# SI 506 Final Project Proposal

Jiaming Yang

## Description

I will crawl and scrape the information of IMDb's top 250 movies to explore the commonalities of those great films. For example, what genres and rating are the most popular ones? Who are the directors, writers, and stars that created the world's most classic films? The target audience would be everyone who is interested in watching films, who is willing to know about what those great films have in common.

## Data Source

The starting point will be the IMDb top rated movies chart. For each movie, I need to crawl one level deeper to get the information about movie's rating, length, genre, and release date. Afterwards, I will need to crawl another level deeper to get the list of the movie's directors, writers, casts, and producers.

The starting page is https://www.imdb.com/chart/top?ref_=nv_mv_250.

Because I will be crawling multiple pages in IMDb, which is a site that we have not used before in SI 507. The challenge score will be 8. The requirement will be satisfied.

## Presentation Option

I am planning on providing an interactive command line prompt for user to choose date/visualization options. The program will support the plot of:

1. Line chart of how the number of top 250 movies grow every year.
2. Histogram of the number of movies in different genres.
3. Histogram of the number of movies with different ratings.
4. Histogram of the number of movies with different lengths.
5. Histogram of the people who directed/wrote/produced/appeared in the most top 250 movies.

## Presentations Tool

I will be using plotly as the tool for data visualization. Because plotly is sufficient to complete the five tasks mentioned in the previous section.