Project 4 Group 18

Ethan Tebbe
Jorge Arroyo
Ben Rosensweig
Hassan Khan

Movie Recommender

Introduction

For our capstone project we had the task of creating and bringing together almost all of the skills we had acquired throughout our six month bootcamp and applying it to a web app that utilized machine learning. Our group decided that the best approach to this would be through a recommender that borrowed from both an IMDB movie dataset, and an API spanning movies from 1920 to 2024. This project required our team to evaluate our own individual strengths and approach our tasks accordingly, ultimately aligning our individual work together to create our final machine learning web app.

Research Questions

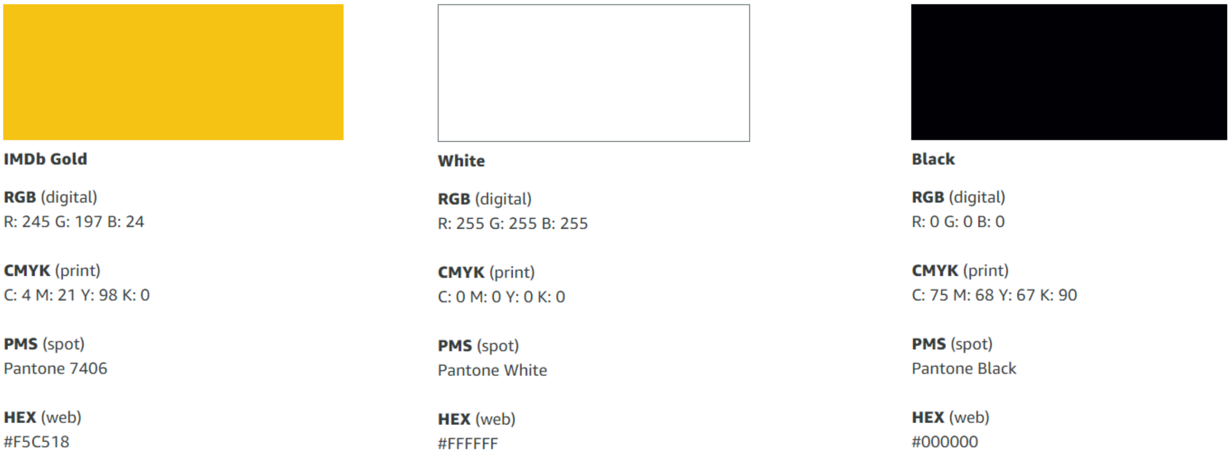
1. Is there a relationship between movie rating and box office sales?
2. Do certain directors tend to produce higher-rated movies?
3. Are movie ratings trending upward or downward over the years?
4. Are certain genres becoming more popular over time?

Data Cleaning

We combined two datasets, one from Kaggle and one from the TMDB API. The dataset comprised several features. For the recommender, we focused on overview, director, cast, genre. While the main focus of our project was not on data cleaning, we did have to do some. We dropped null values in gross (box office), and converted data types in runtime. We also needed to drop all but one string in genre, as there were many rows that had several listed genres. After cleaning the original csv, we were left with about 850 rows of data. We wanted to incorporate more movies to get a broader dataset for the recommender, and used the API TMDB to obtain 2,400 more rows of data with a final set of around 3,400 rows.

Color Design

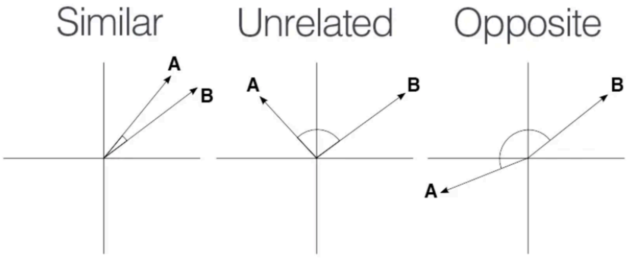
Since the dataset was based on a list of movies on IMDB, we decided to focus on that color palette, particularly gold, black and white.



IMDb Gold	**White**	**Black**
RGB (digital) R: 245 G: 197 B: 24	**RGB** (digital) R: 255 G: 255 B: 255	**RGB** (digital) R: 0 G: 0 B: 0
CMYK (print) C: 4 M: 21 Y: 98 K: 0	**CMYK** (print) C: 0 M: 0 Y: 0 K: 0	**CMYK** (print) C: 75 M: 68 Y: 67 K: 90
PMS (spot) Pantone 7406	**PMS** (spot) Pantone White	**PMS** (spot) Pantone Black
HEX (web) #F5C518	**HEX** (web) #FFFFFF	**HEX** (web) #000000

Machine Learning Recommendation System

We used NLTK (the Natural Language Toolkit Python Library) to process the overview of the movie. We used Term Frequency Inter Document Frequency to generate vectors for each movie. With those vectors, we generated a matrix of similarities using cosine similarity. From the matrix of similarities, we are able to generate a list of recommendations via sorting the list. By using NLTK for text processing and then Term Frequency-Inverse Document Frequency to create vectors, we were able to construct a similarity matrix through cosine similarity. This method allowed us to first identify, and then recommend movies that share similar thematic elements based on user input of any one given film title.



Dashboard & Tableau Design

The Dashboard and Tableau utilize colors and fonts incorporating the IMDB color palette. While the Dashboard, specifically the Movie Recommender page, was the centerpiece of the machine learning portion of the project, the Tableau focused on ratings, box office sales, rating vote counts, and genre besides the obvious inclusion of film title, main and supporting actors, and directors. The Tableau was designed to give a broad scope of visuals and information, while the Dashboard is simplified and concise.

Call to Action

While the recommender's scope did not include all parts of the original research questions, it did provide some interesting insights.

5. Is there a relationship between movie rating and box office sales?
6. Do certain directors tend to produce higher-rated movies?
7. Are movie ratings trending upward or downward over the years?
8. Are certain genres becoming more popular over time?

Limitations and Bias

Our limitations were:

- The dataset was last updated four years ago.
- Given the amount of movies that exist, this dataset will not encompass enough to give a universal representation of all movies and will be leaving out a lot of movies.

Our biases were:

- Ratings could be biased towards the preference of those who review the movies rather than the overall preference. In other words., those that feel the urge to write a review for the movie might have a different opinion on the movie than the general public.

Future Work and Conclusion

Future work could include creating a more robust system of inputs and outputs for the recommender. Instead of the output being just the film and rating, it could include the release year and overview. We could also update the recommender for better error handling to manage inputs where the movie title is either misspelled or not found.

Works Cited

- DataSet: https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows/data
- Slide Deck Template: https://slidesgo.com/theme/movie-awards-ceremony#search-movie&position-4&results-232&rs=search
- Color Palette: https://www.color-hex.com/color-palette/4236, https://brand.imdb.com/imdb