CAN4010 Project 4 – Cam Hayes

The purpose of this document is to provide guidance into the components of this project.

Data:

For the purposes of file transfer, the base dataset is not included with this project, but can be downloaded directly from: https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

Both the rotten_tomatoes_critic_reviews.csv and rotten_tomatoes_movies.csv datasets are used in this project.

Processed meta and sentiment data and model data is found in the data directory:

- movie_metadata.csv: the fully parsed and normalized directory of data derived from the source RT data containing film meta data, review text, and normalized review score.
- movie_metadata-trainer.csv: a randomly generated subset review text and critical scores from movie_metadata. Used for training FilmBERT
- regression_data.csv: finalized accuracy and R2 scores from prediction model training.
- /sentiment: sentiment scores from each sentiment model paired with critical review score
- /sentiment_meta: sentiment scores from each sentiment model paired with critical review score and film metadata
- /models: contains the transformer model directory for the trained FilmBERT model

Source Code:

Project4-collab.ipynb

This is a compiled Jupyter file containing all of the code used for pre-processing data, sentiment generation, FilmBERT model training, and regression model predictions. This file may be opened using Google Collab, VSCode, and other tools that support Jupyter type files.

This compiled file is segmented into four sections: Setup, Sentiment Analysis Models, Prediction Testing, FilmBERT Training. There is no need to execute a new runtime for this project, however you may do so by:

- Running cells 1 and 2 and mounting a drive.
- Changing file reference locations for rotten_tomatoes_critic_reviews.csv and rotten_tomatoes_movies.csv.
- Running the training and sentiment generating **FilmBERT Training** cells.
- Running sentiment generating cells for the remaining models found under **Sentiment Analysis Models**.
- Running the regression prediction models under **Prediction Testing.**