Netflix Movies and TV Shows Clustering

by
Shaik Ahmad Basha
Data science trainee,
AlmaBetter, Bangalore.

Abstract:

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.
.

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.

The main objective behind the project is to explore the dataset to find some key understandings and to cluster similar content by using machine learning clustering algorithms

Data Description:

The dataset has 7787 rows and 12 columns. Those Features are

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 Releaseyear 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 : Genere
12. description: The Summary description

Introduction:

Netflix began experimenting with data in 2006 when they held a competition to create an algorithm to accurately predict how much a viewer would like a movie based on their preferences. Netflix has expanded the use of data beyond rating forecasting and into a variety of areas, including personalized ranking, page generation, search, picture selection, messaging, and marketing. Netflix Recommendation Engine (NRE), is composed of algorithms that select content based on each user's unique profile. The engine filters over 3,000 titles at once utilizing 1,300 recommendation clusters based on user choices. The engine's precise recommendations account for 80% of the Netflix viewer activity. According to estimates, Netflix saves over $1 billion annually thanks to the NRE.

In 2018, Flixable, a third-party Netflix search engine, released a report showing the number of TV shows on Netflix tripling since 2010. Apart from clustering the video content, a comprehensive exploratory data analysis will be performed during this project to understand the trends in the diverse video content catalogue and to verify the findings of Flexible, that Netflix has been focusing on producing TV shows in recent years.

Approach

The first step imported all the necessary libraries like NumPy, Pandas etc. and then collected the data. I started with understanding the data like what are the columns and their meanings and data types. After that, the second step is data preprocessing. Data preprocessing is a process where raw data is converted into clean data. The dataset has null values. After handling the null values, the next step is to analyze the data using Exploratory data analysis techniques. And then I moved on text preprocessing where I have done tokenization, lowercasing, removed stop words and punctuations and stemming. And then I converted words to vectors by using TF-IDF Vectorizer. After converting, the data has so many columns, so in order to reduce dimensionality, I applied Principal component analysis. And the last step is to cluster the data using clustering algorithms. And find the similar movies / TV shows using cosine similarity

Steps Involved

Data Understanding

After the loading and collecting the dataset, understanding the data is very important. I Understand the various features of the dataset.

Data Cleaning and Manipulation:

Cleaning and manipulating the data is very important. Only the cleaned data can be fitted into any machine learning models.

The data has no duplicated values. But has 5 features with null values. They are director, cast, country, rating, date_added.

For director and cast columns, I just replaced null values with unknown

For Country column, I replaced null values with most frequent value.

For rating, date_added columns, I simply dropped the rows with null values.

Created a new column 'month_added' from 'data_added' which contains month

number in which the movie/TV Show released

Created a new column ‘movie_duration’ from ‘duration’ column which contains the duration of movie in minutes.

Created a new column ‘num_of_seasons’ from ‘duration’ column which contains the number of seasons of TV Show.

Created a new column ‘target_ages’ which contains age groups.

Exploratory Data Analysis

After handling with null values, duplicated values, and manipulating the data, we can deep dive into Exploratory Data Analysis. From this analysis, we can draw various insights. From Univariate analysis, the conclusions are:

- There are more movies that tv shows on Netflix
- Jan Suter and Raul Campos directed most number of movies.
- Alastair Fothergill and Ken Burns directed most number of TV Shows.
- Anupam Kher acted in most number of movies followed by Shah Rukh Khan.
- Takahiro Sakurai acted in most number of TV Shows followed by Junichi Suwabe
- United Stated produced most number of movies/Tv Shows followed by India.
- Most number of movies are released in the month of december and january.
- In the year 2018 most number of movies and TV Shows are released.
- In the recent years, netflix has started to increase the TV Shows content as you can see the graph is increasing gradually.
- Most number of movies and TV SHows belongs the rating category of 'TV-MA'
- Most number of movies / TV Shows are from genre of 'International Movies' followed by dramas.
- Most number of movie duration is in the range of 80 to 125 min.
- Most number of TV Shows have only One season.
- Most number of movies and TV Shows are made for adults only.

From Bivariate Analysis, the conclusions are:

- The content for Adults is mostly made from Spain and France countries.
- The content for Teens is mostly made from Egypt and India countries.
- The content for Older Kids is mostly made from Japan country.
- The content for Kids is mostly made from Canada country.
- The content for Adults is mostly released in recent years(2020, 2019, 2018, 2017).
- The content for Teens is mostly released in the year 2010 and 2012.
- The content for Older is Kids released in year of 2014.
- The content for Kids is mostly released in the year of 2020.
- The content for Adults is mostly from the genre of 'stand up comedy' followed by Dramas.

- The content for Teens is mostly from the genre of 'Comedies, Dramas, international movies'
- The content for Older Kids is mostly from the genre of 'Children, Family movies and Comedies'
- The content for Kids is mostly from the genre of Kids TV and Children and Family Movies
- 94% content from Raul Campos and Jan Suter is for adults.
- Steven Spielberg , David Dhawan and Cathy Garcia-Molina are the directors who makes content mostly for Teens and Older Kids.

Text Preprocessing

Text preprocessing involves transforming text into a clean and consistent format that can then be fed into a model for further analysis and learning. Text preprocessing techniques may be general so that they are applicable to many types of applications, or they can be specialized for a specific task.

Tokenization

The tokenization stage involves converting a sentence into a stream of words, also called “tokens.” Tokens are the basic building blocks upon which analysis and other methods are built.
Many NLP toolkits allow users to input multiple criteria based on which word boundaries are determined. For example, you can use a whitespace or punctuation to determine if one word has ended and the next one has started. Again, in some instances, these rules might fail. For example, *don’t*, *it’s*, etc. are words themselves that contain punctuation marks and have to be dealt with separately.

Lower Casing

Lowercasing ALL your text data, although commonly overlooked, is one of the simplest and most effective form of text preprocessing. It is applicable to most text mining and NLP problems and can help in cases where your dataset is not very large and significantly helps with consistency of expected output.

Removing Stop Words

"Stop words" are frequently occurring words used to construct sentences. In the English language, stop words include is, the, are, of, in, and and. For some NLP applications, such as document categorization, sentiment analysis, and spam filtering, these words are redundant, and so are removed at the preprocessing stage.

Removing Punctuations

The next most common text processing technique is removing punctuations from the textual data. The punctuation removal process will help to treat each text equally. For example, the word data and data! are treated equally after the process of removal of punctuations.

We need to take care of the text while removing the punctuation because the contraction words will not have any meaning after the punctuation removal process. Such as ‘don’t’ will convert to ‘dont’ or ‘don t’ depending upon what you set in the parameter.

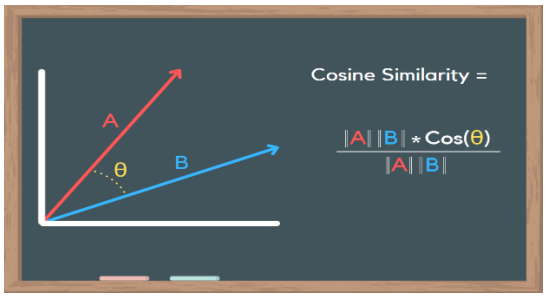
We also need to be extra careful while choosing the list of punctuations that we want to exclude from the data depending upon the use cases. As string.punctuation in python contains these symbols !"#$%&\'()*+,-./:;?@[\\]^_{|}~`

Stemming

The term word stem is borrowed from linguistics and used to refer to the base or root form of a word. For example, learn is a base word for its variants such as learn, learns, learning, and learned. Stemming is the process of converting all words to their base form, or stem. Normally, a lookup table is used to find the word and its corresponding stem. Many search engines apply stemming for retrieving documents that match user queries. Stemming is also used at the preprocessing stage for applications such as emotion identification and text classification.

Cosine Similarity

Cosine Similarity is a measurement that quantifies the similarity between two or more vectors**.** The cosine similarity is the cosine of the angle between vectors. The vectors are typically non-zero and are within an inner product space.



The cosine similarity is described mathematically as the division between the dot product of vectors and the product of the euclidean norms or magnitude of each vector.

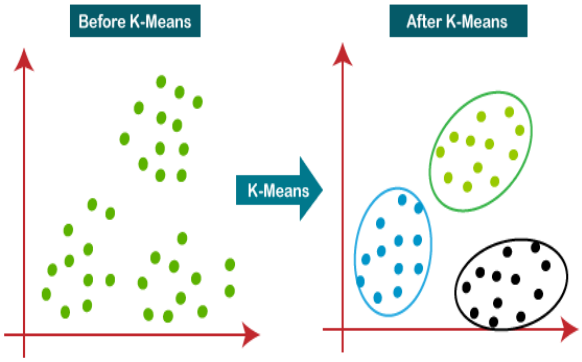
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

By using cosine similarity, we can find similar movies / TV Shows.

K Means Clustering

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.



The working of the K-Means algorithm is explained in the below steps:

Step 1 : Select the number K to decide the number of clusters.

Step 2 : Select random K points or centroids. (It can be other from the input dataset).

Step 3 : Assign each data point to their closest centroid, which will form the predefined K clusters.

Step 4 : Calculate the variance and place a new centroid of each cluster.

Step 5 : Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step 6 : If any reassignment occurs, then go to step-4 else go to FINISH.

Step 7 : The model is ready.

Choosing K Value

Elbow Method

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

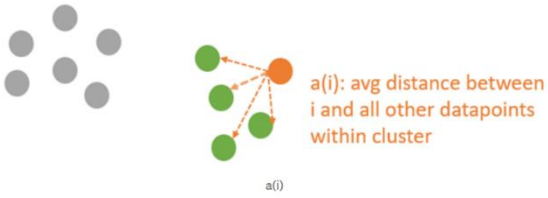
- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

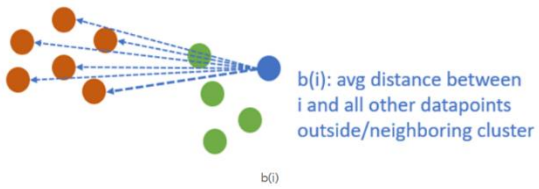
Silhouette Analysis

The silhouette coefficient or silhouette score kmeans is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation).

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

- S(i) is the silhouette coefficient of the data point i.
- a(i) is the average distance between i and all the other data points in the cluster to which i belongs.
- b(i) is the average distance from i to all clusters to which i does not belong.





We will then calculate the average_silhouette for every k.

$$\text{AverageSilhouette} = \text{mean}\{S(i)\}$$

Then plot the graph between average_silhouette and K.

The value of the silhouette coefficient is between [-1, 1].

A score of 1 denotes the best meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.

The worst value is -1. Values near 0 denote overlapping clusters.

Hierarchical Clustering

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as **hierarchical cluster analysis** or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram**.

Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

The hierarchical clustering technique has two approaches:

1. Agglomerative: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

Agglomerative Hierarchical Clustering

The agglomerative hierarchical clustering algorithm is a popular example of HCA. To group the datasets into clusters, it follows the bottom-up approach**.** It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.

This hierarchy of clusters is represented in the form of the dendrogram.

Steps involved in Agglomerative Clustering

Step 1: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.

Step 2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

Step 3: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.

Step 4: Repeat Step 3 until only one cluster left. So, we will get the following clusters.

Step 5: Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

Conclusion:

- There are more movies that tv shows on netflix
- Jan Suter and Raul Campos directed most number of movies.
- Alastair Fothergill and Ken Burns directed most number of TV Shows.
- Anupam Kher acted in most number of movies followed by Shah Rukh Khan.
- Takahiro Sakurai acted in most number of TV Shows followed by Junichi Suwabe
- United Stated produced most number of movies/Tv Shows followed by India.
- Most number of movies are released in the month of december and january.
- In the year 2018 most number of movies and TV Shows are released.
- In the recent years, netflix has started to increase the TV Shows.
- Most number of movies and TV SHows belongs the rating category of 'TV-MA'
- Most number of movies / TV Shows are from genre of 'International Movies' followed by dramas.
- Most number of movie duration is in the range of 80 to 125 min.
- Most number of TV Shows have only One season.
- Most number of movies and TV Shows are made for adults only.
- The content for Adults is mostly made from Spain and France countries.
- The content for Teens is mostly made from Egypt and India countries.
- The content for Older Kids is mostly made from Japan country.
- The content for Kids is mostly made from Canada country.
- The content for Adults is mostly released in recent years(2020, 2019, 2018, 2017).
- The content for Teens is mostly released in the year 2010 and 2012.
- The content for Older is Kids released in year of 2014.
- The content for Kids is mostly released in the year of 2020.
- The content for Adults is mostly from the genre of 'stand up comedy' followed by Dramas.
- The content for Teens is mostly from the genre of 'Comedies, Dramas, international movies'
- The content for Older Kids is mostly from the genre of 'Children, Family movies and Comedies'
- The content for Kids is mostly from the genre of Kids TV and Children and Family Movies
- 94% content from Raul Campos and Jan Suter is for adults.
- Steven Spielberg , David Dhawan and Cathy Garcia-Molina are the directors who makes content mostly for Teens and Older Kids.
- By using the data, a recommender system was created with cosine similarity.

- By using the data, created a clusters using k means and hierarchical clustering.
- By applying silhouette analysis, for k means, the optimal k value is 6.
- And for k = 6, silhouette score is 0.5001759759427156
- For hierarchical clustering, from dendrogram we can choose optimal k value as 3. And silhouette score is 0.5049348034764315