CS 839 Project Stage 2

Team Members:

Shuoxuan Dong sdong34@wisc.edu
Ziyun Zeng ziyun.zeng@wisc.edu
Chushi Shi cshi46@wisc.edu

Introduction:

In stage 2, we select two data sources to crawl movie information. One data source is imdb.com, another source comes from rottentomatoes.com. We use Scrapy to crawl web information, and the movie information with different attributes in two webs is stored into two csv tables for future stage analysis.

Tools Used:

We used an open-source framework Scrapy (https://scrapy.org) to extract the data. It parses any HTML page as a list of tags and its attributes. We can then search for specific tags based on some attribute, and then extract the text field (which corresponds to the text being displayed on the web page).

Data Source:

- Movie information(genres, level, title, url, year) on IMDB: https://www.imdb.com/search/title?title_type=feature&release_date
- Moive information(name, director, genre, level, rating, etc.) on RottenTomatoes: https://www.rottentomatoes.com/top/bestofrt/

Data Extracted:

For imdb.com, we extract the movie released from 1900 to 2019. We use the rules('video_url', 'video_title', 'video_year', 'video_level', 'video_genres') to parse related information and get 8925 tuples with 5 attributes from which we select 3000 tuples.
For rottentomatoes.com, we extract the information of top 100 movie in Best of Rotten Tomatoes. We use the rules('url', 'title', 'rating', 'genre', 'level', 'director', 'writer', 'time', 'runtime', 'studio') to parse related information and get 3010 tuples with 10 attributes.

Table Description:

IMDB.csv has the following attributes:

- video_genres [text]: genre with release date of the movie
- video_level [text]: rating of the movie
- video_tittle [text]: name of the movie
- video_url [url]: crawl url of the movie

- video_year [4-digit]: release year of movie

RottenTomatoes.csv has the following attributes:

- director [text]: director of the movie
- genres [text]: genre of the movie
- level [text]: rating with exact instruction of the movie
- rating [number of percent signs]: rotten tomatoes scores of the movie
- runtime [number with unit]: runtime of the movie
- studio [text]: studio of the movie
- time [month day, year]: release date of the movie
- tittle [text]: name of the movie
- video_url [url]: crawl url of the movie
- writer [4-digit]: writer of the movie

Note that not all attributes are contextually the same in both tables, for example, level, genres and year. Some tuples are dirty and we will process them in later stages. And some attributes might not be useful in entity matching stage, but are of potential interest for analysis in later stages.