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

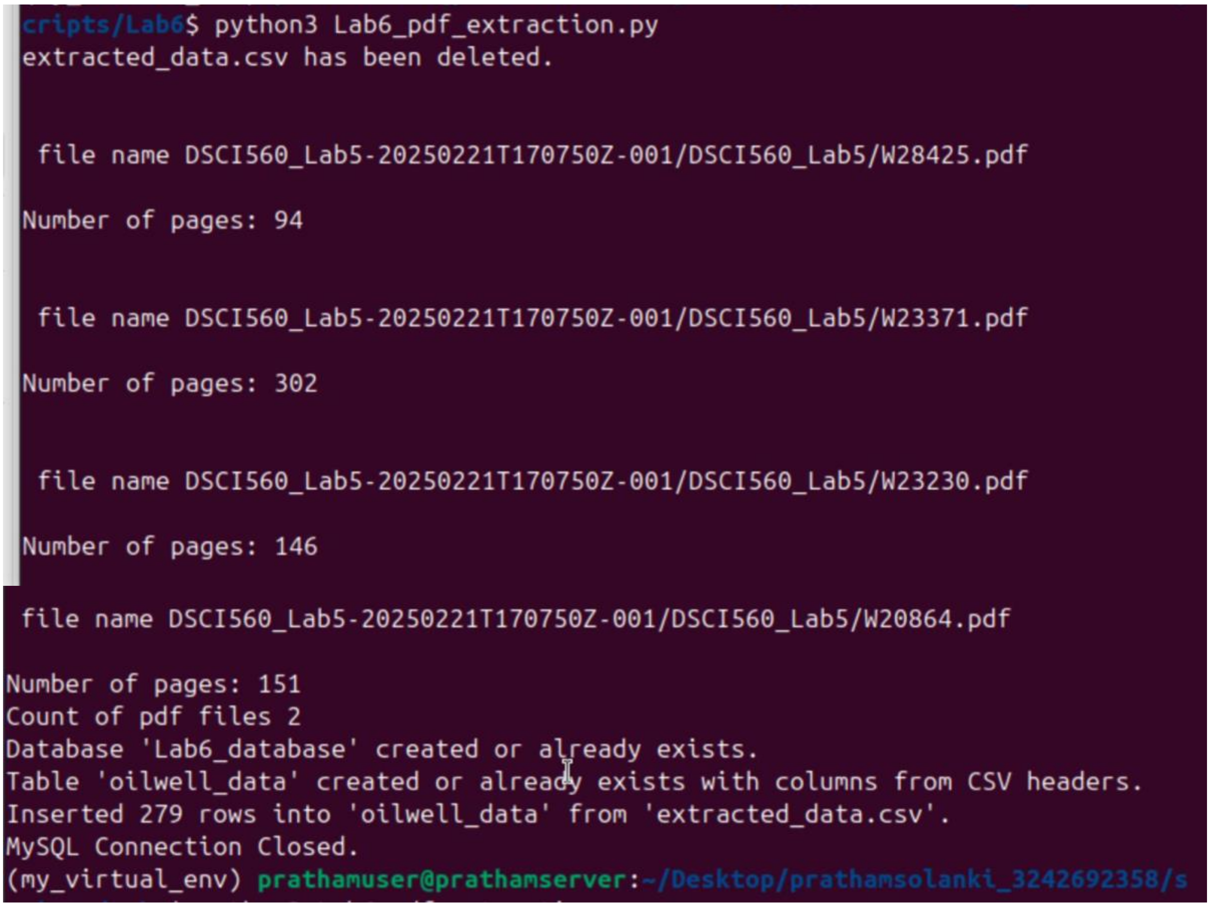
- API numbers from extracted data are used to fetch details from `drillingedge.com`.
- Data is collected and stored in structured format.

4. Data Preprocessing

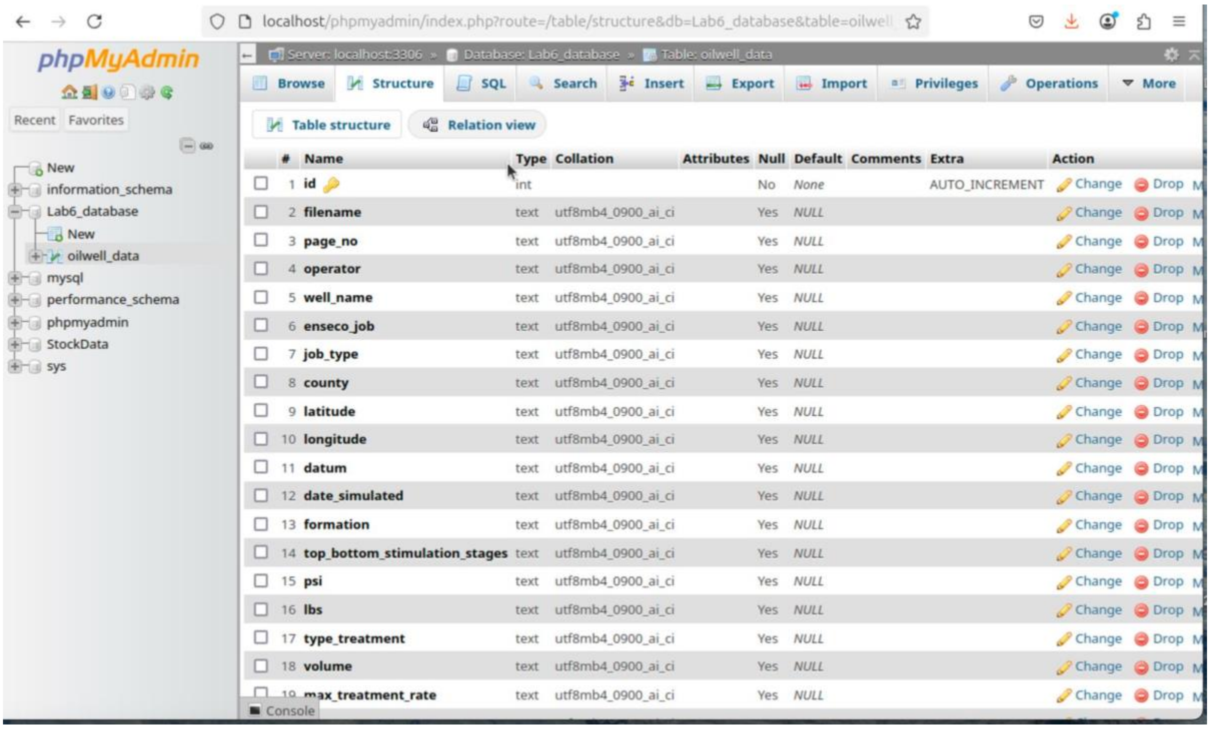
- Removal of missing values.
- Formatting and structuring extracted information.

Screenshots and Output

PDF Extraction and Data Processing Output

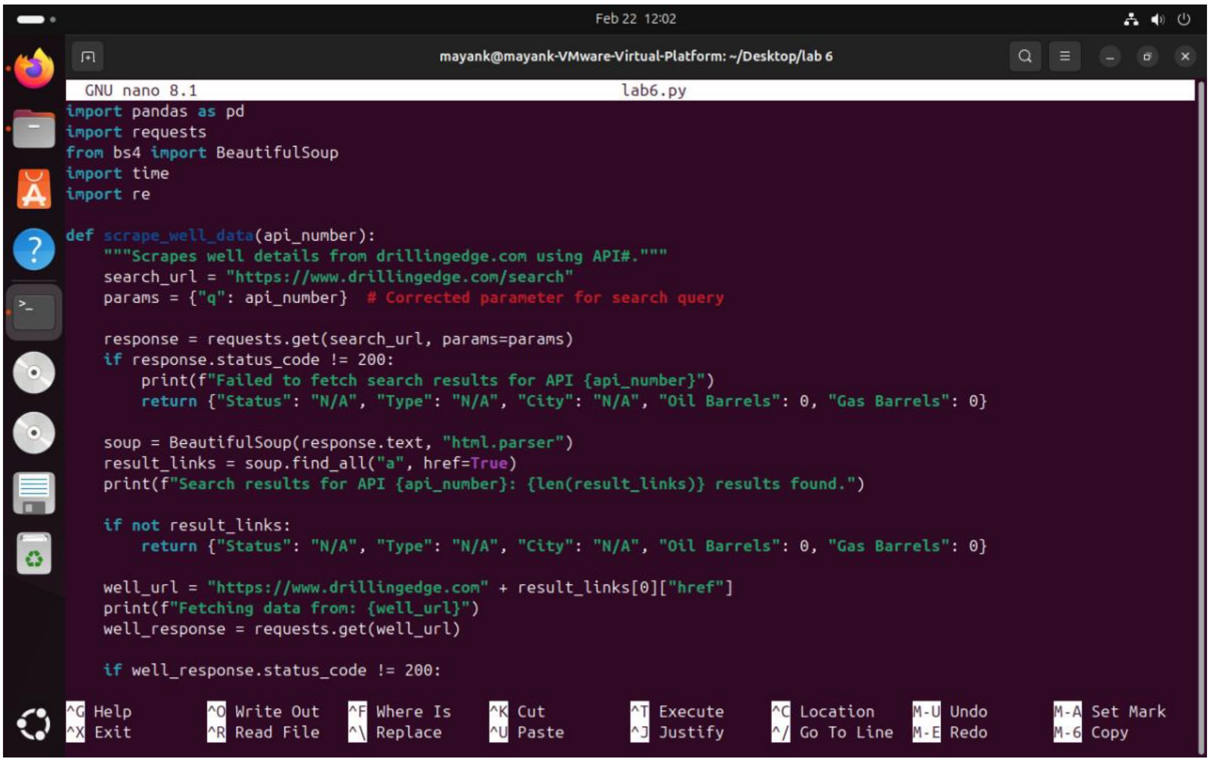


```
cripts/Lab6$ python3 Lab6_pdf_extraction.py
extracted_data.csv has been deleted.


 file name DSCI560_Lab5-20250221T170750Z-001/DSCI560_Lab5/W28425.pdf

Number of pages: 94


 file name DSCI560_Lab5-20250221T170750Z-001/DSCI560_Lab5/W23371.pdf

Number of pages: 302


 file name DSCI560_Lab5-20250221T170750Z-001/DSCI560_Lab5/W23230.pdf

Number of pages: 146

 file name DSCI560_Lab5-20250221T170750Z-001/DSCI560_Lab5/W20864.pdf

Number of pages: 151
Count of pdf files 2
Database 'Lab6_database' created or already exists.
Table 'oilwell_data' created or already exists with columns from CSV headers.
Inserted 279 rows into 'oilwell_data' from 'extracted_data.csv'.
MySQL Connection Closed.
(my_virtual_env) prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/s
```



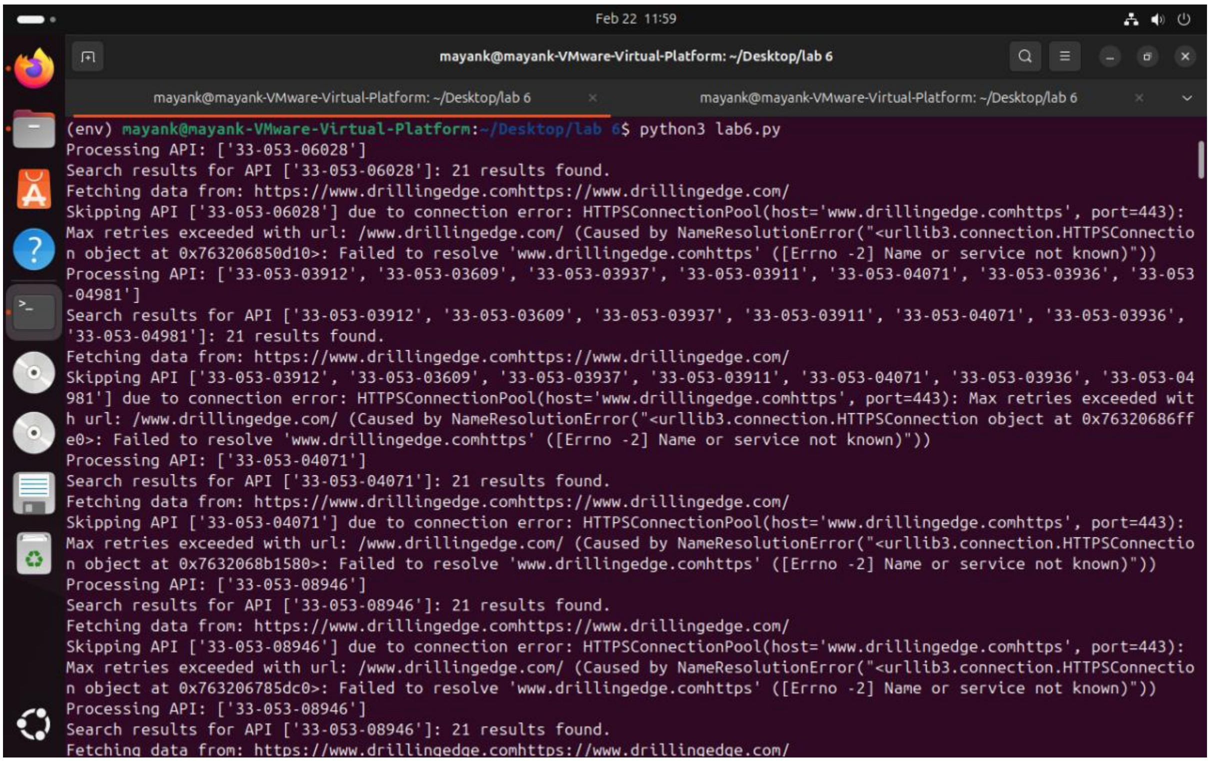
#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
1	id	int			No	None		AUTO_INCREMENT	Change Drop
2	filename	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
3	page_no	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
4	operator	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
5	well_name	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
6	enseco_job	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
7	job_type	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
8	county	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
9	latitude	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
10	longitude	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
11	datum	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
12	date_simulated	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
13	formation	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
14	top_bottom_stimulation_stages	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
15	psi	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
16	lbs	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
17	type_treatment	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
18	volume	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop
19	max_treatment_rate	text	utf8mb4_0900_ai_ci		Yes	NULL			Change Drop

Web scraping code

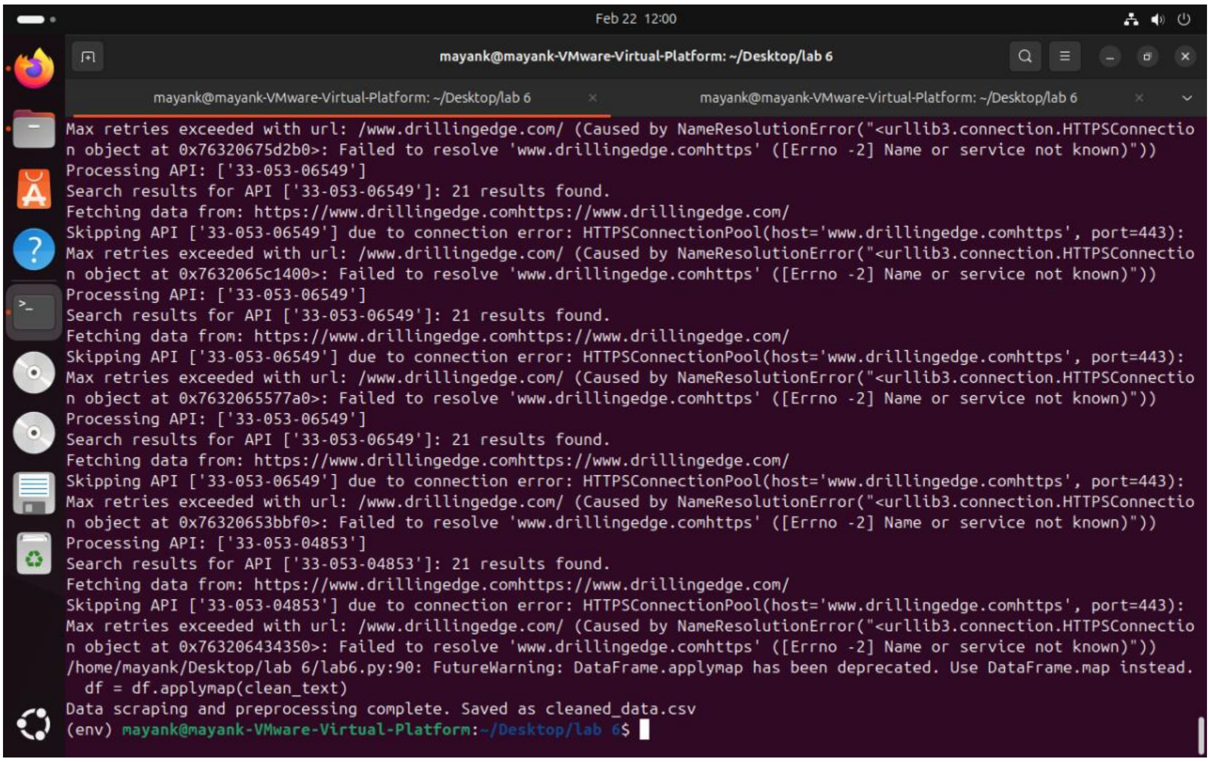


```
  GNU nano 8.1                                  lab6.py
import pandas as pd
import requests
from bs4 import BeautifulSoup
import time
import re

def scrape_well_data(api_number):
    """Scrapes well details from drillingedge.com using API#."""
    search_url = "https://www.drillingedge.com/search"
    params = {"q": api_number}  # Corrected parameter for search query

    response = requests.get(search_url, params=params)
    if response.status_code != 200:
        print(f"Failed to fetch search results for API {api_number}")
        return {"Status": "N/A", "Type": "N/A", "City": "N/A", "Oil Barrels": 0, "Gas Barrels": 0}

    soup = BeautifulSoup(response.text, "html.parser")
    result_links = soup.find_all("a", href=True)
    print(f"Search results for API {api_number}: {len(result_links)} results found.")

    if not result_links:
        return {"Status": "N/A", "Type": "N/A", "City": "N/A", "Oil Barrels": 0, "Gas Barrels": 0}

    well_url = "https://www.drillingedge.com" + result_links[0]["href"]
    print(f"Fetching data from: {well_url}")
    well_response = requests.get(well_url)

    if well_response.status_code != 200:

^G Help       ^O Write Out  ^F Where Is   ^K Cut       ^T Execute   ^C Location   M-U Undo     M-A Set Mark
^X Exit       ^R Read File  ^\ Replace    ^U Paste     ^J Justify   ^/ Go To Line M-E Redo     M-6 Copy
```

Web Scraping Script Execution

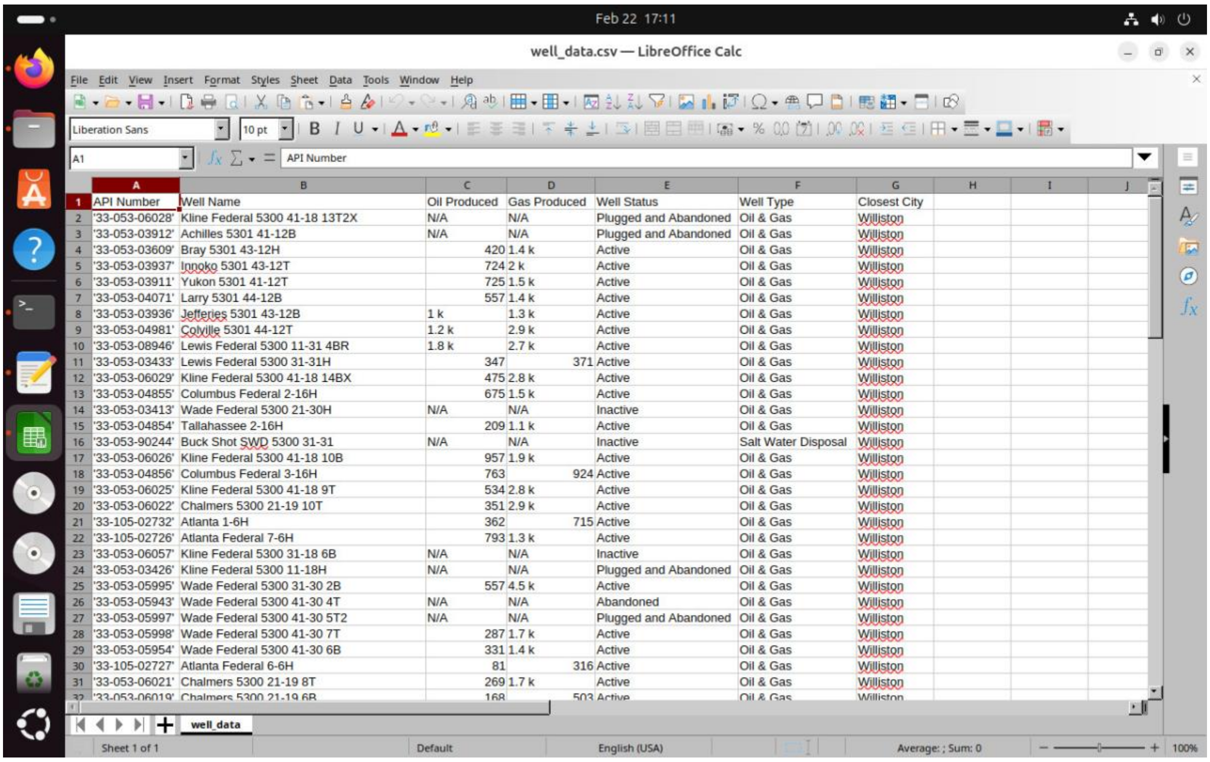

Feb 22 11:59
mayank@mayank-VMware-Virtual-Platform: ~/Desktop/lab 6
mayank@mayank-VMware-Virtual-Platform: ~/Desktop/lab 6
mayank@mayank-VMware-Virtual-Platform: ~/Desktop/lab 6

```
(env) mayank@mayank-VMware-Virtual-Platform:~/Desktop/lab 6$ python3 lab6.py
Processing API: ['33-053-06028']
Search results for API ['33-053-06028']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
Skipping API ['33-053-06028'] due to connection error: HTTPSConnectionPool(host='www.drillingedge.comhttps', port=443):
Max retries exceeded with url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnectio
n object at 0x763206850d10>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
Processing API: ['33-053-03912', '33-053-03609', '33-053-03937', '33-053-03911', '33-053-04071', '33-053-03936', '33-053
-04981']
Search results for API ['33-053-03912', '33-053-03609', '33-053-03937', '33-053-03911', '33-053-04071', '33-053-03936',
'33-053-04981']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
Skipping API ['33-053-03912', '33-053-03609', '33-053-03937', '33-053-03911', '33-053-04071', '33-053-03936', '33-053-04
981'] due to connection error: HTTPSConnectionPool(host='www.drillingedge.comhttps', port=443): Max retries exceeded wit
h url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnection object at 0x76320686ff
e0>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
Processing API: ['33-053-04071']
Search results for API ['33-053-04071']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
Skipping API ['33-053-04071'] due to connection error: HTTPSConnectionPool(host='www.drillingedge.comhttps', port=443):
Max retries exceeded with url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnectio
n object at 0x7632068b1580>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
Processing API: ['33-053-08946']
Search results for API ['33-053-08946']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
Skipping API ['33-053-08946'] due to connection error: HTTPSConnectionPool(host='www.drillingedge.comhttps', port=443):
Max retries exceeded with url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnectio
n object at 0x763206785dc0>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
Processing API: ['33-053-08946']
Search results for API ['33-053-08946']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
```


Feb 22 12:00
mayank@mayank-VMware-Virtual-Platform: ~/Desktop/lab 6
mayank@mayank-VMware-Virtual-Platform: ~/Desktop/lab 6
mayank@mayank-VMware-Virtual-Platform: ~/Desktop/lab 6

```
Max retries exceeded with url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnectio
n object at 0x76320675d2b0>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
Processing API: ['33-053-06549']
Search results for API ['33-053-06549']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
Skipping API ['33-053-06549'] due to connection error: HTTPSConnectionPool(host='www.drillingedge.comhttps', port=443):
Max retries exceeded with url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnectio
n object at 0x7632065c1400>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
Processing API: ['33-053-06549']
Search results for API ['33-053-06549']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
Skipping API ['33-053-06549'] due to connection error: HTTPSConnectionPool(host='www.drillingedge.comhttps', port=443):
Max retries exceeded with url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnectio
n object at 0x7632065577a0>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
Processing API: ['33-053-06549']
Search results for API ['33-053-06549']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
Skipping API ['33-053-06549'] due to connection error: HTTPSConnectionPool(host='www.drillingedge.comhttps', port=443):
Max retries exceeded with url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnectio
n object at 0x76320653bbf0>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
Processing API: ['33-053-04853']
Search results for API ['33-053-04853']: 21 results found.
Fetching data from: https://www.drillingedge.comhttps://www.drillingedge.com/
Skipping API ['33-053-04853'] due to connection error: HTTPSConnectionPool(host='www.drillingedge.comhttps', port=443):
Max retries exceeded with url: /www.drillingedge.com/ (Caused by NameResolutionError("<urllib3.connection.HTTPSConnectio
n object at 0x763206434350>: Failed to resolve 'www.drillingedge.comhttps' ([Errno -2] Name or service not known)"))
/home/mayank/Desktop/lab 6/lab6.py:90: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
  df = df.applymap(clean_text)
Data scraping and preprocessing complete. Saved as cleaned_data.csv
(env) mayank@mayank-VMware-Virtual-Platform:~/Desktop/lab 6$
```

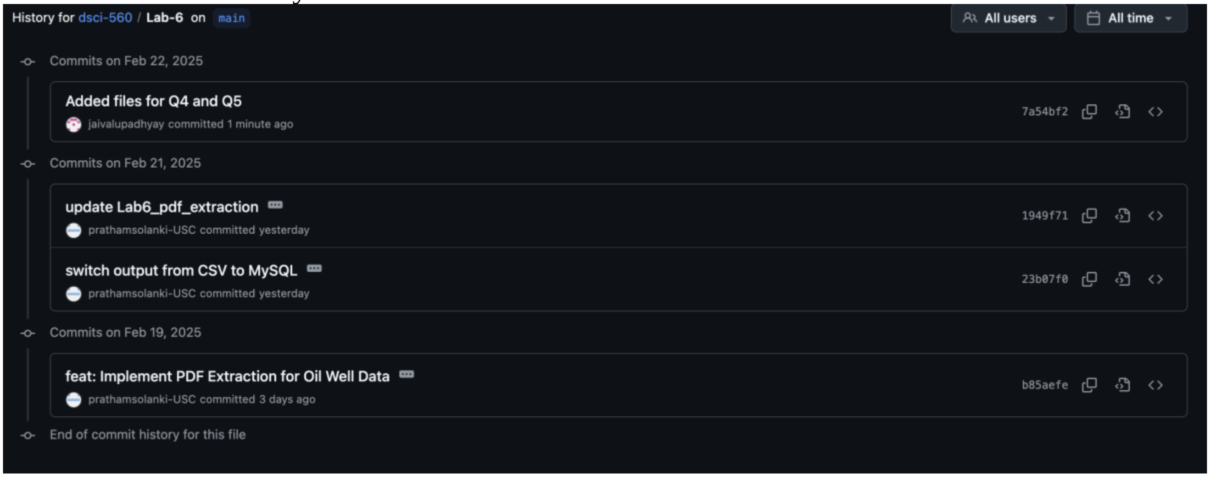
Processed and Scraped Data csv File


Feb 22 17:11
well_data.csv — LibreOffice Calc
File Edit View Insert Format Styles Sheet Data Tools Window Help
Liberation Sans
10 pt
A1

= API Number

API Number	Well Name	Oil Produced	Gas Produced	Well Status	Well Type	Closest City
'33-053-06028'	Kline Federal 5300 41-18 13T2X	N/A	N/A	Plugged and Abandoned	Oil & Gas	Williston
'33-053-03912'	Achilles 5301 41-12B	N/A	N/A	Plugged and Abandoned	Oil & Gas	Williston
'33-053-03609'	Bray 5301 43-12H	420	1.4 k	Active	Oil & Gas	Williston
'33-053-03937'	Innoko 5301 43-12T	724	2 k	Active	Oil & Gas	Williston
'33-053-03911'	Yukon 5301 41-12T	725	1.5 k	Active	Oil & Gas	Williston
'33-053-04071'	Larry 5301 44-12B	557	1.4 k	Active	Oil & Gas	Williston
'33-053-03936'	Jefferies 5301 43-12B	1 k	1.3 k	Active	Oil & Gas	Williston
'33-053-04981'	Colville 5301 44-12T	1.2 k	2.9 k	Active	Oil & Gas	Williston
'33-053-08946'	Lewis Federal 5300 11-31 4BR	1.8 k	2.7 k	Active	Oil & Gas	Williston
'33-053-03433'	Lewis Federal 5300 31-31H	347	371	Active	Oil & Gas	Williston
'33-053-06029'	Kline Federal 5300 41-18 14BX	475	2.8 k	Active	Oil & Gas	Williston
'33-053-04855'	Columbus Federal 2-16H	675	1.5 k	Active	Oil & Gas	Williston
'33-053-03413'	Wade Federal 5300 21-30H	N/A	N/A	Inactive	Oil & Gas	Williston
'33-053-04854'	Tallahassee 2-16H	209	1.1 k	Active	Oil & Gas	Williston
'33-053-90244'	Buck Shot SWD 5300 31-31	N/A	N/A	Inactive	Salt Water Disposal	Williston
'33-053-06026'	Kline Federal 5300 41-18 10B	957	1.9 k	Active	Oil & Gas	Williston
'33-053-04856'	Columbus Federal 3-16H	763	924	Active	Oil & Gas	Williston
'33-053-06025'	Kline Federal 5300 41-18 9T	534	2.8 k	Active	Oil & Gas	Williston
'33-053-06022'	Chalmers 5300 21-19 10T	351	2.9 k	Active	Oil & Gas	Williston
'33-105-02732'	Atlanta 1-6H	362	715	Active	Oil & Gas	Williston
'33-105-02726'	Atlanta Federal 7-6H	793	1.3 k	Active	Oil & Gas	Williston
'33-053-06057'	Kline Federal 5300 31-18 6B	N/A	N/A	Inactive	Oil & Gas	Williston
'33-053-03426'	Kline Federal 5300 11-18H	N/A	N/A	Plugged and Abandoned	Oil & Gas	Williston
'33-053-05995'	Wade Federal 5300 31-30 2B	557	4.5 k	Active	Oil & Gas	Williston
'33-053-05943'	Wade Federal 5300 41-30 4T	N/A	N/A	Abandoned	Oil & Gas	Williston
'33-053-05997'	Wade Federal 5300 41-30 5T2	N/A	N/A	Plugged and Abandoned	Oil & Gas	Williston
'33-053-05998'	Wade Federal 5300 41-30 7T	287	1.7 k	Active	Oil & Gas	Williston
'33-053-05954'	Wade Federal 5300 41-30 6B	331	1.4 k	Active	Oil & Gas	Williston
'33-105-02727'	Atlanta Federal 6-6H	81	316	Active	Oil & Gas	Williston
'33-053-06021'	Chalmers 5300 21-19 8T	269	1.7 k	Active	Oil & Gas	Williston
'33-053-06019'	Chalmers 5300 21-19 6B	168	503	Active	Oil & Gas	Williston

Github Commit History:



Commits on Feb 22, 2025

Added files for Q4 and Q5
jaivalupadhyay committed 1 minute ago
7a54bf2

Commits on Feb 21, 2025

update Lab6_pdf_extraction
prathamsolanki-USC committed yesterday
1949f71

switch output from CSV to MySQL
prathamsolanki-USC committed yesterday
23b07f0

Commits on Feb 19, 2025

feat: Implement PDF Extraction for Oil Well Data
prathamsolanki-USC committed 3 days ago
b85aefe

End of commit history for this file