**Structured (Netflix) Data**

Shefali Kamalnakhawa

Seidenberg School of Computer Science and Information Science, Pace University

CS 619: Data Mining

Charles Edeki

April 30, 2021

**Structured (Netflix) Data**

**For Structured Data (NETFLIX)**

Flixable is a search engine for video streaming services that offers a complete list of movies and shows streaming on Netflix. The search engine released the Netflix Movies and TV Shows data set, which includes the complete list of movies and shows available in 2019.

In this project, we will perform exploratory data analysis on Netflix Movies and TV Shows data set.

# Abstract

For instance, in movie recommendation, we have reviews from different people on different movies. However, it is highly unlikely we have each reviewer review all of the movies in our dataset. This leaves an important and difficult question in analysis. If a reviewer hasn't reviewed a movie, is that an indication that they aren't interested in the movie (and therefore wouldn't recommend it) or simply that they haven't reviewed it yet? Understanding what content is available in different countries. Identifying similar content by matching text-based features. Network analysis of Actors / Directors and find interesting insights. Is Netflix has increasingly focusing on TV rather than movies in recent years.

*Keywords:* social media, netflix, network analysis, movie recommendation.

**Methods and Models**

**For Structured Data (NETFLIX)**

First, let’s import Pandas and read the data into a data frame then print the first five rows of the data. We can see that there are several categorical columns. Let’s define a function that takes as input a data frame, column name, and limit. When called, it prints a dictionary of categorical values and how frequently they appear.

Apply our function to the ‘country’ column and limit our results to the five most common values. Next, it would be useful to generate summary statistics from numerical columns like ‘duration’. Let’s define a function that takes a data frame, a categorical column, and a numerical column. The mean and standard deviation of the numerical column for each category is stored in a data frame and the data frame is sorted in descending order according to the mean. This is useful if you want to quickly see if certain categories have higher or lower mean and/or standard deviation values for a particular numerical column.

We will use boxplots to visualize the distribution in numeric values based on the minimum, maximum, median, first quartile, and third quartile. If you are unfamiliar with them, take a look at the article Understanding Boxplots.

Similar to the summary statistics function, this function takes a data frame, categorical column, and numerical column and displays boxplots for the most common categories based on the limit.

***Affinity analysis for recommendations***

Affinity analysis is the task of determining when similar concept movie watched by person. In this project, we focused on whether the objects themselves are similar - in our case whether the TV shows or Movie were similar in nature. The data for affinity analysis is often described in the form of a transaction. Intuitively, this comes from a transaction at a store—determining when objects are purchased together as a way to recommend products to users that they might purchase.

Affinity analysis is usually much more exploratory than classification. At the very least, we often simply rank the results and choose the top five recommendations (or some other number), rather than expect the algorithm to give us a specific answer. Furthermore, we often don't have the complete dataset we expect for many classification tasks. For instance, in movie recommendation, we have reviews from different people on different movies. However, it is highly unlikely we have each reviewer review all of the movies in our dataset. This leaves an important and difficult question in affinity analysis. If a reviewer hasn't reviewed a movie, is that an indication that they aren't interested in the movie (and therefore wouldn't recommend it) or simply that they haven't reviewed it yet? Thinking about gaps in your datasets can lead to questions like this. In turn, that can lead to answers that may help improve the efficacy of your approach. As a budding data miner, knowing where your models and methodologies need improvement is key to creating great results.

***Recommendation Systems and the inherent challenges***

Making better recommendations leads to better sales. When online shopping is selling to millions of customers every year, there is a lot of potential money to be made by selling more items to these customers.

Product recommendations, including movie and books, have been researched for many years, however, the field gained a significant boost when Netflix ran their Netflix Prize between 2007 and 2009. This competition aimed to determine if anyone can predict a user's rating of a film better than Netflix was currently doing. The prize went to a team that was just over 10 percent better than the current solution. While this may not seem like a large improvement, such an improvement would net millions to Netflix in revenue from better movie recommendations over the following years.

# 

# Conclusion

Analyzed *Netflix Movies and TV Shows*data set to define the functions for generating summary statistics like the mean, standard deviation, and counts for categorical values. We also defined functions for visualizing data with boxplots and histograms. Here we created a movie recommendation system to which will help the customers in creating a user experience that will seek to improve retention rate, which in turn translates to savings on customer acquisition.

To give the customers advantages in two perspectives — 1) As a user, it is more coherent when presented a row of items that are similar, and then decide if he or she is interested in watching something in that category, 2) As a company, it is easier to collect feedback as a right-scroll on a row would indicate interest whilst a scroll-down (ignoring the row) would indicate non-interest (not necessarily irrelevance).

# References

Zhang. J, Wang. Y, Yuan. Z, & Jin. Q, (April 2020),*"Personalized real-time movie recommendation system: Practical prototype and evaluation"*, *Tsinghua Science and Technology*, pp. 180-191, <https://ieeexplore.ieee.org/document/8821512>.

Ortega. F, Rojo. D, Valdiviezo-Diaz. P, & Raya. L, (2018), *"Hybrid Collaborative Filtering Based on Users Rating Behavior,"* in IEEE, pp. 69582-69591, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8533607> .

# Zhou. Y, Wu. J, Chan. T. H, Ho. S, Chiu. D, & Wu. D, (August 2018), *"Interpreting Video Recommendation Mechanisms by Mining View Count Traces,"* in IEEE, pp. 2153-2165, <https://ieeexplore.ieee.org/abstract/document/8170284>

Madanapalli. S. C, Gharakhieli. H. H, & V. Sivaraman, (2019)*"Inferring Netflix User Experience from Broadband Network Measurement,"* IEEE, pp. 41-48, <https://ieeexplore.ieee.org/document/8784609>.