Movie Box Office Data Scraper

1 Project Overview

This Python script (`DataScrapping.py`) scrapes daily box office data for movies from *The Numbers* website. It processes a list of movies (post-2010) from a metadata CSV file, retrieves box office data, and stores the results in structured CSV files.

2 Features

- **Metadata Validation:** Ensures required columns exist and filters movies released after 2010.
- **Parallel Scraping:** Uses multithreading for efficient data collection.
- **Error Handling:** Logs errors for failed extractions.
- **Safe Filenames:** Generates valid filenames for CSV storage.
- **Batch Processing:** Allows incremental data collection.

3 Requirements

Ensure you have the following Python libraries installed:

- `requests`
- `beautifulsoup4`
- `pandas`
- `tqdm`
- `concurrent.futures`

To install missing dependencies, run:

```
pip install requests beautifulsoup4 pandas tqdm
```

4 Usage

4.1 1. Prepare Metadata

Place a CSV file containing movie metadata (with `original_title` and `release_year` columns).

4.2 2. Set Paths in Script

Edit the script and replace:

```
metadata_path = r"path/to/your/movies_metadata.csv"
project_root = r"path/to/your/project/root"
```

4.3 3. Run the Script

Execute in the terminal or command prompt:

```
python DataScrapping.py
```

You will be prompted to enter the starting row, batch size, and number of parallel workers.

4.4 4. Output

- Scraped data will be saved in `data/raw/completed_movies/` inside your project folder.
- A combined dataset (`all_movies_daily_data.csv`) will also be generated.
- Errors will be logged in `error_log.csv`.

5 Notes

- The script includes a delay to avoid overwhelming the website with requests.
- If a movie's data is not found, the next available URL format is attempted.
- Ensure that the metadata file is well-formatted and contains valid movie names.

For any issues, check the error log or verify your metadata file format.