Netflix Movies and TV Shows clustering

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

ABSTRACT:

Netflix is an American technology and media services provider and production company. This project is to recommend the Netflix Movies and Shows by using unsupervised machine learning algorithms. The foremost task is clustering. It is an unsupervised learning task used for exploratory data analysis to find some unrevealed patterns which are present in data but cannot be categorized clearly. Sets of data can be designated or grouped together based on some common characteristics and termed clusters, the mechanism involved in cluster analysis are essentially dependent upon the primary task of keeping objects with in a cluster closer than objects belonging to other groups or clusters. Depending on the data and expected cluster characteristics there are different types of clustering paradigms. 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 all know about Netflix, the world’s largest on-demand internet streaming media and online DVD movie rental service provider. It was founded on August 29, 1997, in Los Gatos, California by Marc and Reed. It has 69 million members in over 60 countries enjoying more than 100 million hours of TV shows and Movies per day. It is a popular entertainment service used by people around the world. This project is to recommend the Netflix TYPE content. This is unsupervised machine learning based project. Through this project, we will introduce the method of clustering. This project and the algorithms would be useful for Netflix when Making recommendations to users.

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

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. The silhouette coefficient is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (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.

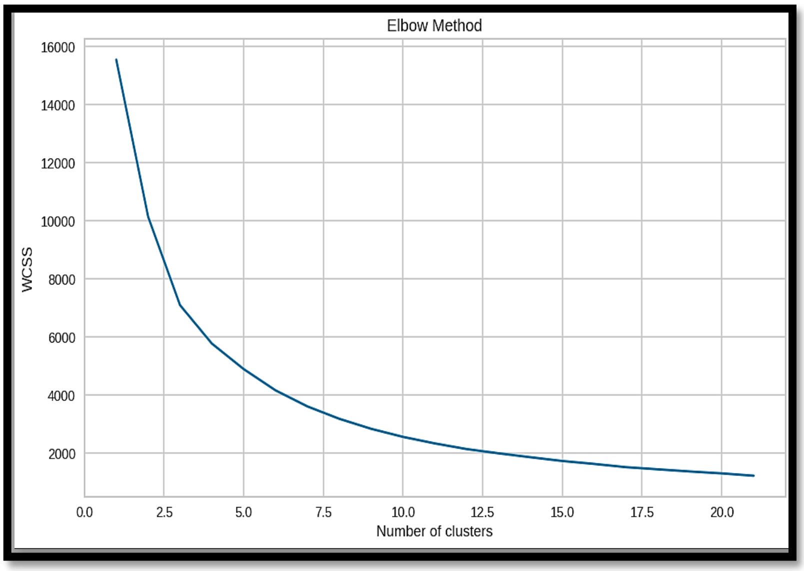
- a(i) is the average distance between i and all the other data points in the cluster to which i belongs.
- b(i) is the average distance from i to all clusters to which i does not belong.
- We will then calculate the average_silhouette for every k.

Average Silhouette = mean{S(i)}

2. ELBOW METHOD:

The elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.
The elbow method runs k-means clustering on the dataset for a range of values of k (say 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").

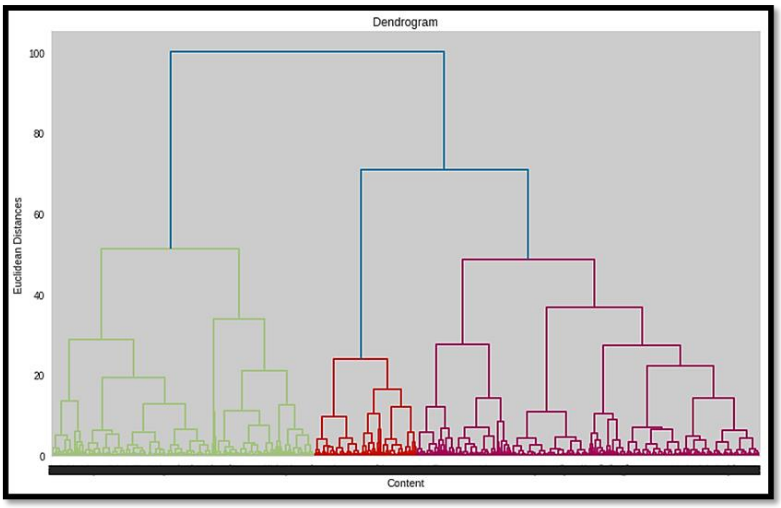


3. DENDOGRAM:

A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data. A dendrogram can be a column graph (as in the image below) or a row graph. Some dendrograms are circular or have a fluid-shape, but software will usually produce a row or column graph.

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

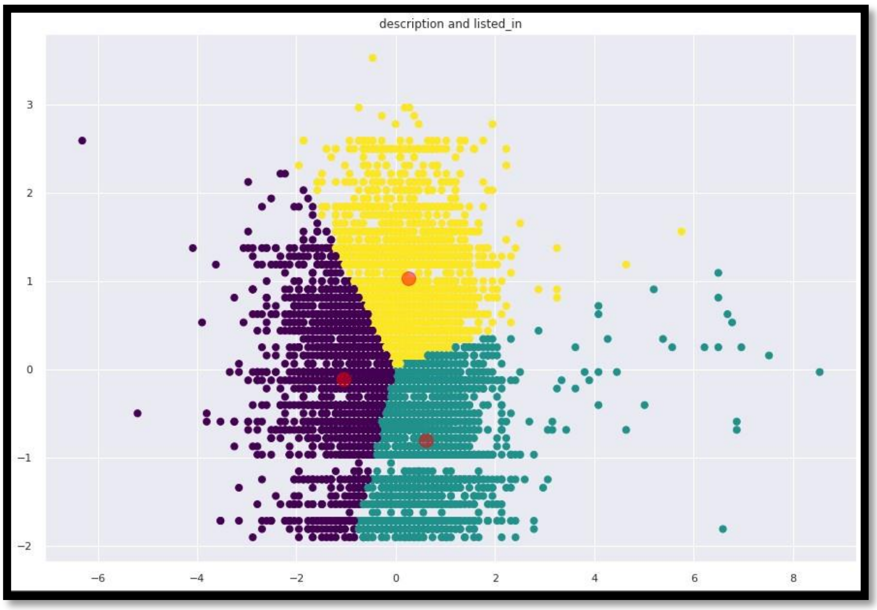
- ➢ The clade is the branch. Usually labelled with Greek letters from left to right (e.g., αβ, δ...).
- ➢ Each clade has one or more leaves. The leaves in the above image are:
- Single (simplicifolius): F
- Double (bifolius): D E
- Triple (trifolious): A B C



CLUSTERING MODELS:

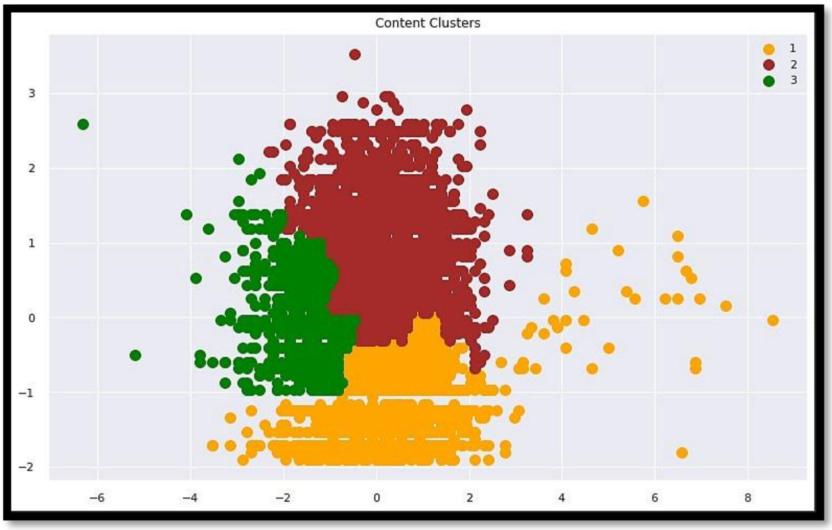
1. K-MEANS CLUSTERING:

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. 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. It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. 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:

Agglomerative Hierarchical Clustering (AHC) is an iterative classification method whose principle is simple. A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data. Agglomerative clustering works in a "bottom-up" manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes).



CONCLUSION:

1. Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we decided to drop this director attribute also duration and show_id attribute because its not useful 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. The United States has the highest number of content on Netflix by a huge margin followed by India.

4. Anupam Kher has acted in the highest number of films on Netflix. Documentaries are the most popular genre followed by Stand-up comedy.

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

6. The number of releases have significantly increased after 2015 and have dropped in 2021 because of Covid 19.

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

8. The second thing we did was feature engineering, which involved removing certain variables and preparing a dataframe to feed the clustering algorithms.

9. By applying the silhouette score method for n range clusters on dataset we got best score which is 0.348 for clusters = 3, it means content explained well on their 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