839 project 2 report

Zhendong Zhou *Yunwen Jiang* *Yixin Chen*

Data Source

In this project, we download action movie data from IMDb and Rotten Tomatoes, which are most famous websites of ratings and reviews for movies. The links below are the first page of our search result:

- https://www.imdb.com/search/title?title_type=feature,tv_movie&genres=action&languages=en&start=1&ref_=adv_nxt
- https://www.rottentomatoes.com/api/private/v2.0/browse?maxTomato=100&maxPopcorn=100&services=amazon%3Bhbo_go%3Bitunes%3Bnetflix_iw%3Bvudu%3Bamazon_prime%3Bfandango_now&genres=1&certified&sortBy=release&type=dvd-streaming-all&page=2

Extraction Method

From each website of search result, we try to find out the structure of item link. For example, on IMDb website, the movie name is in the section of … under the section <h3 class="lister-item-header">…</h3> .

The main method we use here is trying to extract information step by step:

First, we extract a large section containing the information.

Finally, we extract the subset to get the information.

Entity description

As we mentioned above, we extract movies' information from IMDb and Rotten Tomatoes, the entities should be movies, the features we extracted are URL, Name, Stars, Year, Time, Rating, Genre and Director. We get 3100 tuples on IMDb and 3262 tuples on Rotten Tomatoes.

Tools

We used python in this project with packages requests, re, csv and json, which are well-known packages for web crawling. Requests is used to obtain data in given

website, re is used for the regular expression we applied to derive useful information, json is used to deal with json files and csv is used to save a list of dictionary data into json file.