**Stage 2 Report**

Team members: Abbinaya Kalyanaraman, Bob Effinger, Rajan Dalal

Team number: 9

# Web Sources

For this project we decided to gather information on movies. We gathered this information from the following two web sources:

# Metacritic.com

Metacritic aggregates reviews of movies from leading critics. They have a sizeable collection of movies on their website that we decided to extract based on scores.

# imdB

imdB has a page that lists movies based on popularity, with the release year, along with some information about each movie. By extracting the information based on popularity, we can gather a sizeable amount of movie data.

We decided on these 2 web data sources after browsing through various other websites which provided movie information. We looked for 2 web data sources that would provide a large number of common attributes and then we selected the set of attributes that we wanted to extract from the two sources.

# How We Extracted Data

* 1. **Metacritic.com**

To extract structured movie data from Metacritic, the data was initially sorted based on scores.

The first step we took towards extraction was removing the head and tail from the html data. This is so we would have a more focused view on the html files and to make it easier to find the data that we wish to extract. After eliminating a lot of the extra information from our html we then extracted each of the attributes that we were looking for.

For Metacritic we found the attributes

Movie Title

Release\_Date which we changed to Release\_Year to match the imdB data

Meta\_Rating which we removed because it did not match the imdB data

Movie\_Rating

Movie\_Summary

We extracted these by using the regrex package that is built into python.

Meta\_Rating and Movie\_Rating had a few different patterns that could be matched, and so these tags were converted to a single format before extracting the tags. But after they were matched to a single format the values were extracted through basic regex extraction.

# imdB

To extract structured movie data from imdB, the data was initially sorted based on popularity.

# Entity and Table Description

As mentioned above, we decided to extract the entity movies, specifically those released in the last three years.

# Table A - Metacritic

The data from Metacritic contains the movie’s title (String), release year (String), meta rating (Int), movie rating(Float) and summary(String), consisting of 3100 tuples. In order to maintain a common schema with Table B, we decided to remove meta rating from the table.

# Table B - imdB

The imdB data also contains the movie’s title (String), release year (String), meta score(Int) , movie rating(Float), run time(String), age rating, categories(String) and summary(String), consisting of 3250 tuples. In order to maintain a common schema with Table A, we decided to remove meta score, run time, age rating and categories.

Thus our two tables have a common schema which comprises of the movie’s title, release year, movie rating and summary.

# Open Source Tools

We did not use any open source tools for the project but made use of regrex to extract the data, with the help of the re library in python.