Extracting Movie Information via Web Scraping*

Zihan Zhou, Kirin Hong, Mars Hao

April 10, 2019

Abstract

In this project, we'd like to extract comedy movie information by web scraping. We chose IMDB[1] and TMDB[2] as our data sources. Finally, we obtain two tables with the same schema.

1 Background

In this project, we'd like to choose movies as our desired entity since the structure of their information is very clear. There are many movie database websites that can be used in our project. Here we choose IMDB and TMDB as our data sources.

2 Methods

To ensure our results in the two tables could overlap each other for future projects use. We set restrictions with comedy and order by descent popularity. For each movie, the popularity in the two websites are not the same. Therefore, the final results in the two tables won't be all the same.

Firstly, we have to obtain the urls for all the movies that are qualified with our requirements. We extracted these links by web scraping in IMDB. While we have to get these information with api keys in TMDB. The obtained urls text documents are stored in our data folder.

Then we used two types of methods to scrape information from website. For the data from IMDb, we used the idea of HLRT to extract table with the Python built-in package re. We observed that most of the attributes in the table start from its name and end at the nearest '\n'(new line symbol).But for the full credits data, we have to go to the corresponding page and design several senarios to extract data since the format here is very irregular. For example, a table of director either ends at 'writting credits' or ends at 'cast'. After all the tables are extracted, data will be stored in pandas Dataframe.

For the data from TMDB, we used Beautifulsoup[3] package in Python to facilitate our procedure. It could provide idiomatic ways of navigating, searching, and modifying the parse tree and commonly saves programmers hours or days

*This is a CS 839 Project Stage 2 report in UW-Madison.

[1]https://www.imdb.com/

[2]https://www.themoviedb.org/

[3]https://www.crummy.com/software/BeautifulSoup/

of work. We can obtain desired texts from the source code by identifying its type and class. We can easily extract the desired information by finding out its specific pattern with the help of Beautifulsoup.

For the attributes from the two websites. There are various information about a movie, while some information are unique in a specific website(like scores). Therefore, we only selected some shared attributes in our projects to ensure they have the same schema. The final two tables are stored in our data folder.

3 Final results

Finally we obtained two tables about comedy movies. We have 5000 observations in Imdb.csv file and 4000 observations in TMDB.csv file. And both files have the same schema. For the attributes in these files, their meanings are as follows:

- name: the name of a movie
- year: the released year for the movie
- subtext: certification of the movie
- duration: how long the movie lasts
- genres: all the genres that could describe the movie
- director: directors of the movie
- stars: all the casts of the movie