Netflix Movies and TV Shows clustering

Sk Samim Ali,
Sarath Haridas,
Data science trainees,
AlmaBetter, Bangalore

ABSTRACT:

The American production firm and technology and media services provider Netflix. This project uses unsupervised machine learning methods to suggest Netflix movies and TV shows. Clustering takes priority. It is an unsupervised learning task used for exploratory data analysis in order to discover certain hidden patterns that exist in the data but cannot be properly classified. Cluster analysis is primarily dependent on the primary aim of keeping objects inside a cluster closer together than objects belonging to other groups or clusters. Sets of data can be classified or grouped together based on some similar features and are referred to as clusters. There are various sorts of clustering paradigms, depending on the data and anticipated cluster features. In the very recent times, many new algorithms have emerged which aim towards bridging the different approaches towards clustering and merging different clustering algorithms given the requirement of handling sequential, extensive data with multiple relationships in many applications across a broad spectrum. In this project, we used different clustering methods- Silhouette Clustering Method, K-Mean Clustering, Elbow Method, Dendrogram, Agglomerative Clustering. Keywords: Netflix, unsupervised, clustering, groups, algorithms, Silhouette, K-Mean, Elbow method, Agglomerative, Dendrogram.

INTRODUCTION:

We are all familiar with Netflix, the biggest provider of on-demand internet streaming media and online DVD movie rentals. Marc and Reed started it on August 29, 1997, in Los Gatos, California. More than 100 million hours of TV series and movies are watched each day by its 69 million members across 60 countries. People all across the world utilise it as a well-liked entertainment resource. The goal of this project is to suggest Netflix-style material. This project uses unsupervised machine learning. We shall present the clustering process through this project. Netflix might use this technology and the algorithms to help with user recommendations.

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.

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

Data Pre-processing :

- Working on the text-based features (description, listed_in).
- Removing punctuations and stop words from text features.
- Stemming process applied for those text features.

- Applying the count vectorizer on those updated text.

EXPLORATORY DATA ANALYSIS:

Exploratory data analysis was then performed on the clean data set to obtain certain observations like :

- **Percentage distribution of content among all the countries**
- **Value count of TV and movie shows in the Dataset**
- **Total count of type content with respect to unique age rating values**
- **Rating distribution of movies and TV shows**
- **Top Genres and Actors on Netflix**
- **Top releases for last 10 years**

Training Process :

- Silhouette analysis on k-means clustering
- Elbow Method
- Dendrogram
- Agglomerative Clustering

Clustering Methods:

1. SILHOUETTE ANALYSIS ON K MEANS CLUSTERING

The silhouette score measures how effectively samples are clustered with other samples that are similar to them in order to assess the quality of clusters produced by clustering algorithms like K-Means. Each sample of various clusters receives a Silhouette score.
The silhouette coefficient gauges a data point's cohesiveness (similarity to other clusters) with respect to other data points within the same cluster (separation).

- Select a range of values of k (say 1 to 10).
- Plot Silhouette coefficient for each value of K.

The equation for calculating the silhouette coefficient for a particular data point:

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

- S(i) is the silhouette coefficient of the data point i.

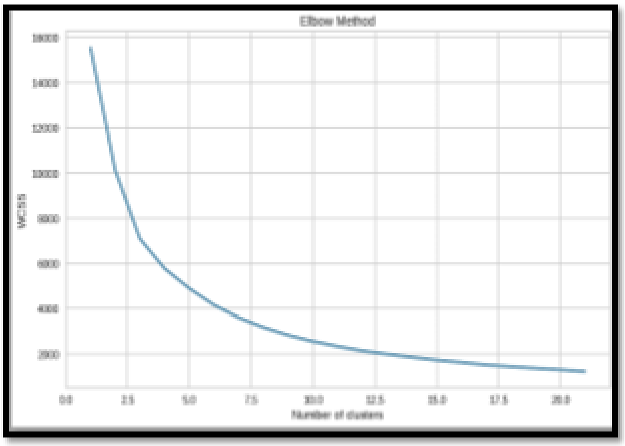
- a(i) is the average distance between i and all the other data points in the cluster to which i belongs.
- b(i) is the average distance from i to all clusters to which i does not belong.
- We will then calculate the average_silhouette for every k.

Average Silhouette = mean{S(i)}

2. ELBOW METHOD:

A heuristic for counting the number of clusters in a data collection is the elbow approach. The process is graphing the explained variance as a function of the number of clusters, then choosing the number of clusters to employ at the elbow of the curve. The number of parameters in other data-driven models, such as the number of primary components to describe a data collection, can be chosen using the same methodology.
The dataset is subjected to k-means clustering using the elbow approach for a range of k values, such as 1 to 10.

- Perform K-means clustering with all these different values of K. For each of the K
- values, we calculate average distances to the centroid across all data points.
- Plot these points and find the point where the average distance from the centroid falls
- suddenly (“Elbow”).



3. DENDOGRAM:

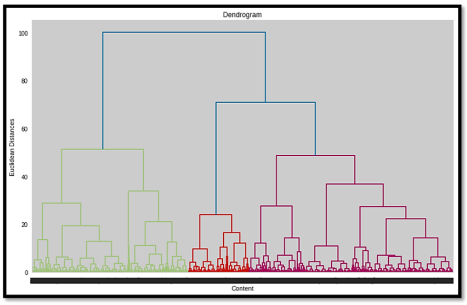
An example of a dendrogram is a tree diagram that demonstrates hierarchical clustering, or the connections between related groups of data. They can depict any kind of clustered data, but they are widely used in biology to illustrate clustering between genes or samples. A dendrogram can be a row graph or a column graph, as shown in the image below. Although some dendrograms are spherical or fluid in shape, software will often generate a row or column graph.

No matter what the shape, the basic graph comprises of the same parts:

The clade is the branch. Usually labelled with Greek letters from left to right (e.g., αβ, δ...).

Each clade has one or more leaves. The leaves in the above image are:

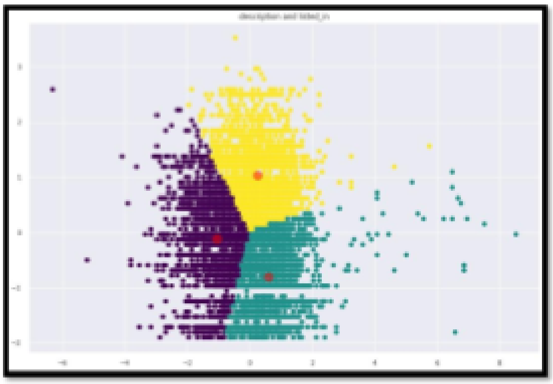
- Single (simplicifolius): F
- Double (bifolius): D E
- Triple (trifolious): A B C



CLUSTERING MODELS:

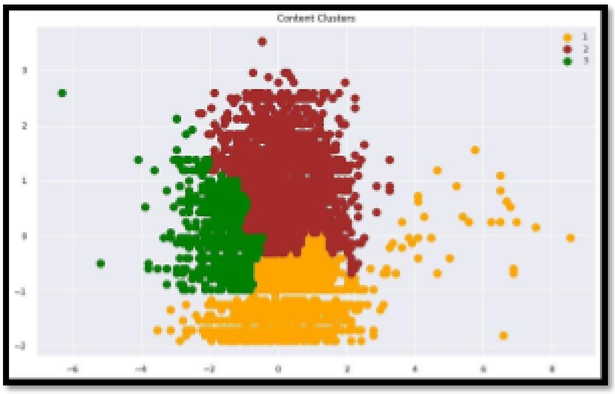
1. K-MEANS CLUSTERING:

The unsupervised learning algorithm K-Means Clustering is used to solve the clustering problems in machine learning or data science. In this case, K indicates the minimum number of pre-defined clusters that must be generated as part of the process; for instance, if K=2, there will be two clusters, if K=3, there will be three clusters, and so on. The unlabeled dataset is divided into k different clusters using an iterative process. Each cluster comprises just one dataset and has a unique set of properties. It provides a straightforward method for categorizing the groups in the unlabeled dataset on our own, without the requirement for any training. It also enables us to cluster the data into several groups. Each cluster has a centroid assigned to it because the algorithm is centroid-based. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.



2. AGGLOMERATIVE HIERARCHICAL CLUSTERING:

The iterative classification method known as Agglomerative Hierarchical Clustering (AHC) operates on a straightforward principal. a structure that provides more useful information than the flat clustering method's unstructured set of clusters. The number of clusters does not need to be predetermined for this clustering process. Bottom-up algorithms start off by treating each piece of data as a singleton cluster, then they gradually combine pairs of clusters until there is only one cluster left that holds all the data. The process of agglomerative clustering is "bottom-up." In other words, each item is originally thought of as a cluster with just one piece (leaf). The two clusters that are the most comparable are joined into a new, larger cluster at each stage of the process (nodes).



CONCLUSION:

1. The data set consists of 7787 rows and 12 columns, and the director and cast features have a significant number of missing values. As a result, we chose to remove the director attribute as well as the duration and show id attributes because they are not necessary for our model.

2. We have two types of content TV shows and Movies (30.95% contains TV shows and 69.05% contains Movies)

3. By a wide amount, Netflix has the most material from the United States, followed by India.

4. Anupam Kher has appeared in the most Netflix movies of any actor. The most popular genre is documentaries, followed by stand-up comedy.

5. Most films were released in the years 2018, 2019, and 2020.

6. Because of Covid 19, the number of releases dramatically increased after 2015 and decreased in 2021.

7. By analysing the content added over years we get to know that in recent years Netflix is focusing movies than TV shows.

8. In order to prepare a dataframe for the clustering methods, we performed feature engineering as the second action. This involved deleting specific variables.

9. By using the silhouette score approach for n range clusters on the dataset, we obtained the best score of 0.348 for clusters = 3, which indicates that the content was adequately explained on its own clusters.

10. For the clustering algorithm, we utilised "description" and "listed_in" attributes

11. Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangements

12. Speaking about other different cluster methods, K mean, hierarchical, agglomerative clustering on data, we got the best cluster arrangements**.**

Optimal number of cluster = 3

REFERENCES:

- Geeksforgeeks
- Towardsdatascience
- analyticsvidhya