Lab 5 Part 1

Team name: Cleanup Crew
Team members: Jonathan Kusuma, James Susanto

Introduction

This lab marks the commencement of a two-part assignment aimed at enhancing our understanding and skills in PDF text extraction, web scraping, data preprocessing, and visualization. Focusing on oil wells' data, we aim to collect, organize, and visualize information, laying the groundwork for a comprehensive web interface in the subsequent part.

Requirements

- Python 3.x
- Libraries: OCRMYPDF, PyPDF2, PyTesseract, requests, selenium, beautifulsoup4, mysql-connector-python, numpy, pdf2image, cv2
- MySQL Server
- Linux Environment

Installation

Ensure Python 3.x and MySQL Server are installed. Install the necessary libraries using pip:

1. pip install pytesseract PyPDF2 pdf2image requests selenium beautifulsoup4 mysql-connector-python numpy opencv-python-headless

For OCR functionality, install Tesseract-OCR:

1. sudo apt-get install tesseract-ocr tesseract-ocr-eng

Running the Application

1. Start the MySQL Server.
2. Adjust database configuration in both ocr.py and additionalinfo.py scripts to match your setup.
3. Execute the OCR script to extract data from PDFs and store it in the database:
    a. python3 ocr.py
4. Run the web scraping script to fetch additional information and update the database:
    a. python3 additionalinfo.py

Notes

- The ocr.py script processes PDF files to extract relevant information using OCR when necessary and stores it in a MySQL database.
- The additionalinfo.py script fetches additional details from the web using the extracted API numbers and well names, appending this information to the existing database records.
- **Warning:** The program is resource-intensive. Running with 5 max workers for OCR processing can consume approximately 23GB of RAM due to multithreading. Adjust the number of workers based on available system resources.