**User Guide for IMDb Movie Ratings Analysis**

**Introduction**

The IMDb Movie Ratings Analysis application is designed to help users analyze IMDb movie datasets by merging, cleaning, and filtering data. Users can gain insights into movie ratings, votes, and genres with just a few steps. This guide explains how to set up, use, and explore the application.

**System Requirements**

* **Python Version**: 3.7 or higher
* **Operating System**: Windows, macOS, or Linux
* **Dependencies**: Listed in requirements.txt

Ensure Python and the required packages are installed before running the application.

**Installation Guide**

**1. Clone the Repository**

Clone the project to your local machine using the following command:

git clone https://github.com/<your-username>/imdb-movie-ratings-analysis.git

Replace <your-username> with your GitHub username.

**2. Install Dependencies**

Navigate to the project directory and install required libraries using:

pip install -r requirements.txt

**3. Download IMDb Data**

Place the required IMDb datasets (title\_basics.csv and title\_ratings.csv) into the data/ folder.

**How to Use**

**Step 1: Run Data Preprocessing**

Use the script data\_preprocessing.py to load and preprocess the data. This step cleans and merges the datasets to prepare for analysis.

python scripts/data\_preprocessing.py

**Step 2: Perform Analysis**

Run the analysis.py script to analyze the processed data. This includes generating basic statistics, filtering top-rated movies, and more.

python scripts/analysis.py

**Step 3: Customize Filters**

You can customize the filtering criteria by editing the threshold parameter in the analysis.py file:

top\_rated = filter\_top\_rated(cleaned\_df, threshold=1000)

Change 1000 to your desired threshold for the minimum number of votes.

**Example Outputs**

**Basic Statistics**

When you run the analysis script, you will see outputs like:

Average Rating: 7.2

Total Votes: 50,000,000

**Filtered Data**

The script can display a list of top-rated movies with details such as:

tconst primaryTitle averageRating numVotes

1 tt0111161 The Shawshank Redemption 9.3 2500000

2 tt0068646 The Godfather 9.2 1700000

**File Structure**

* **data/**: Folder for raw IMDb datasets.
* **scripts/**: Contains Python scripts for data preprocessing and analysis.
  + data\_preprocessing.py: Handles data loading, cleaning, and merging.
  + analysis.py: Performs statistical analysis and filtering.
* **output/**: Stores any generated results or reports.

**Troubleshooting**

1. **Missing Dataset Error**:  
   Ensure the required files (title\_basics.csv, title\_ratings.csv) are placed in the data/ directory.
2. **Dependency Issues**:  
   If the application fails to run, ensure all dependencies are installed using:
3. pip install -r requirements.txt
4. **Python Version**:  
   Use Python 3.7 or higher to avoid compatibility issues.

**Extending the Application**

You can extend the analysis by:

* Adding new datasets, such as title\_crew.csv or title\_principals.csv.
* Enhancing the filtering logic in analysis.py.
* Visualizing the results using libraries like Matplotlib or Seaborn.

For assistance or contributions, feel free to submit issues or pull requests on the repository.