***An Examination of Movies***

**Abstract:** We wished to take a deeper dive into movies and film history. Intuitively, we all have an understanding of what movies we enjoy and don’t enjoy as well as at least a cursory level of knowledge with film history and the mass of movies that have been made. The means of investigation detailed in this paper are largely through the use of visualizations: with graphs and tables. We also created a model which recommends movies similar to an input movie. This is achieved by itemizing details of a film (director, genre, description, etc) and finding its cosine similarity compared to other films. The lower the cosine similarity, the more similar the films are and, therefore, it’s likely that a person who enjoyed the input movie would enjoy the output movies as well.

**Introduction:** We’ve examined the IMDb movies extensive data set from kaggle (<https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>) in order to acquire a better understanding of the landscape of film that’s available for viewing. Movies are a big part of our lives but we have a surprisingly underdeveloped method of finding movies we may enjoy, or understanding aggregate trends in the over-century long history of filmmaking.

By leveraging this dataset, we visualized and analyzed the movies and ratings as well as the people involved in making the films in a more consistent, well-founded fashion.

**Data:** The dataset linked above, the IMDb movies extensive data set, includes 4 tables: movies, names, ratings, and title principals.

The movies table includes details about the movie itself such as the title, genre, and description as well as information about its production such as the production company, who directed it and who wrote it. It also contains details about the reception it received from both critics and IMDb reviewers, and financial information such as budget and revenue (categorized as USA gross income and worldwide gross income).

The names table itemizes the people who work on films. It also includes personal data such as birth date and location, death date and location (if applicable), and spousal details, even divorce. It also mentions any children they may have.

The ratings table itemizes ratings for each movie in the movies table and has some derived statistics: median rating and mean rating. It also has a wealth of information on reviews based on splits of demographic conditions of age and gender. So a movie may have an average rating but do very well among women under 18, yet poorly with men over 45.

Finally is the title principals table which breaks down the credits of each movie and lists out who had what role in making the film. It includes actors and actresses but also crew members such as the cinematographer, director, production designer, and producer.

**Model:** With information taken from the movies database. We analyzed each film’s cosine similarity to each other in order to make a movie recommendation system. The steps done to achieve this were to create the feature space of name, genre, actors, director, and description of the film, then convert this into a matrix of token counts. From there, the cosine similarity gives a good estimation on the similarity between two films.

**Analysis and Results:** With the help of the model, any film can be input and the most similar films will be returned. For instance, if Pride and Prejudice is put in, then the model will return La Femme de Nulle Part and The Rainbow Jacket as the two most similar films. Separate from the model, visualizations helped glean information about the landscape of movies. We now have answers to simple questions such as “who directed the most movies?”, “who has the best reviewed catalog?”, and “what movies made the most money?”

**Future Research Directions:** As with any model, some improvements can be made in order to adjust for a variety of factors such as only including movies in one particular language or specifically returning movies that have a budget above a certain amount. Furthermore, new movies are going to be made which may affect the top cosine similarity movies for films already released. The model would have to be updated with fresh data on a regular basis in order to achieve accurate results and stay contemporary.