Extraction Of User Defined Information from Pdf

Final Year Project

PROPOSAL

****

**Supervisor**

Tahir Iqbal

**Submitted by (BSCS-S24-017)**

|  |  |  |
| --- | --- | --- |
| 03-134202-102 |  | Rohaan Nadeem |
| 03-134211-033 |  | Noor Fatima |

**Bachelor of Science in Computer Science**

**Department of Computer Sciences**

Bahria University Lahore Campus

February 2024

Abstract

Problem and Significance: Extracting specific information from PDF documents, including textual content and graphical data, poses a significant challenge due to the diverse formats and structures of these files. This project addresses the need for an efficient tool capable of extracting user-defined information from PDFs and images without requiring manual reading. The significance lies in providing users with a streamlined method to obtain targeted information from PDFs, enhancing research, analysis, and information retrieval processes across various domains.

Method/Tool/Technology and Solution: Leveraging natural language processing techniques, including large language models, combined with large language models, the proposed tool automates the extraction process. By dividing the text into manageable chunks and integrating user prompts, the tool interfaces with large language models to extract relevant variables. Utilizing tools such as Jupiter Notebook for development, React for front-end interface, and Visual Studio Code for development environment enhances the efficiency and user-friendliness of the tool. The solution aims to offer users a convenient means of obtaining desired information from PDFs and images efficiently, thereby reducing the time and effort required for manual extraction and analysis.

Table of Contents

[1. Introduction 1](#_Toc159877523)

[2. Problem Description 1](#_Toc159877524)

[2.1. Primary Scope 1](#_Toc159877525)

[2.2. Final Deliverable of the Project and Beneficiaries 1](#_Toc159877526)

[2.3. Optional Scope 1](#_Toc159877527)

[2.4. Objectives 1](#_Toc159877528)

[2.5. Novelty 1](#_Toc159877529)

[3. Methodology 2](#_Toc159877530)

[4. Feasibility Plan 2](#_Toc159877531)

[4.1. Resource Requirement 2](#_Toc159877532)

[4.1.1. Expertise of the Team 2](#_Toc159877533)

[4.1.2. Tools / Technology 3](#_Toc159877534)

[4.1.3. Budget 3](#_Toc159877535)

[4.2. Risks Involved 3](#_Toc159877536)

[5. Key Milestones and Schedule 4](#_Toc159877537)

[5.1. Key Milestones 5](#_Toc159877538)

[5.2. Gantt Chart 6](#_Toc159877539)

# Introduction

The proposed project aims to develop a tool named "Extraction of User Defined Information from PDF" to automate the process of extracting specific information from multiple PDF documents and images. This website will facilitate users in efficiently obtaining targeted information without the need for manual reading, benefiting researchers, analysts, and professionals across various domains.

# Problem Description

## Primary Scope

The primary scope of this project is to develop a website capable of extracting user-defined information from multiple PDF documents and images. The project addresses the challenges associated with manually extracting relevant information from diverse PDF formats and graphical content. By automating this process, the website aims to streamline information retrieval tasks for users.

## Final Deliverable of the Project and Beneficiaries

The final deliverables of this project encompass a web-based application meticulously crafted to streamline the extraction of information from multiple PDF documents. This application boasts user-friendly interfaces, facilitating seamless input of variables and ensuring the extracted information is presented clearly and comprehensively.

The beneficiaries of this innovative tool are diverse, including students seeking efficient information extraction solutions, researchers engaged in comprehensive data analysis, data analysts navigating complex datasets, and healthcare professionals requiring swift access to pertinent information within PDF documents.

## Optional Scope

1. Implementation of user accounts for personalized experiences and data storage.

2. Exploration options for storing and retrieving previously extracted information for faster responses and inexpensive information extraction.

## Objectives

1. Create a user-friendly interface that allows users to specify variables for extracting information from multiple PDF files.
2. Present the meta-analysis of multiple PDF documents.

## Novelty

1. There is no similar FYP already approved in BULC.
2. Introduction of meta-analysis for enhanced output presentation including text, tabular, and graphical formats.

# Methodology

**[1/3] Pre-Development Stages:**

1. Research on text extraction libraries
   * PyMuPDF
2. Exploration of NLTK for text preprocessing
3. Investigation of image/table recognition algorithms

**[2/3] Data Collection and Preparation:**

1. Acquiring PDF documents for testing purposes
2. Source: PubMed database - accessible at <https://pubmed.ncbi.nlm.nih.gov/>
3. Conducting PICO (Population, Intervention, Comparison, Outcome) analysis
4. Performing data cleaning and organization on the gathered documents
5. Extracting relevant metadata from PDFs for further analysis

**[3/3] Development Steps/Phases:**

1. Implementing text extraction using PyMuPDF
2. Integrating NLTK for text preprocessing
3. Developing image recognition algorithms for graphical content extraction
4. Incorporating LLMs for user defined information extraction
5. Presents the meta-analysis.

**A diagram of a process

Description automatically generated**

# Feasibility Plan

## Resource Requirement

### Expertise of the Team

The team possesses the necessary knowledge and skills in natural language processing, image recognition, and software development required for the successful completion of the project. Relevant courses have been studied, and the project is of equal interest to all team members.

### Tools / Technology

The project relies on key Python libraries:

1. PyMuPDF for PDF manipulation and text extraction from both text and images within PDFs,
2. NLTK for natural language processing tasks.

These libraries, combined with image recognition algorithms, form the backbone of the project's functionality. Additionally, development tools like Jupyter Notebook / Google Colab, React, and Visual Studio Code facilitate the development process, offering accessibility and efficiency for project implementation.

The following hardware and software tools/technologies will be utilized for the project:

**Hardware:**

* Standard laptops/desktops for development and testing.

**Software:**

* Python programming language for backend development.
* PyMuPDF library for PDF processing.
* React for front-end development.
* Text editor or Integrated Development Environment (IDE) for coding (e.g., Visual Studio Code, PyCharm).
* Web browsers for testing and debugging (e.g., Google Chrome, Mozilla Firefox).

### Budget

The budget includes expenses for large language models usage and domain hosting. Other tools and technologies are primarily open-source and do not incur significant costs.

## Risks Involved

The proposed project involves the risk of technical challenges in implementing image recognition algorithms and integrating large language models for user interaction. However, the team is prepared to address these challenges through research and collaboration with experts in the field.

# Key Milestones and Schedule

1. **First Semester:**

**Idea Selection & Project Initiation (1-5 Weeks)**

* Selecting Idea and approving from FYP Supervisor.

**Data Collection & Analysis (6-11 Weeks)**

* Acquiring PDF documents for testing purposes
* Source: PubMed database - accessible at <https://pubmed.ncbi.nlm.nih.gov/>

**Designing and Prototyping (10-13 Weeks):**

* Both of us will leverage Canva for prototype designing, as we possess expertise and familiarity with the platform.

**Development of Text Extraction Algorithms (Month 1-2):**

* Will use PyMuPDF for extracting text from PDF documents and images.
* We'll verify the presence of user-provided variables while designing the prompt in the output to ensure accuracy.

**Implementation of Natural Language Processing (NLP) Techniques (Month 3-4):**

* Integrate NLP techniques, such as NLTK for text preprocessing.
* Develop methods to handle various text formats and languages.

**Initial Integration of large language models (4 Weeks):**

* Begin integration of large language models for user interaction and information extraction.
* Set up initial prompts and responses for testing purposes.

**Further Integration and Refinement of large language models (4 Weeks):**

* Continue integrating large language models for enhanced user interaction.
* Refine prompts and responses based on user feedback.

1. **Second Semester:**

**Development of user interface (4 Weeks):**

* Design and develop a user-friendly interface using React.
* Implement features for input and output interaction.

**Testing, Feedback, and Optimization (4 Weeks):**

* Conduct extensive testing of the tool's functionality and performance.
* Gather user feedback and make necessary optimizations and improvements.

**Finalization and Deployment (4 Weeks):**

* Finalize the tool based on user feedback and testing results.
* Release the tool for public use.
* Finalize documentation and prepare for project submission.

## Key Milestones

Following are the major milestones of the project and their related information.

**Table 5‑1: Breakdown of work in form of milestones**

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | S.No. of Predecessor Milestone | Key Milestone Name / Description | Duration (person-hours[[1]](#footnote-1)) |
| 1 | - | Idea Selection & Project Initiation | 30 |
| 2 | 1 | Data Collection & Analysis | 20 |
| 3 | 2 | Designing and Prototyping | 50 |
| 4 | 3 | Implementation of PDF processing algorithms | 30 |
| 5 | 4 | Integration of information extraction features | 50 |
| 6 | 5 | Development of user interface | 40 |
| 7 | 6 | Testing and refinement | 40 |
| 8 | 7 | Finalization and submission | 28 |

Total: 288

## Gantt Chart

Following is the timeline divided into milestones mentioned (A sample Gantt chart; you may attach a Gantt. Chart produced by a CASE tool)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **Weeks** | | | | | | | | | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Idea Selection & Project Initiation |  |  |  |  |  |  |  |  | Midterm Exam Week |  |  |  |  |  |  |  |  | Exam Week |
| Data Collection & Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Designing and Prototyping |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementation of PDF processing algorithms |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Figure 5‑1: SEMESTER 1 (Spring 2024)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **Weeks** | | | | | | | | | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Integration of information extraction features |  |  |  |  |  |  |  |  | Mid Exam Week |  |  |  |  |  |  |  |  | Exam Week |
| Development of user interface |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Testing and refinement |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Finalization and submission |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Figure 5‑2: SEMESTER 2 (Fall 2024)**

1. [↑](#footnote-ref-1)