Netflix Movies And TV Shows Clustering

BAISHNAVEE MAHATO

Data science trainee,
AlmaBetter, Bangalore

1. INTRODUCTION:

Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies on an internet-connected device. Depending on your plan, you can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection.
Netflix content varies by region and may change over time. We can watch a variety of award-winning Netflix originals, TV shows, movies, documentaries, and more.

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled
dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of
similar data points. The objects with the possible similarities remain in a group that has less or no
similarities with another group."
It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color,
behavior, etc., and divides them as per the presence and absence of those similar patterns.

2. Problem Statement:

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.

The given challenge is to build a classification model to predict the sentiment of Covid-19 tweets. The tweets have been pulled from Twitter and manual tagging has been done. I have given information like Location, Tweet At, Original Tweet, and Sentiment.

In this project, we are required to do
1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix has increasingly focusing on TV rather than movies in recent years.

4. Clustering similar content by matching text-based features

3. Data Description:

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
description: The Summary description

5. Data Summary:

The original dataset has 12 columns and 7787 rows.

In this project, we are required to do
1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix has increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features

5. Basic Exploratory Data Analysis:

- The number of releases have There are more number movies on Netflix than TV shows. It is evident that there are more movies on Netflix than TV shows.
- Netflix has 4673 movies, which is more than double the quantity of TV shows significantly increased after 2015 and have dropped in 2021 due to Covid 19.
- TV-MA ratings are the highest in movies and TV shows, because TV-MA standing for Mature Audience Only. Because this programme is intended for adults, it may not be appropriate for children under the age of 17.
- As stated previously regarding the top genres, it's no surprise that the most popular directors on Netflix with the most titles are mainly international as well.

6. Data Preprocessing:

Modeling Approach:

1.Select the attributes based on which you want to cluster the shows

2.Text preprocessing: Remove all stopwords and punctuation marks, convert all textual data to lowercase.

3.Lemmatization to generate a meaningful word out of corpus of words

4.Tokenization of corpus

5.Word vectorization

6.Dimensionality reduction

7.Used different algorithms to cluster the movies, obtaining the optimal number of clusters using different techniques.

We will cluster the shows on Netflix based on the following attributes:

Director

Cast

Country

Listed in (genres)

Description

7. Dimensionality reduction using PCA:

1.I found that 100% of the variance is explained by about ~4500 components.

2.Also, more than 80% of the variance is explained just by 3000 components.

3.Hence to simplify the model, and reduce dimensionality, I can take the top 3000 components, which will still be able to capture more than 80% of variance.

8. Clusters Implementation:

K-Means Clustering:

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

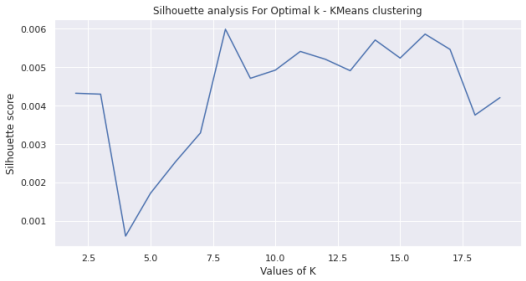
It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum.

The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

Building clusters using the K-means clustering algorithm.

Visualizing the elbow curve and Silhouette score to decide on the optimal number of clusters for

K-means clustering algorithm.



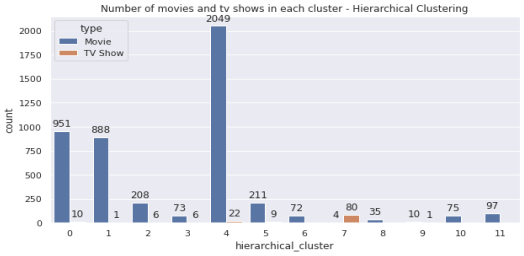
The highest Silhouette score is obtained for 8 clusters.

Hierarchical clustering:

Hierarchical cluster algorithm is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom.

The algorithm groups similar objects into groups called clusters.

Hierarchical clustering takes away the problem of having to pre-define the number of clusters.



12 clusters using the Agglomerative clustering algorithm:

9. Conclusions:

1.The dataset contained about 7787 rows, and 12 columns.

2.I began by dealing with the datasets missing values and doing exploratory data analysis (EDA).

3.There are more number movies on Netflix than TV shows. It is evident that there are more movies on Netflix

than TV shows. Netflix has 4673 movies, which is more than double the quantity of TV shows.

4.The number of releases have significantly increased after 2015 and have dropped in 2021 due to Covid 19.

5.TV-MA ratings are the highest in movies and TV shows, because TV-MA standing for Mature Audience

Only. Because this programme is intended for adults, it may not be appropriate for children under the age of

17.

6.I used Principal Component Analysis (PCA) to handle the curse of dimensionality. 3000 components were

able to capture more than 80% of variance, and hence, the number of components were restricted to 3000.

7.I first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to

be 8. This was obtained through the elbow method and Silhouette score analysis.

8.Then clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters

came out to be 12. This was obtained after visualizing the dendrogram.

10. References:

https://en.wikipedia.org/wiki/Netflix

https://www.javatpoint.com/clustering-in-machine-learning

https://www.kaggle.com/code/onyonixch/netflix-movies-tv-shows-eda-and-clustering