Muhammad Subhan
20I-0873
DS - N

MLOPS Assignment 2

Overview

Task 1: Extract Links and Metadata

The two sources of news that are extracted for this assignment are 'https://www.dawn.com/latest-news' and 'https://www.bbc.com/news'. The title and description of each article are then extracted by processing it. The article is skipped if the description or title are too brief or absent. To ascertain the sentiment of the description text, a pre-trained model is employed for sentiment analysis. The processed data is gathered and provided back, together with the article ID, sentiment, title, description, and source.

Task 2: CSV data storage

The information gathered in Task 1 is saved in a CSV file by this task. It determines whether the needed directory already exists and, if not, creates it. After that, it gets the information from the XCom that Task 1 passed. It raises an error if no data is received. 'id, sentiment, title, description, and source' are the field names that are sent to a CSV file with the data. Every row represents one article. A success message is printed at the end.

Task 3: Push to Git

The CSV file containing the data that was scraped is pushed to a Git repository by this task. It first verifies the current directory before changing to the proper directory (assuming the directory is part of a project hierarchy). After that, a script is created to add the CSV file to Data Version Control (DVC), push it, add all modifications to Git, commit a message, and push the file to the Git repository. Any errors that arise when the command is being executed are recorded and printed.

The DAG started one day ago and is planned to run every day. Task 1 (extract_links) initiates Task 2 (store_data_in_csv), which in turn initiates Task 3 (push_to_git). This is how the tasks are linked sequentially.

1. CSV file created after DAG run

	A	B	C	
	id	sentiment	title	description
	1	negative	justice babar sattar denounces lawyers strike, blocking litigants entry to courts	justice babar sattar denounces lawyers strike, blocking litigants entry to courts lawyers
	2	neutral	formation of na standing panels still in limbo	formation of na standing panels still in limbo na secretariat declares 82 pti-backed mna
	3	neutral	seven security personnel martyred in north waziristan	seven security personnel martyred in north waziristan improvised explosive device targ
	4	neutral	opposition plans fresh campaign to protect constitution	opposition plans fresh campaign to protect constitution six-party coalition to approach c
	5	positive	govt hints at raising wheat purchase target	govt hints at raising wheat purchase target drive to buy 400,000 tonnes of grain kicks o
	6	neutral	1 police officer dead, over 90 injured in ajk clashes	1 police officer dead, over 90 injured in ajk clashes kotli ssp blames "attacks by miscre
	7	neutral	england great anderson will make last test appearance in west indies clash	england great anderson will make last test appearance in west indies clash the 41-year-
	8	neutral	pti issues show-cause notice to sher afzal marwat for harming party interests, violating discipline	pti issues show-cause notice to sher afzal marwat for harming party interests, violating
	9	neutral	men must take the lead in standing up against sexual harassment, says hamza ali abbasi	men must take the lead in standing up against sexual harassment, says hamza ali abb
	10	neutral	first extreme solar storm in 20 years strikes earth, brings spectacular auroras	first extreme solar storm in 20 years strikes earth, brings spectacular auroras social me
	11	negative	india poll watchdog s inaction lets pm modi commit brazen violations, opposition says	india poll watchdog s inaction lets pm modi commit brazen violations, opposition says r
	12	positive	sultan azlan shah cup: pakistan fall as dead sport recaptures audience once again	sultan azlan shah cup: pakistan fall as dead sport recaptures audience once again japa
	13	neutral	president zardari granted immunity from criminal proceedings by islamabad accountability court	president zardari granted immunity from criminal proceedings by islamabad accountabil

2. Google Drive Folder:
https://drive.google.com/drive/u/1/folders/11mmC-ZG7wlXiN3VaoiBm2iVyFMeTkydA

3. Airflow History