Netflix Movies and TV Shows Clustering

V. Bhavya Reddy

***Abstract*—This project aims to develop a Netflix movie recommendation system based on similarities within movies/ TV shows. The purpose of this study is to analyze a dataset from the Netflix database in order to analyze the similarities among the movies/ TV shows. We have all found ourselves scrolling endlessly looking for what to watch so this project might help with that. There is no information provided about the preferences of users, so what would be one's recommendation after watching a certain movie is what we are estimating.**

***Index Terms*—K-Means clustering, Natural Language Processing, Topic Modelling.**

I. Problem Statement

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

In this project, you are required to do:

- Exploratory Data Analysis
- Understanding what type content is available in different countries
- Is Netflix has increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features

A. Attribute Information

- show_id : Unique ID for every Movie / Tv Show
- type : Identifier - A Movie or TV Show
- title : Title of the Movie / Tv Show
- director : Director of the Movie
- cast : Actors involved in the movie / show
- country : Country where the movie / show was produced
- date_added : Date it was added on Netflix
- release_year : Actual Release year of the movie / show
- rating : TV Rating of the movie / show
- duration : Total Duration - in minutes or number of seasons
- listed_in : Genre

II. Introduction

Netflix has TV shows and movies which can be viewed online anytime. The platform is a monthly subscription service and customer subscriptions can, however, be canceled at any time. The platform must therefore keep users interested and keep them hooked on it. Providing users with valuable recommendations is vital.

A Netflix recommendation system increases revenue for the company by recommending what users will enjoy watching. In addition, it gains the loyalty of users, offers a wide selection of films, etc. The user can make a more informed decision when he or she has a wide variety of movie recommendations at their disposal. Streaming services provided by Netflix are the most popular in the world and have the most value in the Internet broadcast industry. A digital success story about Netflix would not be complete without mentioning Netflix's recommended systems. A list of recommendations can be created in several ways based on your preferences.

In Netflix Recommendation model, Netflix movies and TV shows are recommended based on a user's favorite movies and TV shows. Given that no personal data is provided, we can only find a recommendation from one movie to another idividually. Natural Language Processing (NLP) and K-Means Clustering are used in this project to make these recommendations. Movie and TV show suggestions are based on plot descriptions, type, cast, genre, etc. To gain a better understanding of recommendation systems, this project will develop a model capable of clustering similar material based on numerical and categorical attributes.

III. Approach

For the given problem statement, understanding what type content is available in different countries - I used Topic Modelling and several plots. For Clustering similar content by matching text-based features - I applied stemming and tf-idf vectorizer on movie plot description. K-Means clustering was also used for clustering.

A. Tools

Python was used throughout the project, in Google Collaboratory. For analyzing and visualizing the data, as well as building the Netflix cluster model, the following libraries were used.

- Pandas : Manipulates dataset.
- Numpy : Array and vector manipulation.
- Matplotlib : Visualising the data.
- Seaborn : Visualising the data.

- Warning : For ignoring the warnings.
- Sklearn : Machine learning library to perform several classification, regression algorithms.
- nltk : Natural language programming libraries.

B. Handling Null values:

In general, null values are dropped, but in this case, there are many null values, and dropping all of them would result in an imbalanced dataset. So, the only null values that were dropped were from rating and date_added features. No director, No cast, and country not available were replaced with null values in director, cast, country features respectively.

C. Exploratory Data Analysis

Analysis of the dataset using pandas, numpy, and statistical methods is an important process to understand the variety's key characteristics. This is exactly what EDA does. We can use it to understand the given dataset, clean up the dataset, identify outliers, identify essential features, and provide a clear picture of features and their relationships. This is a time-consuming process and the following activities are performed on the datasets.

- Proportion of type of content.
- Ratings for Movies and TV Shows.
- Heatmap for date_added and year_added.
- Total content added across all years (up to 2020)
- duration for TV Shows seasons and Movies.
- Top 20 countries on Netflix.
- Top 20 genres on Netflix.
- Top 10 director on Netflix.
- Top 10 cast on Netflix.
- International content across all years (up to 2020).

D. Understanding what type content is available in different countries

As a first step, we compare the top 20 countries by type - Movies vs TV Shows. From the dataset, we select country-specific data and compute LDA and Document Term Matrix(DTM). The following graphs illustrate the specific country data (in this case the United States).

1. Count plot for Movies and TV Shows.
2. Count plot for ratings of Movies and TV Shows.
3. Count plot for Top 10 cast.
4. Count plot for Top 10 countries.
5. Count plot for Top 10 TV Shows Seasons.
6. Displot for Duration Distribution for Netflix Movies.

E. TF-IDF vectorization

TF-IDF stands for Term Frequency Inverse Document Frequency. TF-IDF is used to fit a meaningless representation of text into a meaningful representation of numbers.

Movie plot description is transformed into TF-IDF vectorizer, then converted into an array for our model.

Cast, director, and genre are also converted into TD-IDF vector array.

F. K-Means Clustering

Using K-Means clustering, similar groups are separated and assigned a cluster number. Based on the Elbow method and the Silhouette score, 12 clusters appear to be the optimal number.

Unsupervised machine learning can be performed using K-means clustering. Based on similarities (k), it can classify unlabeled data into clusters.

G. Natural Language Processing (NLP) Model

1) Removing Punctuation: By removing punctuations from the data, we help to get rid of data that is unhelpful, or noise.

2) Removing stop-words: In any language, stop-words are common words used frequently. It is easier to get better recommendations if we remove the very commonly used words from a language.

3) Stemming: In stemming, a word is reduced to its stem or root by removing a part of it. Stemming reduces the words to their simplest form, or stem, whether or not that term is legitimate.

H. Natural Language Processing and K-Means Clustering

One of the most popular unsupervised machine learning algorithms is K-means clustering. It is typical for unsupervised algorithms to use only input vectors without labelled data outcomes when making inferences from datasets.

1) K-Means algorithm works: Using k-means++, we started the algorithm by selecting a cluster of centroids at random. This process is repeated until an optimal solution is found.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then optimizes the centroids' positions through iterative (repetitive) calculations.

IV. Silhouette Coefficient or silhouette score and Elbow Curve

An indicator of how well a clustering technique performs is done by the silhouette coefficient, or silhouette score. The value ranges from -1 to 1. 1 - There is good separation between clusters and they can be easily distinguished.

The Silhouette Coefficient s for a single sample is then given as:

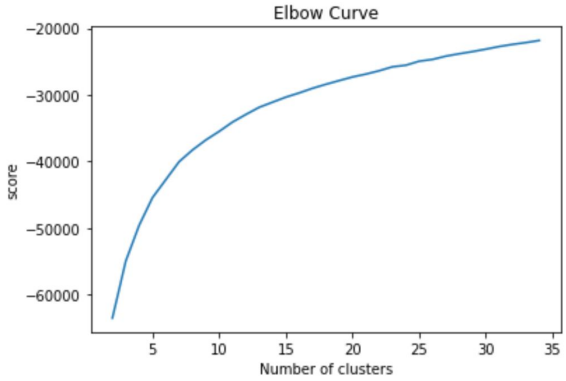
$$s = \frac{b - a}{max(a, b)}$$

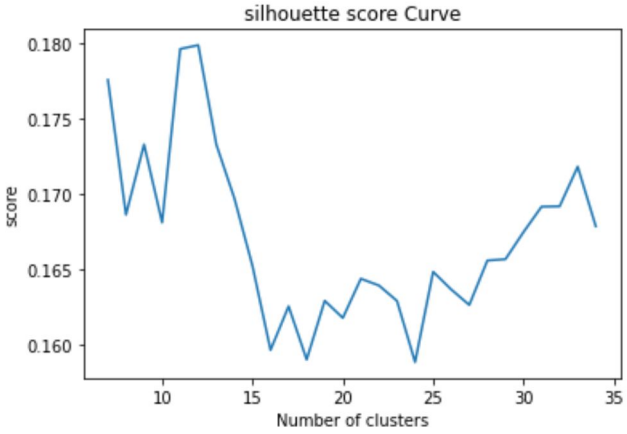
Where,

Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.

Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.

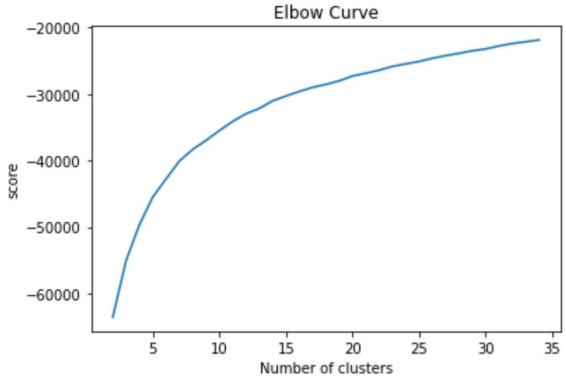
When I considered - type, director, cast, country, year_added, month_added, release_year, rating, duration, listed_in features the following are the graphs obtained.

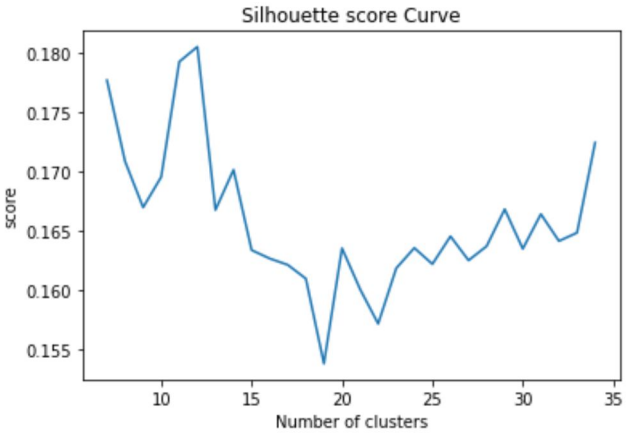




From elbow curve the optimal number of clusters seemed to be in between 10 to 15 and from silhouette score curve the optimal value is 12 clusters.

When I considered - type, director, cast, country, year_added, month_added, release_year, rating, duration, listed_in and description features the following are the graphs obtained.





From elbow curve the optimal number of clusters seemed to be in between 10 to 20 and from silhouette score curve the optimal value is 12 clusters.

V. Overview

- We started by replacing Nan values with No Director, No Cast, and Country Not Available for director, cast, and country respectively. Nan values were dropped from date_added and rating features.
- 'date_added' feature is used to obtain the 'month_added' and 'year_added' features.
- As a first step, we compare the top 20 countries by type - Movies vs TV Shows. From the dataset, we select country-specific data and compute LDA and Document Term Matrix(DTM). Several plots were taken to evaluate the available content for specific countries. There were several conclusions made regarding the United States, as an example.
- In the feature engineering - the rating values were reassigned, ordinal encoding was used on the type, the string value was dropped from the duration, etc.
- K-Means Clustering was performed on type, director, cast, country, year_added, month_added, release_year, rating, duration, listed_in features. The silhouette_score was used to calculate the number of clusters. A test on '13 Reasons Why' was then conducted to determine the recommendations.
- In NLP, we used only the description to determine recommendations.
- A K-means clustering dataset with TF-IDF vectoriser from NLP is then combined for clustering. As a result, better recommendations were made.

Conclusion

- A recommendation system with the description column works well.
- In the case of K-means, the optimal number of clusters are 12.
- When K-means is applied to the description sum column, the optimal number of clusters was also 12.
- Clustering with the description column had better recommendations than clustering without description.

- The optimal number of clusters was calculated using silhouette_score.

A. EDA Conclusion

- The most content type on Netflix is movies. The largest count of Netflix content is made with a 'TV-MA' rating.
- After 2014 the amount of content added has been increasing significantly.
- The number of movies in 2020 have reduced compared to its previous year. However the TV shows have increased.
- While most TV seasons have only 1 season, movie lengths follow a normal distribution with a mean of 100 minutes.
- According to the amount of content produced, the United States is the top country.
- International Movies are a genre mostly found on Netflix.
- In terms of titles, Jan Suter is the most popular director on Netflix.
- Anupam Kher is the most popular Netflix cast member, according to number of movies made.
- In 2018, 2019, and 2020, the majority of films were released.
- A large number of movies and TV Shows were released in October, November, December, and January.