**DSCI-560: Data Science Practicum Laboratory Assignment 6**

**Team Name: Guardians of the Algorithm**

**Names:** Jaival Chintankumar Upadhyay

Pratham Solanki

Mayank Patil

Github link: https://github.com/jaivalupadhyay/dsci-560/tree/main/Lab-6

This lab focuses on text extraction, web scraping, data preprocessing, and visualization from scanned PDF files. You assigned tasks include writing and running scripts that efficiently and effectively collect and organize data from PDFs and create a web interface to visualize the collected information. You will work with your team on pdf text extraction in this lab. Additionally, you'll preprocess the data to remove missing values, fetch additional data from the web, and store them in a database.

**Project Workflow**

**1. PDF Text Extraction**

* Script reads multiple PDF files and extracts relevant data.
* Extracted data is stored in a CSV file.

**2. Storing Data in MySQL Database**

* A database Lab6\_database is created.
* Table oilwell\_data stores extracted well information.
* Data is inserted into MySQL for structured access.

**3. Web Scraping for Additional Data**

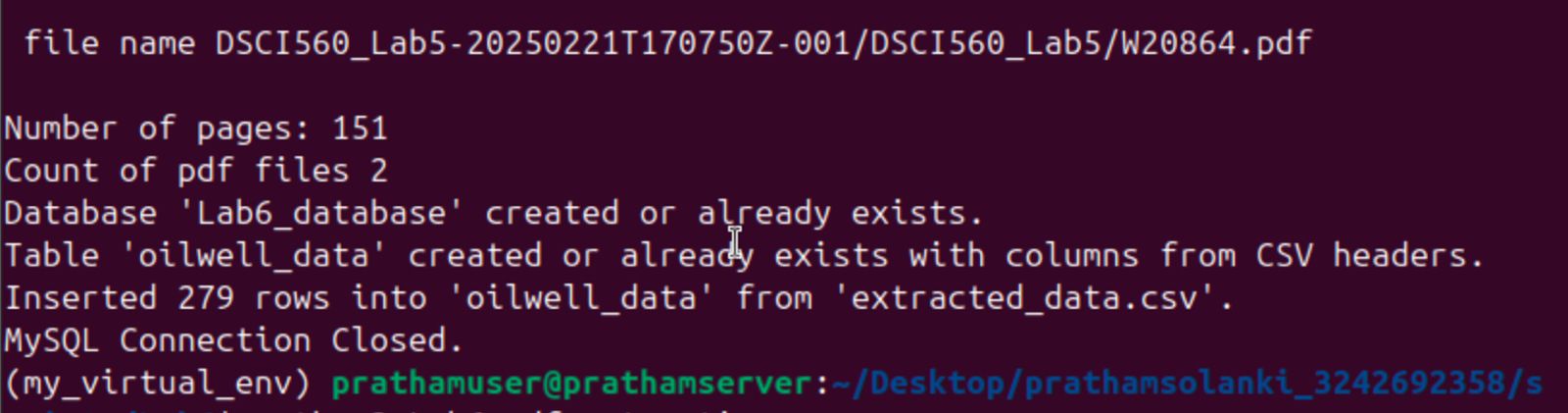
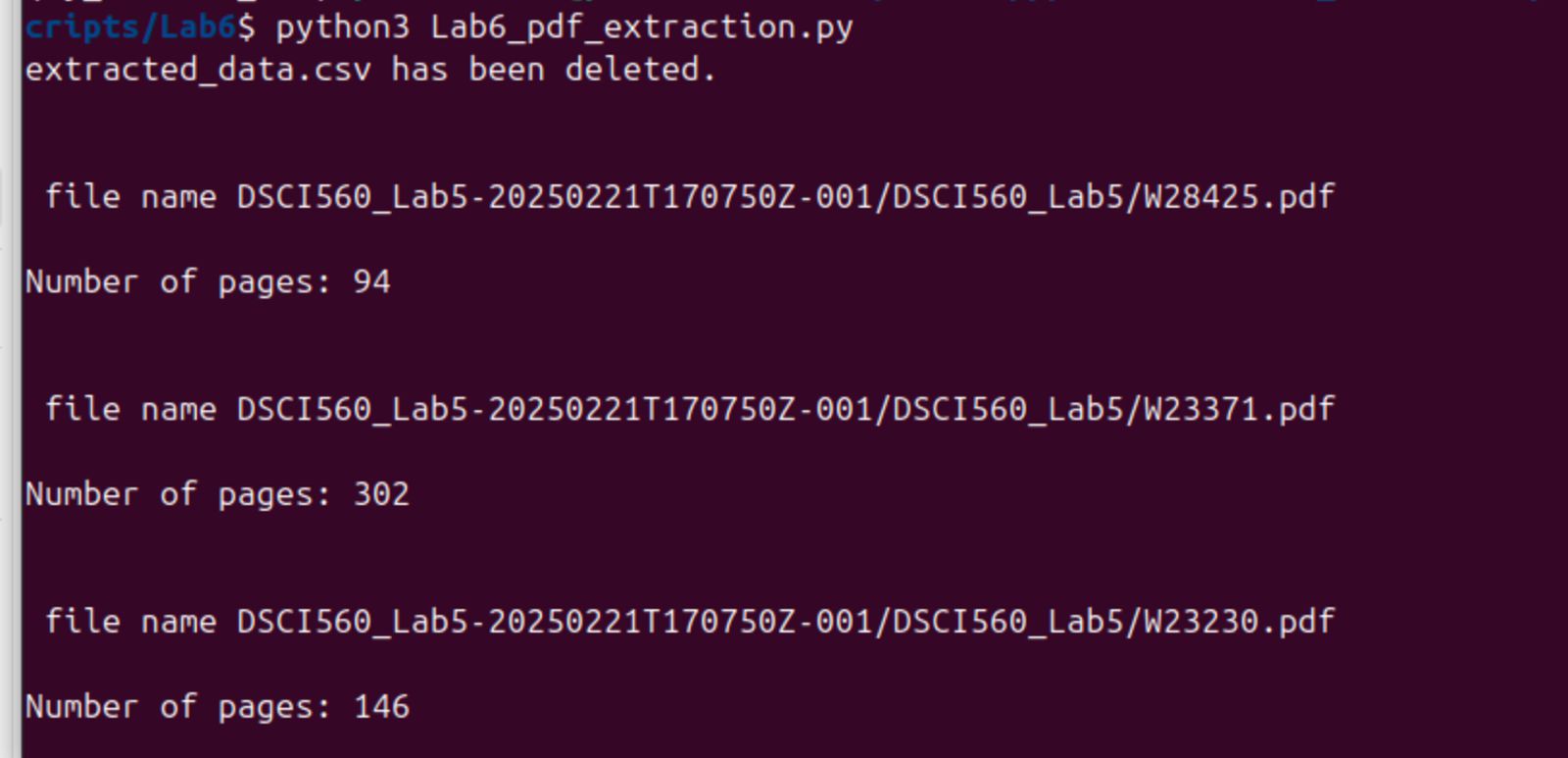
* API numbers from extracted data are used to fetch details from drillingedge.com.
* Data is collected and stored in structured format.

**4. Data Preprocessing**

* Removal of missing values.
* Formatting and structuring extracted information.

**Screenshots and Output**

**PDF Extraction and Data Processing Output**

****

**A screenshot of a computer

AI-generated content may be incorrect.**

Web scraping code

A screenshot of a computer

AI-generated content may be incorrect.

**Web Scraping Script Execution**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

**Processed and Scraped Data csv File**

A screenshot of a computer

AI-generated content may be incorrect.

Github Commit History: A screenshot of a video game

AI-generated content may be incorrect.