**NETFLIX MOVIES AND TV SHOWS CLUSTERING**

**Name: Mohd Aun Farooqui**

**INDEX**

**SL.NO Topics**

1. **Abstract**
2. **Introduction**
3. **Project Goal**
4. **Data Description**
5. **Exploratory Data Analysis**
6. **Data Visualization**
7. **Algorithms**
8. **Conclusion**
9. **Challenges**
10. **Abstract**

This project aims to develop a recommendation system for Netflix using content-based filtering and clustering techniques. The dataset used in the project contains information about movies and TV shows available on Netflix, including genres, ratings, and descriptions.

The project starts with exploratory data analysis to gain insights into the data and identify patterns and trends. The analysis reveals the most popular genres and the countries that produce the most content on Netflix.

Next, the project applies content-based filtering to develop a recommendation system based on the similarity between movies and TV shows. The system recommends ten similar titles to a user based on the description of a movie or TV show they like.

In addition to content-based filtering, the project applies clustering techniques to group movies and TV shows based on their attributes, such as ratings and genres. The clustering model identifies twelve distinct groups of movies and TV shows, each with its own unique set of attributes.

The project concludes with a discussion of the strengths and limitations of the recommendation system and suggestions for future improvements. Overall, the recommendation system provides users with personalized recommendations based on their interests and preferences, improving their overall viewing experience on Netflix.

1. **Introduction**

Netflix is one of the largest streaming platforms in the world with over 230 million subscribers as of 2023 in more than 190 countries. It offers a wide range of movies, TV shows, and documentaries to its subscribers. Due to this extensive range of content, it becomes difficult for a user to choose what they want to watch. This problem is known as the "paradox of choice." In this project, we have developed a content-based recommendation system that suggests similar movies and TV shows based on the user's selection.

The recommendation system is built using the Netflix dataset, which consists of TV shows and movies available on Netflix as of 2021. The dataset contains information such as title, director, cast, description, rating, etc. The recommendation system uses natural language processing techniques to extract relevant features from the movie/TV show descriptions and then uses machine learning algorithms to recommend similar movies/TV shows.

1. **Project Goal**

The goal of this project is to build a recommendation system for Netflix movies and TV shows using content-based recommended system. The recommendation system will take in a movie or TV show title as input and output 10 similar titles. We will also perform exploratory data analysis to gain insights into the data and use K-Means and Hierarchical clustering algorithms to cluster similar titles based on their attributes. Finally, we will evaluate the performance of the recommendation system using various metrics. The project aims to help Netflix users discover new titles that match their interests and preferences.

1. **Dataset Description**

The dataset used for this project is the Netflix Movies and TV Shows dataset, obtained from AlmaBetter. The dataset contains information of the TV shows and movies available on Netflix as of 2021. The dataset includes information such as show title, director, cast, country of origin, release year, rating, duration, type (TV show or movie), and a description of the show or movie. The dataset contains a total of 7787 rows and 12 columns. This dataset is used to develop a content-based recommendation system for Netflix shows and movies.

**Variable Description:**

show\_id: Unique id for every movies/Tv shows

type: Identifier - A movie or Tv show

title: Title of the movie/show

director: Director of the show

cast: Actors involved in the show

Country: Country of production

date\_added: Date is what added on Netflix

release\_year: Actual release year of the show

rating: TV rating of the show

duration: Total duration in minutes or number of seasons.

listed\_in: Genre

Description: The summary description

1. **Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a critical first step in any data analysis project. It helps to understand the data and identify patterns, relationships, and outliers in the dataset. For the NETFLIX MOVIES AND TV SHOWS CLUSTERING, EDA was performed on the dataset to gain insights into the data and understand its structure.

First, we explored the distribution of movies and TV shows on Netflix, and found that the majority of content on Netflix is movies. We also examined the distribution of content by country and discovered that the United States produces the most content, followed by India and the United Kingdom.

Next, we looked at the rating system on Netflix and found that the majority of content is rated TV-MA, followed by TV-14 and R. We also analyzed the distribution of content by release year and discovered that the number of movies and TV shows on Netflix has been increasing rapidly since 2008.

Finally, we analyzed the relationships between different attributes of the dataset. We found that there is a positive correlation between the year added and the release year of movies and TV shows, and a negative correlation between the release year and the duration.

Overall, EDA helped us to gain insights into the structure of the Netflix Movies and TV Shows dataset, and provided us with the information we needed to develop a content-based recommendation system.

1. **Data Visualization**

Data visualization is the graphical representation of data and information using visual elements like charts, graphs, and maps. It helps to communicate complex data and information in an easily understandable and accessible format.

* Pie chart
* Bar Plot
* Box plot
* Heatmap
* Pair plot
* Count plot
* Word Cloud
* Stacked Bar

1. **Algorithm**

The main algorithm used in this project are as follows-

1. **Hierarchical Clustering:** A clustering algorithm used to group similar movies and TV shows together based on their descriptions. The algorithm creates a hierarchy of clusters by successively merging or splitting them until a stopping criterion is met. In this project, the stopping criterion was the number of clusters.
2. **K-Means Clustering:** K-means clustering is a type of unsupervised learning algorithm that is used to group similar data points into clusters. In this project, we used K-means clustering to group the movies and TV shows based on their descriptions and create clusters of similar items.
3. **Content-Based Recommendation System:** A content-based recommendation system is a type of recommendation system that suggests items to users based on their preferences and interests. In this project, we used a content-based recommendation system to recommend movies and TV shows based on their descriptions.
4. **Conclusion**

* After performing exploratory data analysis on the Netflix dataset, we can conclude that:
* Netflix's content has increased over the years, with a significant increase in the number of TV shows since 2008.
* The majority of the content on Netflix is movies, followed by TV shows and documentaries.
* The United States is the country with the most content on Netflix, followed by India and the United Kingdom.
* The most common genre on Netflix is drama, followed by comedy, Documentaries and Action & Adventure.
* Netflix has a diverse range of content from different countries, with a significant portion of content being produced in the United States.
* Most TV shows on Netflix have one season, with a few having two, three or more seasons.
* FROM MODEL IMPLEMENTATION:
* The data was clustered based on the attributes: director, cast, country, genre, rating, and description.
* TFIDF vectorizer was used to tokenize, pre-process, and vectorize the values in these attributes, creating a total of 20000 attributes.
* The k-means clustering algorithm was implemented to group similar movies and TV shows together based on the extracted features. It was found that a value of 6 was optimal for the number of clusters.
* The content-based recommendation system was built using cosine similarity on the Netflix dataset. The dataset was pre-processed, cleaned and transformed into a structured format. A bag-of-words approach was used to create a term frequency-inverse document frequency (TF-IDF) matrix, which was then used to compute cosine similarities between pairs of movies and TV shows.
* Finally, the content-based recommendation system based on cosine similarity was able to recommend 10 similar movies/TV shows based on a selected title. This system could be used to suggest movies or TV shows to Netflix users based on their preferences.
* I hope that these recommendation systems can be used to help users find TV shows and movies that match their interests. By providing recommendations that are tailored to the user's preferences, we can enhance their overall experience of using Netflix.

1. **Challenges**

There were several challenges encountered during this project:

**1. Data Cleaning:** One of the biggest challenges in this project was cleaning and pre-processing the data. The data contained missing values, duplicates, and inconsistent formats, which had to be addressed before it could be used in the analysis.

**2. Feature Extraction:** Extracting meaningful features from the text data was another challenge. The process of converting the textual data into numerical vectors using techniques like Count Vectorization and TF-IDF requires careful consideration of parameters such as stop words and tokenization.

**3. Choosing the Right Algorithm:** Choosing the right algorithm to build the recommendation system was a challenge. There are several algorithms available, and each has its own strengths and weaknesses. Ultimately, the decision to use a Content-Based approach was based on the nature of the data and the project goals.

**4. Overcoming Dimensionality:** With the large number of features in the dataset, the issue of dimensionality had to be addressed. The high dimensionality of the data makes it computationally expensive to perform calculations on the entire dataset. To overcome this, techniques such as Principal Component Analysis (PCA) were used to reduce the dimensionality of the data.

**5. Evaluating the Model:** Evaluating the performance of the recommendation system is difficult as there are no clear metrics to measure its effectiveness. In addition, there is no single "correct" recommendation for a user, making it challenging to evaluate the accuracy of the recommendations. Therefore, the evaluation of the model can be done based on the user's feedback and the relevance of the recommendations.

**10. References**

1. <https://www.almabetter.com/notes>
2. <https://www..python.org.com>
3. <https://www.stackoverflow.com>