CS 839 Project-2

by Vishal Agarwal, Roshan Lal and Pratham Desai

We extracted movie information from the following two sources:
Source-1: www.imdb.com/ (Internet Movie Database)
Source-2: https://www.allmovie.com/

Our Method:

IMDB

1. Obtained url from advanced search on imdb
2. Manipulated parameters like genre and num_votes to get a page with list of movies
3. Extracted movie urls and accessed the movie pages
4. Extracted information from the movie page

Allmovie

1. Obtained a list of webpage URLs for all movies from genre-specific webpages.
2. Removed duplicates in cases where one movie appears in multiple genre-lists
3. Using this list, crawled and downloaded ~8000 HTML pages containing movie details
4. Extracted information from the web pages

Entity:

As mentioned earlier, we extracted movie information from the above sources. The fields we extracted can be found in the section below.

Number of tuples from imdb: 7846
Number of tuples from allmovie: 7258

Schema:

Column Name	Description	Data type
Key	Unique id on respective source website	string
Name	Title of movie	string
Year	Year of release	int
Certificate	MPAA rating (adult, universal, etc.)	string
Runtime	Length of movie in minutes	int
Genre	Type of movie (comedy, romance, etc.)	pipe separated strings
SubGenre*	Type of movie (urban comedy, etc.)	pipe separated strings
Rating	Rating from respective source	float
Votes	Number of reviewers	int
Director	Name of director of the movie	string

Producer*	Name of producer of the movie	string
Country*	Countries where the movie was released	pipe separated strings
Gross**	Revenue in dollars	int
OtherRating*	Other movie rating on respective source	float
Metascore**	Movie rating from metacritic.com	int

* information not present on IMDB website (src1)
** information not present on Allmovie website (src2)

Open Source Tools:

- Python selenium chrome driver (headless mode) to access web pages from allmovie.com, we had to use a headless browser since some information was generated using JS
- Python package urllib to access web pages from imdb.com
- Python package BeautifulSoup to find html tags from the crawled web pages