CS839 Project Stage 2 Report

Pan Wu, Dongqiangzi Ye, Sek Cheong

Data Source

We extracted movie descriptions and review information from one on line movie review site and one movie retail website:

- https://www.imdb.com
- https://www.walmart.com

The IMDB contains information about movie reviews while the Walmart contains movie DVDs for sale. Both web sites contain information such as movie title, synopsis, genre, duration, actors/actresses, etc.

Entity

We collected movie information from IMDB and Walmart and obtained the following information:

Schema:

movies_IMDb(name, duration, genre, release_date, directors, stars)
movies_walmart(name, duration, genre, release_date, directors, stars)

Since both websites essential provide a lot of same movie attributes, we were able to come up with the common schema for both tables. The movies_IMDb table contains 3043 tuples and the movies_walmart table contains 3140 tuples. To make sure that we can contain mare than 100 same tuples in two tables, we extract all the movies from these two websites with same genres. To make sure we contain different tuples in these two tables, we also extract different genre movies in two websites.

Data Extraction Techniques

As mentioned above, both websites display the information with a structured HTML element. Specially, the IDMb website uses multiple to display the details about the movies, while the Walmart website uses <table> to display movie details. Based on the formats of HTML tags, we were able to extract the information we need.

Our initial web page(url) is a movies list page, which contains all movies of a same genre.

1. Firstly, we extract all the movie links in current page, put them in a links list and we extract all the movies details in this links list. Given a web page (url), our web crawler converts the web page into an HTML string. And we use BeautifulSoup to transform a

complex HTML document into a complex tree of Python objects which enable us to access the content easily by the query the tags. For every movie page, we extract the name, the duration, the genres, the release date, directors, and the stars.

2. Secondly, we extract the next page link from the current page.

3. Do step 1 and 2 iteratively until we get more than 3000 tuples.

Open Source Tools Used

We use the BeautifulSoup python package. BeautifulSoup allows you to pull data out of HTML and XML files. It provides simple methods for navigating, searching, and modifying the parse tree.