IST652 Scripting for Data Analysis

Netflix Movie and TV Shows Analysis

Final Project Report

Group 2 Members:

Colton Moyer

Ritesh Sabhapati Verma

Objective

The goal of this project was to analyze data from Netflix to understand trends relating to different countries and how data is used to target the interests of clients in those countries. For relaying the process in the writeup, we did not subsample and compare between countries, but this would be easy to do by simply extracting rows with 'country', running parallel analyses, and comparing visuals.

Data Description

To do this we downloaded data from Kaggle, a data sharing website: Netflix Movies and TV Shows (kaggle.com). This data was provided in a structured, .csv format and is composed of columns of strings and vectors. Most columns are categorical variables such as country shared within, name of director, name(s) of the cast, and title, however, there are also integer variables such as duration and release year. Each row is a different show or movie, and there are 8,807 rows and 12 columns.

Data Preprocessing

Preprocessing the data for analysis was minimal but included converting the data frame from .csv format to reducing the dataset to only countries of interest and removing rows with missing values (NA's) from columns used for analysis, such as those that do not list a director.

Methods of Analysis

The following is a list of research questions we sought to answer:

1. What are the top 5 actors and directors for content on Netflix?
2. When was most Netflix content released and what time period were the most movies and shows released? When was/is the peak of streaming?
3. Do users prefer content with positive or negative sentiment?
4. How does Netflix use artificial intelligence to recommend additional content based on user input?

Analyzing the top 5 Actors and Directors on Netflix:

Our analysis started by removing NAs from the data frame. We first created an empty data frame and split the cast column from the original to create multiple columns for each cast member using Pandas and Numpy. In some instances we no cast or director was specified. We used the .fillna command in pandas and filled these columns with 'No cast specified' or "No director specified' within the respective columns, and removed instances where none were listed. We counted the number of instances each actor was mentioned in the data frame, after confirming that there were not variations in spelling or capitalization between mentions of actors or directors. Then we plotted the top five actors and directors respectively using bar charts using the

package ‘plotly’. The top actor across all of Netflix was Anupam Kher and the top director was Marcus Raboy.

Analyzing the content produced on Netflix based on each years:

To visualize when Netflix content was released, we subset the data frame by the columns content ‘type’ which was of categories either ‘movie’ or ‘show’ and the column ‘release_year’. We then used the ‘groupby’ function to count the number of shows and movies released in each year. This analysis would not account for shows that had additional seasons in subsequent years. For instance, ‘Survivor’ is a current show (2024) on Netflix and has had over 50 seasons but first aired in the early 2000’s. We then subset the dataset to only data after 2000 because there is significantly fewer digital data available prior to then. We used the px.line command to create a line graph for the number of shows in each year and colored each line by ‘type’ to create a line for movies and tv shows respectively using ‘plotly’. The highest number of movies uploaded to Netflix was in 2017, and tv shows peaked around 2020, however, since this data was compiled in 2021, it is likely not complete for 2021 and is probably higher today.

Sentiment Analysis of Netflix Content:

To analyze the titles of tv shows and moves for their sentiment we used textblob. We wrote a loop to iterate through the descriptions of each row and create a textblob object, then used the function sentiment.polarity to see if descriptions tended to be positive, negative, or neutral based on their polarity value. We then grouped the data frame by the release year and sentiment values and filtered out rows prior to 2005. We used the px.bar function in plotly to plot a bar graph of sentiment in each year and stacked sentiment. Descriptions of content were composed of mostly positive sentiment in every year which tracked the amount of content that was released in each year (question 2).

Content-Based Recommendation System:

To assess how Netflix uses content-based recommendation to improve their user’s experience and suggest additional content for users to watch, we installed the package ‘re’ for regex expressions and the artificial intelligence package ‘sklearn’. We first replaced all instances of hyphens, colons or apostrophes with spaces using the apply function.we created a TfidfVectorizer object using fit_transform on the description column of the data frame and returned the feature names as a matrix, which we turned into a numpy array. We used the cosine_similarity function to generate values of similarity between feature names in a matrix and generated recommendations based on a similarity index from the array, which works as a library. Through inputting the name of a movie or show, a function calls its index, which retrieves the

numpy array of similarity values (based on their descriptions) to everything else in the dataset, which is put in a dataframe, and sorted based on score, and returned by title. For “Peaky Blinders” the program returned the five highest similarity scores of “Our Godfather”, “My Stupid Boss”, “Don” and “Jonathan Strange & Mr Norrell” meaning these five listings have the most similar descriptions among everything else on Netflix, which is what Netflix would use to visualize the “Since you watched Peaky Blinders:” command after finishing a title. I did a similar call for “Tiger King: Murder, Mayhem and Madness” and the program returned “Club Friday To Be Continued - Friend & Enemy”, “Deep”, “Hell and Back”, “Aggretsuko: We Wish You a Metal Christmas”, “The Son”, and “The Life Ahead”. The first return is a drama from Thailand released in 2016 about a live triangle that gets out of control! So, the program works!

Conclusions

Our results suggest that despite being a global company that has projects all over the world, some of the largest contributors are English directors and actors. Marcus Raboy was most abundant director on the list, he is the video director for many large American musicians and singers such as The Dixie Chicks, Rihanna, and Shakira. He is probably at the top of the list because of music videos or documentaries about some of the musicians he works with. Surprisingly, movies seemed to peak on Netflix in 2017. This might be due to more content becoming series, and fewer people opting to view content from home rather than at movie theatres.

Movie series such as Star Wars are beginning to transition into show series composed of more episodes of shorter length. We can see in the line graph that the number of movies uploaded appears to have dropped off but tv shows are still increasing in quantity (Image 1).

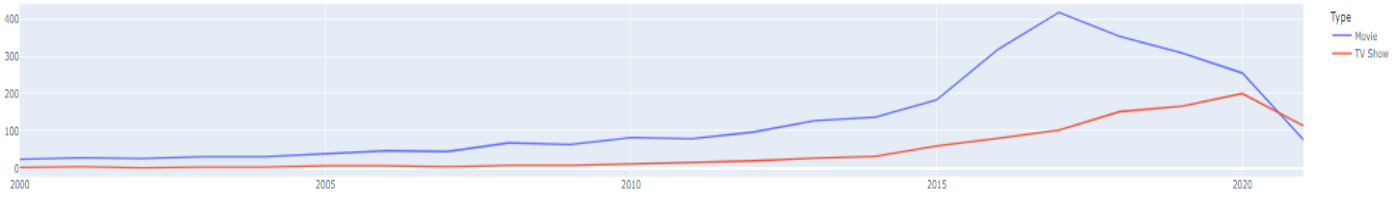

Image 1

It appears most viewers appreciate movies that have positive sentiment. Most descriptions do not have neural sentiments, this makes sense because content is supposed to illicit an emotional reaction, and what would be the point of it otherwise.

The recommendations program is very useful for understanding likeness between different video works on Netflix. It was surprisingly easy to create a program to analyze descriptions of movies

and it is interesting that one of the central features to a huge streaming service is simple to run on a local computer given a semester's worth of python experience.

Group Work

Both members developed the idea for the Netflix project including the questions that could be answered using the data we had collected. Ritesh developed most of the python program for analyzing the Netflix data and created some of the slides for the presentation. Colton prepared some of the slides for the presentation and wrote up the description of the project for the report. The overall workload was shared equally.