DSCI-560 Assignment No. 4 – Part 1

Team Formation

Team name: HealthEdBot

Names & USC ID

Cici Chang (8706354957)

Daoyang Li (7168949829)

Yifan Yang (8386626967)

https://github.com/DSCI560

Minutes of the offline meeting on February 19

Participants: project team members

Conference topic: Initial Setup and Planning

Ensure all required software and libraries and installed and functional. Test each tool with simple scripts to confirm they're ready for use. Design the database schema focusing on strong well-specific information and stimulation data. Create the tables in MySQL. Plan the Python script structure for PDF extraction and web scraping, detailing functions, and libraries to be used.

Minutes of the online meeting on February 20

Participants: project team members

Conference topic: Data Collection and PDF Extraction Setup

Download the PDF folder from the provided Google Drive link and organize the files locally for processing. Develop a Python script to iterate over PDF files, using PyPDF2 for text-based PDFs and PyTesseract in conjunction with OCRMYPDF for scanned image PDFs. Extract preliminary data form a couple of PDFs to test the script's effectiveness and adjust code as necessary.

Minutes of the offline meeting on February 21

Participants: project team members

Conference topic: PDF Data Extraction and Initial Database Storage

Run the Python script to extract data from all PDFs. Monitor for any errors or issues with specific files and adjust the script as needed. Begin inserting extracted data into the database. Ensure data integrity and correct any anomalies encountered during insertion. Review the database entries for completeness and accuracy. Adjust the extraction script if required.

Minutes of the online meeting on February 22

Participants: project team members

Conference topic: Web Scraping for Additional Well Information

Develop a Python script using requests/selenium and beautifulsoup4 to scrape well information from drillingedge.com based on API# and well name. Test the web scraping script with a few database entries to refine the search queries and information extraction process. Run the script to scrape data for all entries in the database. Monitor for any blocks or issues and adjust the script as necessary.

Minutes of the online meeting on February 23

Participants: project team members

Conference topic: Data Preprocessing and Finalization

Preprocess the collected data to remove HTML tags, special characters, and irrelevant information. Convert timestamps and replace missing data with 0 or N/A. Update the database with preprocessed web-scraped data. Perform a final check on the database for consistency and completeness. Compile documentation detailing the project setup, scripts, database schema, and any challenges faced. Ensure the documentation includes steps for running the scripts and replicating the project.

Team name: HealthEdBot
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang (8386626967)