**Netflix Movies and TV Shows Clustering**

**By Harshavardhan M. Shete**

OTT platforms have become a normal and usual mode of entertainment in today’s world. There are many popular OTT platforms like Amazon Prime, BigFlix, Arre, Discovery+, etc., competing with each other to increase their user base. Netflix is one of the leading OTT platforms, not only in India but also internationally. The success of the OTT platforms depends on two things- the variety of content and appropriate recommendations to the users. Clustering is a useful technique to achieve the best possible recommendations and increase the viewership of the platform.

**Problem Statement:**

The dataset consists of tv shows and movies available on Netflix till 2019. The dataset is collected from Flixable which is a third-party Netflix search engine. The task is to perform EDA (exploratory data analysis) to understand patterns and features of the dataset and cluster the content into a suitable number of clusters using an appropriate clustering algorithm.

The dataset provided contains 7787 rows and 12 columns.

The following are the columns in the dataset:

* Show id: Unique identifier of the record in the dataset
* Type: Whether it is a TV show or movie
* Title: Title of the show or movie
* Director: Director of the TV show or movie
* Cast: The cast of the movie or TV show
* Country: The list of the country in which a show/ movie is released or watched
* Date added: The date on which the content was onboarded on the Netflix platform
* Release year: Year of the release of the show/ movie
* Rating: The rating informs about the suitability of the content for a specific age group
* Duration: Duration is specified in terms of minutes for movies and in terms of the number of seasons in the case of TV shows
* Listed in: This columns species the category/ genre of the content
* Description: A short summary about the storyline of the content

**Solution:**

1. Exploratory Data Analysis
2. Clustering of the dataset

**1. Exploratory Data Analysis**

Exploratory Data Analysis (EDA) as the name suggests, is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. It also helps to understand the relationship between the variables (if any) and it will be useful for feature engineering. It helps to understand data well before making any assumptions, to identify obvious errors, as well as better understand patterns within data, detect outliers, anomalous events, find interesting relations among the variables.

Explorations and visualizations are as follows:

1. Proportion of type of content
2. Country-wise count of content
3. Top 10 countries with highest content production
4. Proportion of TV shows and movies in top 10 countries with maximum content
5. Rating-wise content count
6. Count of content appropriate for different ages
7. Age appropriate content count in top 10 countries with maximum content
8. Proportion of movies and TV shows content appropriate for different ages
9. Trend of year-wise content release
10. Trend of year-wise content on-boarded on Netflix
11. Relation between month and content onboarding on Netflix
12. Top 10 genres in movie content
13. Top 10 genres in TV content
14. Number of seasons vs. number of shows

**2. Clustering of the dataset**

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. Clustering has a large number of applications in various domains. A recommendation engine, market segmentation, social network analysis, search result grouping, image segmentation, anomaly detection, etc. are the few most popular applications. There are various clustering algorithms like K-Means clustering, Mean-shift clustering, Density-based spatial clustering of applications with noise (DBSCAN), Expectation maximization clustering using Gaussian Mixture Models (GMM), agglomerative hierarchical clustering, etc. These algorithms differ in the way in which they define similarity between the data points.

**K-Means Clustering**

K-Means is probably the most well-known clustering algorithm.

1. To begin, we first select a number of classes/groups to use and randomly initialize their respective center points. To figure out the number of classes to use, it’s good to take a quick look at the data and try to identify any distinct groupings.
2. Each data point is classified by computing the distance between that point and each group center and then classifying the point to be in the group whose center is closest to it.
3. Based on these classified points, we recompute the group center by taking the mean of all the vectors in the group.
4. Repeat these steps for a set number of iterations or until the group centers don’t change much between iterations. You can also opt to randomly initialize the group centers a few times, and then select the run that looks like it provided the best results.

K-Means has the advantage that it’s pretty fast, as all we’re really doing is computing the distances between points and group centers; very few computations! It thus has a linear complexity O(n). On the other hand, K-Means has a couple of disadvantages. Firstly, you have to select how many groups/classes there are. This isn’t always trivial and ideally, with a clustering algorithm, we’d want it to figure those out for us because the point of it is to gain some insight from the data. K-means also starts with a random choice of cluster centers and therefore it may yield different clustering results on different runs of the algorithm. Thus, the results may not be repeatable and lack consistency. Other cluster methods are more consistent.

**The optimal value of “K”**

The number of clusters (k) is the most important hyperparameter in K-Means clustering. If there is no idea about the optimal value of k, then there are various methods to find the optimal/best value of k. Two such widely used methods are Elbow method and the Silhouette method.

**Elbow Method:**

Elbow Method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. Pick the value of k, where the average distance falls suddenly.



According to graph above, elbow is at 4, thus optimal value of ‘k’ is 4.

**Silhouette Method:**

The silhouette Method is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. by providing a succinct graphical representation of how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Silhouette coefficients are computed for each of point, and averaged out for all the samples to get the silhouette score.

******

According to graph above, the peak is at ‘4 and thus optimal value of ‘k’ is 4.

**References:**

* Analyticsvidhya.com
* Towardsdatascience.com
* Gdcoder.com