MLOps Assignment 2 - Documentation

Umer Mukhtar i20-0696

Department of Computer Science, National University of Computer and Engineering, Islamabad, Pakistan

1 DAG

1.1 Extract Task

Extracts data from a list of URLs (urls) using the `extract_data` function. Each URL is processed sequentially, and the extracted data is combined.

1.2 Preprocess Task

Preprocesses the extracted data using the `clean_data` function. This task ensures that the text data is cleaned and formatted consistently.

1.3 Save Task

Saves the preprocessed data to a CSV file specified by filename. The data is saved in the format of 'id', 'title', 'description', and 'source' columns.

1.4 DVC Push Task

Adds the CSV file (`data/extracted.csv`) to the DVC repository using `dvc add` command and pushes the changes to the remote gdrive storage using `dvc push`.

1.5 Git Push Task

Performs Git operations to push the changes made to the Git repository. It includes commands to pull changes, add files, commit changes, and push to the remote repository.

2 DAG Execution Order

The tasks are executed sequentially in the following order:

```
extract_task >> preprocess_task >> save_task >> dvc_push_task >> git_push_task
```

The DAG is configured to run manually (schedule=None) and does not have a specific schedule for automatic execution.

3 Encountered Challenges

3.1 Challenge: Changing the airflow configuration to detect dags present in locations other than airflow folder

Solution: Make a dags directory in your current folder, places dags there and set environment variable `export AIRFLOW_HOME="current folder"`

3.2 Challenge: How to automate git and dvc commands?

Solution: Use python os library

3.3 Challenge: Installing airflow in python virtual environment.

Solution: `pip install apache-airflow`

3.4 Challenge: Dataset not being saved using airflow

Solution: Use absolute path for dataset

4 Points To Note

- Place all/any dags in the /dags folder for airflow to detect.
- Run `airflow dags list` to check if airflow properly picks up dags from the dagbag (dag folder which contains all dags).