Python Project 2 – Web Scraping

This project consists of web scraping for top 500 most popular movies on the website - https://www.metacritic.com/. We would be fetching primary information like movie name, directors, cast etc. from the website. These information are a stored onto a .csv file, sqllite3 database and are also stored locally. Using the locally stored data vitals information for any movie in the top 500 movies can be displayed and not only this but a few manipulations are performed to calculate the cosine similarity between directors, actors depending on the common supporting cast that they consist of.

Director Analysis – we perform a function on all the directors and return most successful directors which are decided depending on the number of movies that they are involved in. In our case the with an initial threshold of 4 movies we got a result of 20 directors. As we increase the threshold to 6, we find that the number of directors in that category reduces to $1/5^{th}$ of its original of 20. And Finally, after all the filtering possible standing strong with more than 9 movies in the top 500 movies on Metacritic, Alfred Hitchcock is on the No. 1 position in the successful director's table. He has directed totally 12 movies making it a probability that every 3 in 125 movies in the top 500 movies is directed by Alfred Hitchcock.

Actor/Actress Analysis – Here we perform a function on all the cast from all the top 500 movies on Metacritic website and return most successful actors/actresses who are decided depending on the number of movies that they are involved in. In our case the with an initial threshold of 4 movies we got a result of 90 actors/actresses. As we increase the threshold to 6, we find that the number reduces to 15 which is $1/6^{th}$ of the initial value. Additionally, in the top 15 successful the Actor to Actress ratio is 13:2 even after the fact that each movie's cast consists of a good number of actresses. According to the data of the most appearing actors and actresses that we have scraped and analyzed for the top 500 movies we can say that no actor/actress has been a part of 9 or more movies and 8 being the maximum number of appearances which are achieved by Harvey Keitel & Robert de Niro bringing the most successful Actor to Actress ratio to 1:1.

General Analysis – of the 500 movies information that we scrapped from Metacritic website most of the movies have one director and some of the movies have more than 1 directors, the ratio of Movie to Directors is 1:1.18. The total count of cast including the general cast and the principle cast in the top 500 most popular movies is 8049 meaning on an average there are 17 cast members in each movie. Since the actor/actress appearance chance is only 90 for a threshold of 4 movies this points to the remnant of the cast members of 7959 which are in the range of 1-3 movies. This showcases the fact that only (90/8049 * 100) ~ 1% percent are the successful actors/actresses and rest 99% of the entire cast members of the top 500 most popular movies are still struggling to get good opportunities to give their best and to be a part of that 1% successful actor/actresses.

Moving ahead with the analysis of growth in director's careers even though the ratio of movies to directors is over 1 it does not clearly indicate reappearances of directors in different movies and is just an indicative of multiple directors for movies. In the 411 directors from the top 500 movies just 20 of them appear in more than 4 movie's information meaning (20/411*100) ~ 5% pointing to the fact that still 95% of the directors lack support and opportunities.

Conclusion – We can conclude that majority of the top 500 most popular on Metacritic website movie's directors and cast are still in search of good opportunities to showcase their talent.