Data Preprocessing steps and DVC setup

# Data Preprocessing

1. Data Extraction

* Utilized `requests` library to fetch HTML content from `dawn.com` and `bbc.com`.
* Utilized `BeautifulSoup` library to parse the HTML content and extract links from the landing pages.
* Extracted titles and descriptions from articles displayed on the homepages.

2. Data Transformation

* Loaded the extracted data into a Pandas DataFrame.
* Cleaned the text data by:
* Lowercasing all letters.
* Removing unnecessary special characters using regular expressions.
* Transformed the data by creating a new column with cleaned descriptions.
* Saved the transformed data to a CSV file named `dataset\_news\_new.csv`.

# DVC Setup

1. Initialization

* Initialized Git repository using the BashOperator with the command: `git init`.

2. DVC Initialization

* Initialized DVC repository with the force option using the BashOperator with the command: `dvc init --force`.

3. Remote Setup

* Added Google Drive as a remote for DVC using the BashOperator with the command: `dvc remote add -d my\_google\_drive gdrive://1i3kGKsxTKMw3-tN9wpJ6fHjrIOoFaO0j -f`.

4. Data Tracking and Version Control

* Used the BashOperator to add the transformed CSV file (`dataset\_news\_new.csv`) to DVC with the command: `dvc add dataset\_news\_new.csv`.
* Pushed the data with DVC using the BashOperator with the command: `dvc push`.

# DAG Structure

* The DAG `Data\_Extraction\_Transformation\_Loading\_5` automates the processes of data extraction, transformation, and storage.
* Task dependencies are handled effectively, ensuring sequential execution of tasks.
* Error management is implemented to handle failures gracefully.