CS839 Stage2 Report

Yunang Chen, Yuanfang Song, Hongqian Xia

1 WEB DATA SOURCES

We extracted movie information from these two websites:

- **IMDb** (Internet Movie Database, https://www.imdb.com): an online database that contains information related to films, including cast, production crew, plot summaries, trivia, and fan reviews and ratings.
- **Metacritic** (https://www.metacritic.com): a website that aggregates critics and publications' reviews of media products (including films).

2 METHOD

To extract movie information, we started with web pages where a bunch of movies have been structurally displayed. For IMDb website, we got the base URL by searching for feature films and sorting by US box office descending. For Metacritic website, we got the base URL by sorting the movie Metascores descending. Then we use python script to crawl data with the following steps:

1. From the base searching result page, pull the URLs of movies that appears in the result list.
2. For each movie page, send HTTP GET requests to get the source HTML code, find the class names and locations of the tags that contain interested information, apply parser to unwrap all the information, and save that in a dictionary. We also used a thread pool to parallelize this process.
3. Increase the page number in the base searching result URL repeat the above steps until there is enough data, and finally save the data into a .csv file.

3 DATA DESCRIPTIONS

3.1 ENTITY TYPE

Our entity type is **movie information**.

3.2 DATA SIZE

Table 1 Number of tuples in each table

	IMDb	Metacritic
Number of tuples	3094	3345

3.3 TABLE SCHEMA

Table 2 Descriptions of each attribute in the tables

Attribute	Description
id	The index assigned to each movie, starting from 1 for each table
title	The name of the movie
release year	The year in which the movie was released
rating	The rating of the movie (e.g. PG-13, R)
runtime	The length of the movie
genres	The genres of the movie
director	The directors of the movie
starring	The main cast of the movie
countries	The places where the production companies for the movie are based in
languages	The languages that are spoken in the movie
production company	The companies that produced the movie
writers	The writers of the movie, as appeared in the credit
score	The user/media score for the movie

4 OPEN-SOURCE TOOLS

To extracted data from websites, we used the following open-source packages

- **BeautifulSoup 4**, which is a Python library for parsing data out of HTML strings.
- **requests**, which is a Python HTTP library and is used to send GET requests to website.
- **pandas**, which is used to generated .csv files from Python's in-memory data structure.