CSCI461 Big Data Assignment #1

Overview

This project processes a dataset using a pipeline built with Docker and Python. It includes steps for data loading, preprocessing, exploratory data analysis (EDA), visualization, and K-means clustering. The output files are stored in the res directory.

Requirements

- Docker
- Dataset (iris.csv used in this example)

Project Structure

- **Dockerfile**: Defines the Docker container setup with necessary packages.
- **load.py**: Loads the dataset and saves it as loaded_data.csv.
- **dpre.py**: Performs data cleaning, transformation, reduction, and discretization on the dataset. The output is saved as res_dpre.csv.
- **eda.py**: Generates insights from the data and saves them in eda-in-1.txt, eda-in-2.txt, and eda-in-3.txt.
- **vis.py**: Creates a visualization and saves it as vis.png.
- **model.py**: Applies K-means clustering with k=3 and saves the cluster counts in k.txt.

Setup and Execution

1. Build the Docker Image

In the bd-a1 directory (where the Dockerfile is located), build the Docker image:

bash

Copy code

```
docker build -t bd-a1-image .
```

2. Run the Docker Container

Run the container interactively to access the bash shell:

bash

Copy code

```
docker run -it --name bd-a1-container bd-a1-image
```

3. Execute the Pipeline

Inside the Docker container, navigate to /home/doc-bd-a1/ and execute each Python script in the following order:

1. **Load the Data**:

bash

Copy code

```
python3 load.py /home/doc-bd-a1/iris.csv
```

2. **Preprocess the Data**:

bash

Copy code

```
python3 dpre.py
```

3. **Perform EDA**:

bash

Copy code

```
python3 eda.py
```

4. **Generate Visualization**:

bash

Copy code

```
python3 vis.py
```

5. **Apply K-means Clustering**:

bash

Copy code

```
python3 model.py
```

4. Copy Output Files to the Local Machine

After executing the pipeline, copy the generated files from the container to the res directory on your local machine:

bash

Copy code

```
docker cp bd-a1-container:/home/doc-bd-a1/res_dpre.csv C:\Users\Salma\Desktop\bd-a1\res\

docker cp bd-a1-container:/home/doc-bd-a1/eda-in-1.txt C:\Users\Salma\Desktop\bd-a1\res\
docker cp bd-a1-container:/home/doc-bd-a1/eda-in-2.txt C:\Users\Salma\Desktop\bd-a1\res\

docker cp bd-a1-container:/home/doc-bd-a1/eda-in-3.txt C:\Users\Salma\Desktop\bd-a1\res\

docker cp bd-a1-container:/home/doc-bd-a1/vis.png C:\Users\Salma\Desktop\bd-a1\res\

docker cp bd-a1-container:/home/doc-bd-a1/k.txt C:\Users\Salma\Desktop\bd-a1\res\
```

Output Files

- **res_dpre.csv**: Preprocessed data.
- **eda-in-1.txt, eda-in-2.txt, eda-in-3.txt**: EDA insights.
- **vis.png**: Visualization image.
- **k.txt**: Cluster counts from K-means.

Troubleshooting

- Ensure that all Python scripts are copied into the /home/doc-bd-a1/ directory in the container before execution.
- If you encounter any issues with file paths, ensure they are specified in Unix-style (e.g., /home/doc-bd-a1/iris.csv).

Bonus (Optional)

1. **Push Docker Image to Docker Hub**:
   - Tag and push the Docker image:

bash

Copy code

```
docker tag bd-a1-image salmaheshamsalem123/bd-a1-image

docker push salmaheshamsalem123/bd-a1-image
```

2. **Push Project to GitHub**:
   - Create a repository on GitHub, add your files, commit, and push.