Team name: HealthEdBot

Names & USC ID

Cici Chang (8706354957)

Daoyang Li (7168949829)

Yifan Yang (8386626967)

https://github.com/DSCI560

README for Lab5 Part1

This document provides a detailed overview of a Python script designed for data processing, PDF text extraction, web scraping, and visualization. The script utilizes several popular Python libraries to accomplish tasks such as data cleaning, collect and organize data from PDFs and create web interface to visualize the collected information. The Python script scrape oil wells information from pdfs and drillingedge website, preprocess the data, and save to MySQL database.

Imported Libraries

- re: Regular expression operations for string searching and manipulation.
- fitz: Used for reading, writing, rendering, and manipulating PDFs and other documents.
- PyPDF2: Split, merge, and transform PDF pages, as well as extract text from PDFs.
- pytesseract: An OCR tool that converts images of text into strings of text, using the Tesseract-OCR engine.
- pdf2image: Converts PDF documents into images, allowing each page of the PDF to be turned into an image file.
- mysql.connector: Allow Python to connect to MySQL databases, execute queries, and handle database operations.
- requests: Used for making HTTP requests to web servers, which is useful for downloading web pages or consuming web APIs.
- Beautifulsoup: Used for Parsing HTML and XML documents, making it easier to scrape, parse, and manipulate data from web pages.

Main Parts

OilWells_PDFExtraction.py

The provided code is a Python script designed to connect to a MySQL database, extract text from PDF files, parse specific information from the extracted text, and save the information to both the MySQL database and a CSV file (oil_wells_pdf_data.csv).

OilWells_WebScraping.py

The script is designed to read API numbers from the database or the CSV file, scrape well details for each API number from a website, and save the scraped data into a new CSV file. This process involves network requests to fetch web pages and text processing to extract relevant data. (oil_wells_details_scraped.csv).

User Interaction

The script allows for user interaction through a command-line interface.

Team name: HealthEdBot
Names & USC ID
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang (8386626967)
https://github.com/DSCI560