. This project addresses current document management challenges and lays the groundwork for future enhancements, making it a valuable tool for efficiently managing personal documents.

References

[1] A. Silva, M. Gomes, C. André da Costa, R. Righi, J. Barbosa, G. Pessin, G. Doncker, and G. Federizzi, "Intelligent personal assistants: A systematic literature review," *Expert Systems with Applications*, vol. 147, p. 113193, 06 2020.

[2] N. Zierau, C. Engel, M. Söllner, and J. M. Leimeister, "Trust in smart personal assistants: A systematic literature review and development of a research agenda," 03 2020, pp. 99–114.

[3] F. Rusli, K. Adhiguna, and H. Irawan, "Indonesian id card extractor using optical character recognition and natural language post-processing," 12 2020.

[4] X. Ling, M. Gao, and D. Wang, "Intelligent document processing based on rpa and machine learning," pp. 1349–1353, 2020.