Capstone 2: Initial Project Ideas

First idea: Network Analysis Using Clustering Models for Movie Casting

Filmmaking can be a very uncertain business with many moving parts. A production company must mediate between a great assortment of personalities and artistic visions. Few tools are available to help them decide which people will have the best chemistry for making a particular film.

This project will construct a graph composed of many types of movie workers. This graph will be used to find the best clustering model for grouping these workers based on movie revenue figures. This model can be used to predict the most profitable crew to hire for future works. These choices can be refined depending on various movie characteristics, such as genre and budget.

The dataset for this project has already been acquired through API requests from the TMDb database.

Second idea: Tree Regression for Predicting the Success of Actor's Movies

A properly fitted linear regression model can be a good baseline for making accurate predictions on movie data. Inadvertently, the financial part of this data consists of many outliers. Other machine learning models, such as regression trees, are more robust to these outliers.

This project will perform a thorough search for the best regression tree model for making predictions about the levels of success of movies, given a particular actor's participation. This model will assist the people who hire actors toward making the best casting choices with respect to the characteristics of their productions.

The dataset for this project has already been acquired through API requests from the TMDb database.

Third idea: Sentiment Analysis of Film Reviews to Determine Revenue Influence

Before the internet age, the source of film criticism was mainly in the hands of the professional critic. These people had great influence when it came to a moviegoer's decision to buy a ticket for a particular film. Today, the number of available opinions that can be considered before seeing a movie is exponentially larger. With so many user reviews posted on websites, it is important to ask if these reviewers have taken the place of the traditional critic in determining the success of a film.

This project will use Natural Language Processing techniques to classify movies based on the text of professional critic reviews and those of the audience. After each movie has been classified using these two corpora, the classifications will be compared to the revenues of the movies to determine which group is better suited to predict the financial success of a film.

The dataset for this project has already been acquired through API requests from the TMDb database, which will be supplemented by data from user reviews from an IMDb dataset.