**Netflix Movies & TV Shows Clustering**

By – Ranajay Biswas

**Data Science Trainee,**

**Almabetter, Bangalore**

**Abstract:**

Netflix is a streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more – on thousands of internet-connected devices. New TV shows and movies are added every week!

Netflix has an extensive library of feature films, documentaries, shows, award-winning Netflix originals, and tons of other contents.

We have data about Netflix contents. We perform EDA on this data and clustering algorithms. Our thorough analysis will be useful for future applications like categorizing similar movies and shows.

We created a content based filtering and recommendation system. With that we are able to make recommendations to users with the contents of their taste.

**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, it we are required to do -**

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

**Data Summary & Attributes:**

Data is in csv format. We used pandas’ read\_csv() function to import the data in our work environment.

Dataset has 7787 rows and 12 columns. And it contains object, integer and float data types.

**Attributes:**

1. Show Id : Unique ID for every Movie / TV Show
2. Type : Identifier - A Movie or TV Show
3. Title : Title of the Movie / TV Show
4. Director : Director of the Movie
5. Cast : Actors involved in the movie / show
6. Country : Country where the movie / show was produced
7. Date added : Date it was added on Netflix
8. Release year : Actual Release year of the movie / show
9. Rating : TV Rating of the movie / show
10. Duration : Total Duration - in minutes or number of seasons
11. Listed in : Genre
12. Description : The Summary description

**Approach:**

Our approach to solve this problem is going to be -

* Understanding the dataset, different rows and columns.
* During the EDA, we tried to find popular actors and directors, what are the top content producing countries. We will get an idea about different genres and runtimes for different contents. Statistical methods and Visualizations are going are very important in this EDA process.
* In the pre-processing step, we shall filtered the most important features, make necessary transformations, create or omit features as needed. Depending on the features we choose, we will find best approaches for text pre-processing. Then we have the data, ready for implementing clustering algorithms or building a recommender system with this data.
* Using Topic Modelling, we tried to find popular genres. We tried different clustering algorithms to see which fits the data best. The best performing algorithm will be selected and then we analyzed different clusters made by the algorithm to find patterns inside those clusters.

**Steps Involved:**

* **Loading & Understanding the Data**

The csv file containing the Email Data was loaded in our work environment as a data frame using pandas. We checked top 5 rows and bottom 5 rows to get an initial idea about the data.

* **Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

* There were many null values in the data. **Director** column has roughly 31%, **Cast** column has 9%, **Country** column has 6.51%, **Date** **Added** column has 0.13% and **Rating** column has 0.09% null values.
* Checking for duplicates, we found no duplicates in the data.
* During the EDA, we used different plots and charts to visualize the patterns and trends that are present in the data.
* Using pieplot, we found that there is roughly 69% movies and 31% TV shows present in the data. We found that 36 % of total content are ‘Netflix Originals’.
* We used histogram and distribution plot to see how the numbers of releases of movies and shows have changed throughout the years. We found the years with most releases and also noticed the numbers going down after 2019 because of Covid-19 pandemic.
* Using countplots, we found the top content providing countries or Netflix. We visualized how different countries prefer different genres and contents of different ratings.
* Using histogram we found most movies in the data have 90-120 minutes runtime and most shows are 1-2 seasons long.
* We found actors and directors with the most credits on Netflix.
* With the help of barplots, we visualized the trend of how movies and shows are added on Netflix throughout different months.
* We discovered which are the popular ratings and genres among the audiences.
* **Pre-Processing & Feature Engineering:**

**Feature Selection:**

Our first objective is to create a content based recommender system which will need information about movie description, genre, rating, cast and directors.

During topic modelling, we utilize the movie descriptions and genres. And in the clustering stage, in order to cluster similar contents, we need description, genres, and ratings columns. So, these are the columns that we kept as the important columns.

**Null value Treatment:**

For Director, Cast and Content Rating columns, we replaced the null values with ‘Unknown’.

**Text Pre-processing:** We used punctuation removal, lower casing, stop words removal and lemmatization to acquire clean texts. Then we passed that clean text through the **TFIDF** vectorizer to get the words’ vector forms.

TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus (data-set).

* **Modelling:**
* Recommender System: To create a content based recommender system, we found the **Cosine similarities** between the data points. Based on the movie or show that user gives as input, our system calculates the similarities and recommends top 5 most similar content available on Netflix.
* Topic Modelling:Using LSA, we found the most important topics that are present in the data. We visualized it using Word Cloud.

Latent semantic analysis (LSA) is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text.

* Clustering**:** Clustering is an unsupervised technique of finding the most similar contents in the data. The main concept is that similar datapoints will have more similarity, so the distance between them will be less compared to non-similar data points.

We used **K-means** and **Agglomerative Hierarchical Clustering** methods to cluster similar contents.

* **Evaluation:**

For evaluation of the clusters, we used Silhouette Score and Davis Bouldin Score. We found the optimal number of clusters using the Elbow method and Dendogram.

**Silhouette score** is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. Positive value Means clusters are well apart from each other and clearly distinguished. With the increasing value of the score, we can be confident to have better clusters.

* **Model Selection:**

**K-Means** clustering gave better results in terms of silhouette score and Davis Bouldin score. So, that is going to be the best choice of model for this data.

* **Clustering analysis:**

The clusters that were formed by K-Means algorithm, we performed EDA on those clusters, and found why they are similar. We discovered the genres, content ratings difference among the clusters’ contents. We made discoveries about how each cluster had similar movies and shows and how they are different from other clusters.

* **Conclusion:** In the conclusion, we can say that with our thorough analysis of the data, we found many patterns and insights that can be used for business advantage for targeting audiences.

Recommendation system will prove to be useful for making recommendations similar to user’s taste.

With the help of topic modelling, we can find the most popular contents’ genres. That can help us understand what type of content is most popular.

Clustering techniques can be used in future when we are trying to categorize movies or shows. It can be helpful to decide what kind of content should be added on the platform based on what type audience we are dealing with.

**References-**

1. AlmaBetter
2. Kaggle
3. MachineLearningMastery
4. GeeksforGeeks
5. Analytics Vidhya