Tre’on Russell
Caleb Collins
Thomas Macdougall

Final Project Proposal

Introduction

With more media available to us than ever it's easy to get lost in the sheer amount of content that we're getting each year, and the backlog only gets larger. Each year has many classics and amazing projects; however, many fall short and as a group we would like to understand why. We will focus on movies and use a dataset that covers the lowest rated movies and with that we will apply clustering to see which factors may have led to lower review scores for the films. This could be useful for analyzing market trends in taste of the general population, which will be great for the more risk averse as they could see their likelihood of success.

Data

https://www.kaggle.com/datasets/octopusteam/imdb-top-1000-worst-rated-titles

We are using a dataset from Kaggle called “IMDB Top 1000 Worst Rated Titles”. This data set goes all the way back to movies from 1957 up until 2024. The featured columns include a unique IMDb ID value, the title of the movie, the genre it falls under, the average rating, the number of votes and the release year. This data set is simple but makes up for it with the large number of entries which will allow us to zero in on the data and produce more in-depth results.

Methodology Plan

What we will do is use clustering to identify patterns in the data set relating to the genre and the user score. We will also use the release date to track trends over time. There have been conversations about movie quality going down, the fall of certain genres and people being nostalgic for certain genres that were more prominent in the past. We want to use a KNN graph to find out whether movies are generally getting worse or if that's just nostalgia talking. We will see if there's a correlation between the lowest rated movies and the number of votes as well to determine if it's just a bunch of B movies being rated low or if there is merit to the claim.

Evaluation Plan

What we want to answer is if the genre correlates to low review scores and if movies are getting worse over time. We can examine if certain genres perform lower than others by using clustering and a KNN graph. We can check which genres had the greatest number of entries and which year or set of years had the highest number of entries in the lowest review movies category. We can determine if more movies were rated lower in the past by checking how many entries each year, decade, or other set of time had. The graphs used can be color coded and shown over a period of time using a line graph as well. A pie chart can be used to determine which genre took the largest chunk of low score reviews.

GitHub Repository and Initial Tasks

Final Project GitHub Link

https://github.com/TARus-1/ITCS-3162-002-Final-Project

Initial Tasks

- Data Preprocessing/Cleaning - [Caleb Collins] (Due November 30th)
- Analyzing Data to Determine Trends/Pre-Clustering - [Tre'on Russell] (Due December 2nd)
- Clustering and KNN - [Tristan Jarvis/Other] (Due December 5th)
- Visualizations - [Thomas Macdougall] (Due December 7th)
- Upload to GitHub - [Group/All] (Due December 8th)
- Presentation Slides/Formatting - (Due December 11th)

Group Expectations

We will encourage each other to remain close to deadlines and communicate with each other in case one of them cannot be met. If someone is unable to complete their specified portion(s) of the assignment and/or we are unable to reach them, we will reach out either to the TA or professor for intervention if possible. After which if we are unable to reach a resolution, request for their grade to either be lowered or for them to be omitted from the project.