Fasih Abdullah i200432 DS-N

Data preprocessing:

1) All of the titles and descriptions were filtered to remove non ascii characters.
2) Endlines, carriage returns, and tab spaces were removed from the data.
3) Empty titles and descriptions were removed.
4) Titles and descriptions were converted to lowercase.

```python
# this function extracts the titles and decriptions then preprocesses them
def transform(**kwargs):

    links_dataframe = Variable.get("links_dataframe", deserialize_json=True) # data from step 1

    if type(links_dataframe) != pd.DataFrame:
        links_dataframe = pd.read_json(links_dataframe)

    print("Transformation")

    # process the titles
    links_dataframe['Title'] = links_dataframe['Title'].apply(lambda x: x.encode('ascii', 'ignore').decode('ascii'))
    links_dataframe['Title'] = links_dataframe['Title'].apply(lambda x: x.replace('\n', ' '))
    links_dataframe['Title'] = links_dataframe['Title'].apply(lambda x: x.replace('\r', ' '))
    links_dataframe['Title'] = links_dataframe['Title'].apply(lambda x: x.replace('\t', ' '))

    # preprocess description to remove symbols
    links_dataframe['Description'] = links_dataframe['Description'].apply(lambda x: x.encode('ascii', 'ignore').decode('ascii'))
    links_dataframe['Description'] = links_dataframe['Description'].apply(lambda x: x.replace('\n', ' '))
    links_dataframe['Description'] = links_dataframe['Description'].apply(lambda x: x.replace('\r', ' '))
    links_dataframe['Description'] = links_dataframe['Description'].apply(lambda x: x.replace('\t', ' '))


    # remove any rows with empty titles or descriptions
    links_dataframe = links_dataframe[links_dataframe['Title'] != '']
    links_dataframe = links_dataframe[links_dataframe['Description'] != '']

    # make titles lowercase
    links_dataframe['Title'] = links_dataframe['Title'].apply(lambda x: x.lower())

    # make descriptions lowercase
    links_dataframe['Description'] = links_dataframe['Description'].apply(lambda x: x.lower())

    print(links_dataframe.head())
```

DVC Setup:

1) Created a google drive folder
2) dvc init in repo folder
3) Add dvc remove → dvc remote add -d storage gdrive://1cBko_FDMEWlldZDYaPg0lN53Uq25Zbk1
4) Perform authentication and give permissions
5) Add dataset to dvc for tracking → dvc add ./data/links.csv
6) Push dataset to remote → dvc push
7) Add tracking of dataset metadata in git → git add data.dvc
8) Make commit → git commit -m “Update dataset”
9) Push to remote repository → git push origin main