David Kimmel

SI 206

Final Project Proposal

4/11/18

For my final project, I will be web crawling and scraping the webpage IMDB (international movie database), a website that contains all information regarding movies, television shows, and other entertainment (http://www.imdb.com/). This information includes all movies, all television shows, all actors and actresses, directors, etc. For my 4 data sources, I will be crawling through the IMDB webpage and scraping it for data on movies. Specifically, my project will retrieve data from the 250 top rated movies rated by IMDB users. From the data I retrieve, I will be extracting the movie titles with the years and the number ranking based on IMDB users, movie descriptions, movie ratings, movie genres, IMDB ratings out of 10, list actors in the movies (first billed actors only) or just the main actor/actress in the movie and movie directors. Because I have never worked with IMDB before and I will be web scraping and crawling multiple pages of this website, my project will have a challenge score of 8.

From the data I collect scraping and crawling IMDB, I will create a database with 3 different tables. One table with contain each movie with information for each movie, another table will contain the directors, and the third table with contain actors and actresses. There will be a category in the movie table that gives each movie a unique identifier and this will be the primary key. The other two tables will have unique identifiers with each director and each actor/actress as the foreign keys. These tables will then be presented using an interactive command line to prompt the user into retrieving 4 different graphs, all being generated using plotly. I’m not a hundred percent certain how I am going to display the data, but thinking towards displaying number of each genre in the top 250 movies, number of times actors/actresses appear in these 250 movies (displaying the top 5 or 10), and then two other displays that I still need to specify.

\*\*\*\*The screenshot I have attached displays how my code is able to retrieve the movie titles from IMDB and displays my ability to crawl through imdb to get these movie titles. \*\*\*\*