A picture containing text

Description automatically generated

**MLOPS**

Assignment No.2

Section: CS-B

Submitted to: Pir Sami Ullah Shah

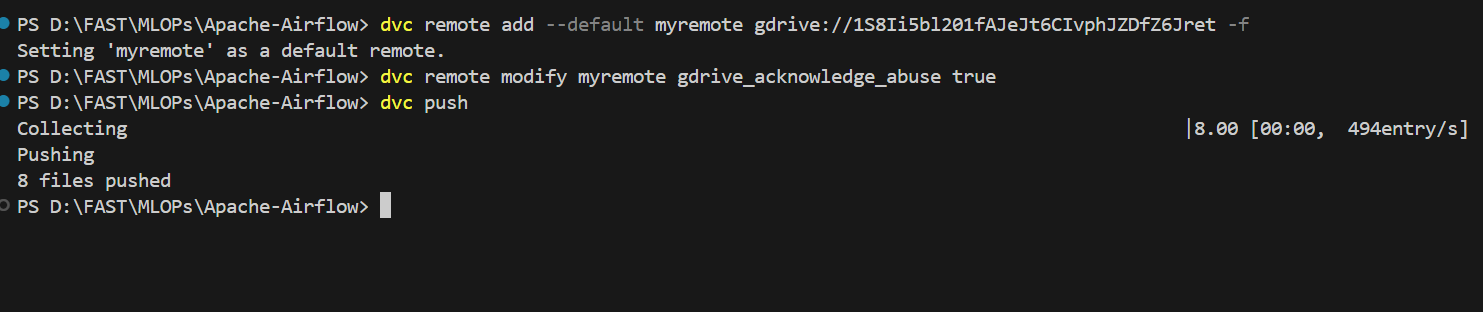
Submitted on: 12/05/2024

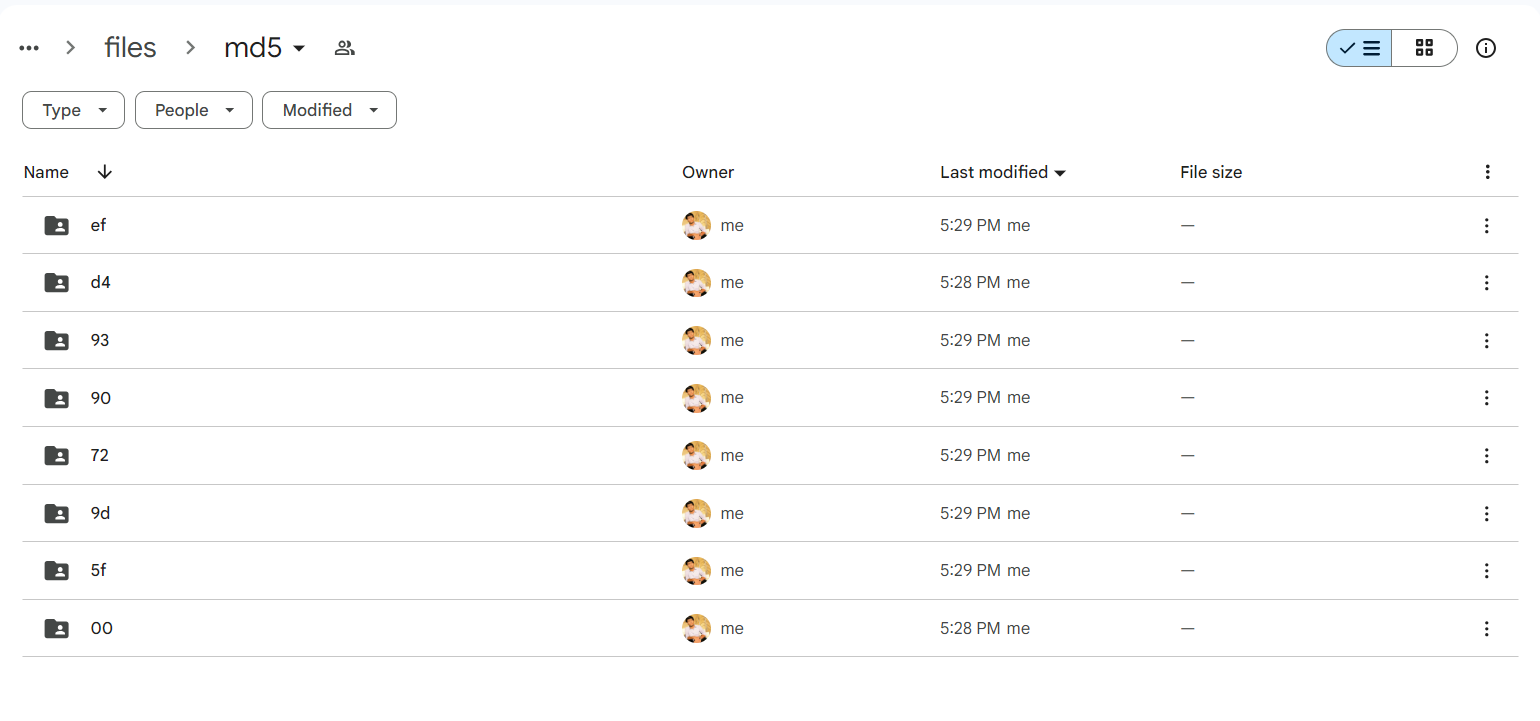
**DVC Setup:**

First of all I initialized my project with DVC.

Then I linked the data folder of the DVC with the folder on my google drive by creating a folder there.

I copied the unique Id of the folder and linked the gdrive folder with the local data folder.





**Data Extraction and Cleaning:**

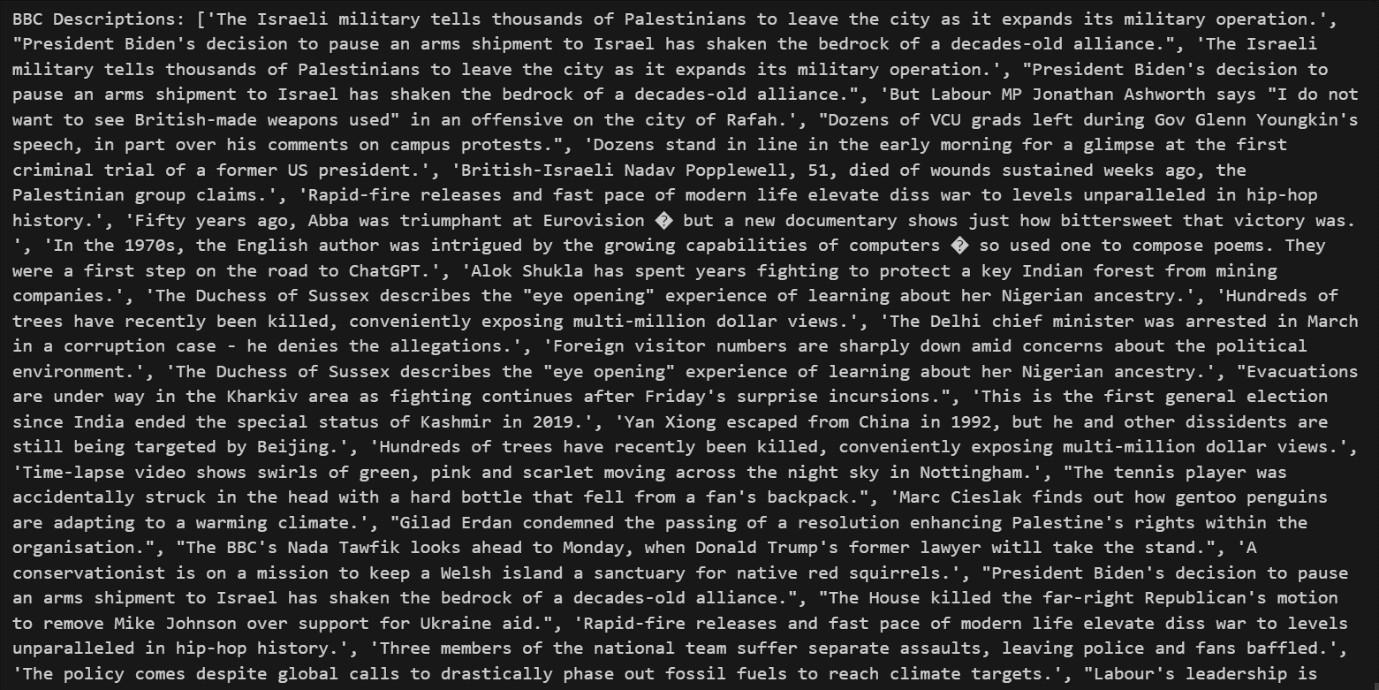
I used the beautifulsoap library to extract data from the websites like bbc.com and dawn.com.

I extracted the news titles, descriptions and links from the websites.

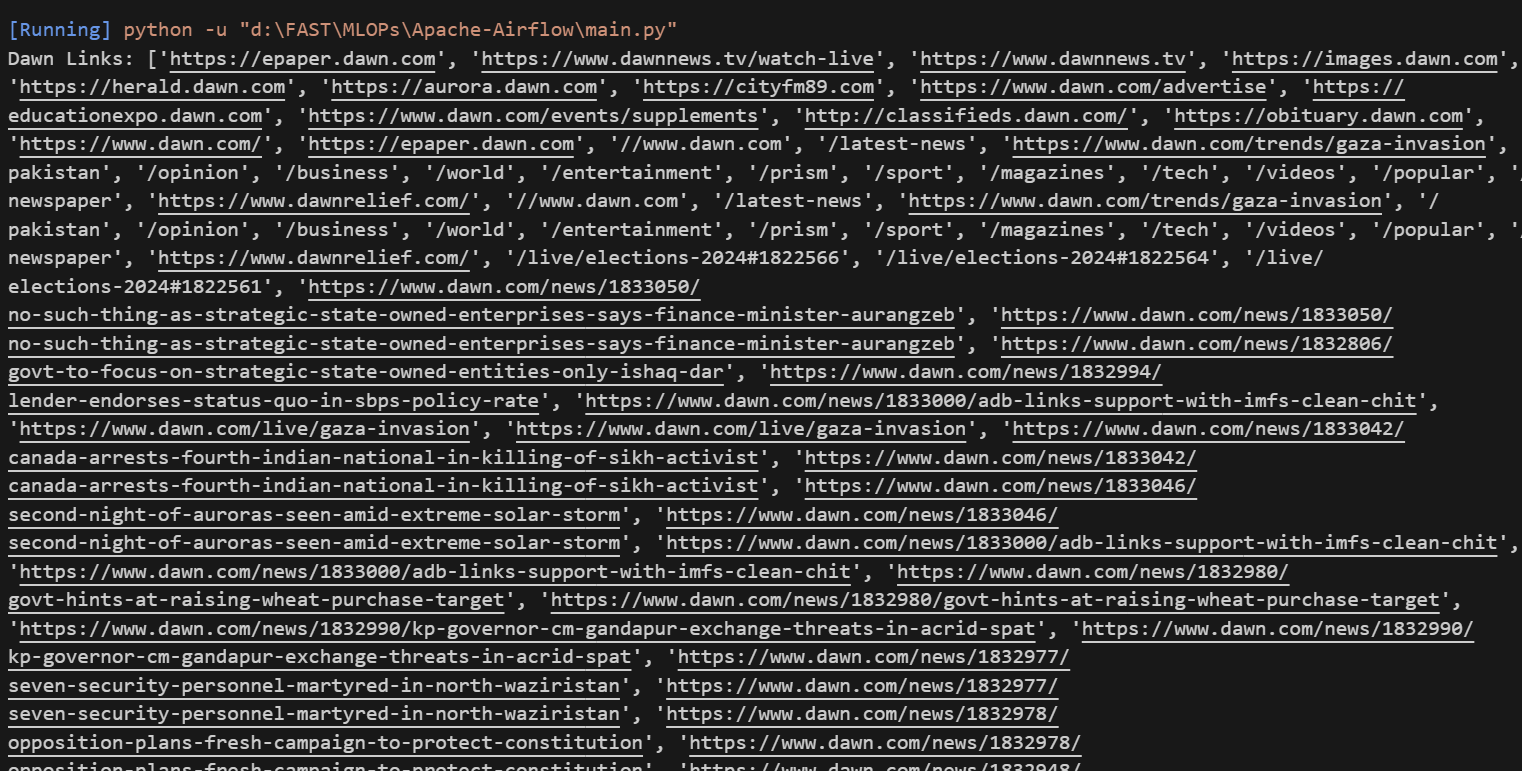
The extracted data had unreadable characters that I removed by writing a script for data cleaning.

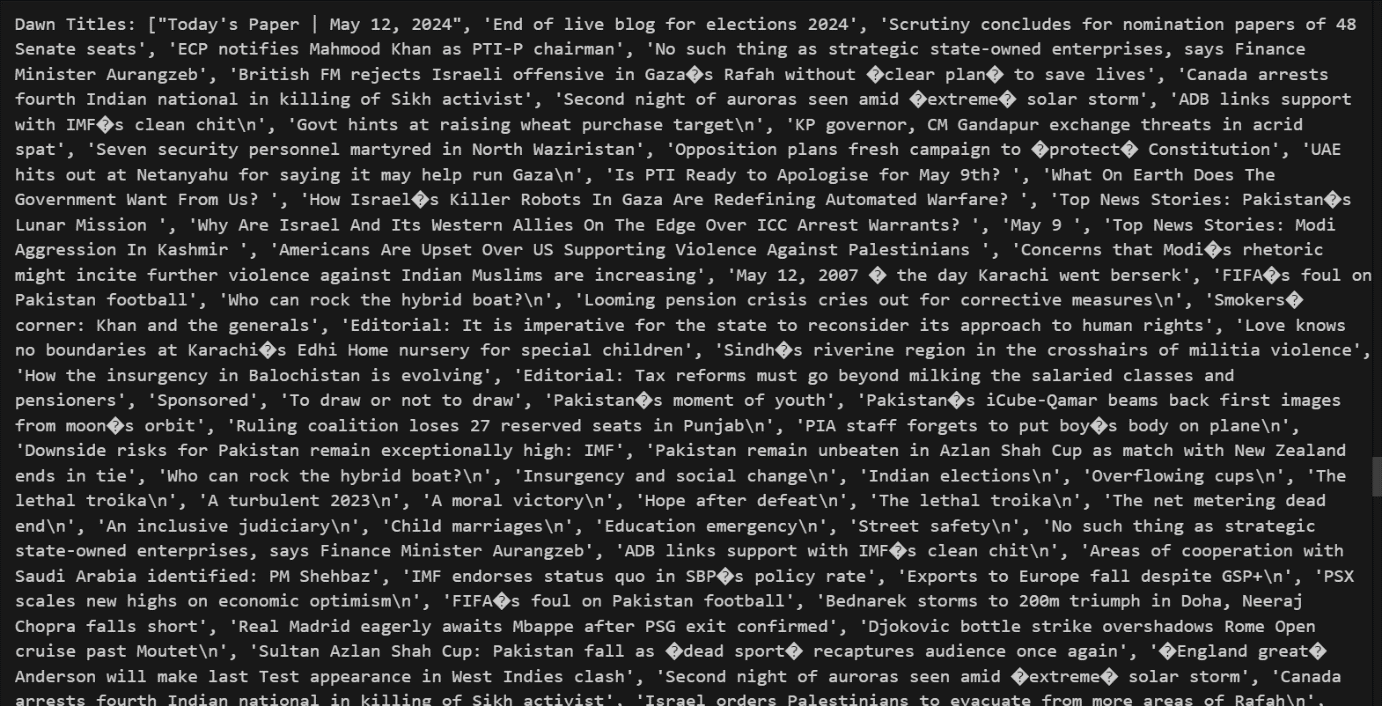
Also made the data more visually appealable my formatting it.

As the data is fetched, the data is stored in the dvc data folder.

****

****

****

****

**Airflow Dag Setup:**

I setup docker to install the airflow setup.

I downloaded the prebuilt airflow image from apache website.

I did docker compose up to start the container.

The container is mapped with port 8070 of local machine.

Then I wrote DAG file to automate all the processes.

Here is the break down of DAG,

1. **Define Default Settings**: Set up default setting for task that when they should start.
2. **Define Functions**: write Python functions to do different jobs, like getting information from websites, cleaning it up, and saving it to files.
3. **Define the DAG**: created a Directed Acyclic Graph (DAG) named 'web\_scraping\_and\_preprocessing'. This DAG organizes our tasks and their order.
4. **Define Tasks**: defined each task in the DAG using PythonOperator.
5. **Define Task Dependencies**: defined which task follow the other.

