Capstone Project- Submission

NETFLIX MOVIES & TV SHOWS CLUSTERING

GitHub Link: -

Ranjit Biswal: -https://github.com/Ranjitcnb/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING

Suvendu Dey: - https://github.com/devsuvendu/Netflix-Movies-and-TV-Shows-Clustering.git

Abhishek Kumar: - https://github.com/Developer-AD/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING

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 Fixable 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.

Introduction

Netflix's recommendation system helps them increase their popularity among service providers as they help increase the number of items sold, offer a diverse selection of items, increase user satisfaction, as well as user loyalty to the company, and they are very helpful in getting a better understanding of what the user wants. Then it's easier to get the user to make better decisions from a wide variety of movie products. With over 139 million paid subscribers (total viewer pool -300 million) across 190 countries, 15,400 titles across its regional libraries and 112 Emmy Award Nominations in 2018 — Netflix is the world's leading Internet television network and the most-valued largest streaming service in the world. The amazing digital success story of Netflix is incomplete without the mention of its recommender systems that focus on personalization. There are several methods to create a list of recommendations according to your preferences. You can use (Collaborative-filtering) and(Content-based Filtering) for recommendation.

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 increasingly focused on TV rather than movies in recent years?

. Clustering similar content by matching text-based features.

Objective

Netflix Recommender recommends Netflix movies and TV shows based on a user's favourite movie or TV show. It uses a Natural Language Processing (NLP) model and a K-Means Clustering model to make these recommendations. These models use information about movies and TV shows such as their plot descriptions and genres to make suggestions.

The motivation behind this project is to develop a deeper understanding of recommender systems and create a model that can perform Clustering on comparable material by matching text-based attributes. Specifically, thinking about how Netflix create algorithms to tailor content based on user interests and behaviour.

Data Description

Attribute Information:

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
- **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
- **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

Approach

As the problem statement says, understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that we used Affinity Propagation, Agglomerative Clustering, and K-means Clustering.

Tools Used

The whole project was done using python, in google Collaboratory. Following libraries were used for analysing the data and visualizing it and to build the model to predict the Netflix clustering

- Pandas: Extensively used to load and wrangle with the dataset.
- Matplotlib: Used for visualization.
- Seaborn: Used for visualization.
- Nl-tk: It is a toolkit build for working with NLP.
- Datetime: Used for analyzing the date variable.
- Warnings: For filtering and ignoring the warnings.

- NumPy: For some math operations in predictions.
- Word cloud: Visual representation of text data.
- Sklearn: For the purpose of analysis and prediction.

1. Handling missing values:

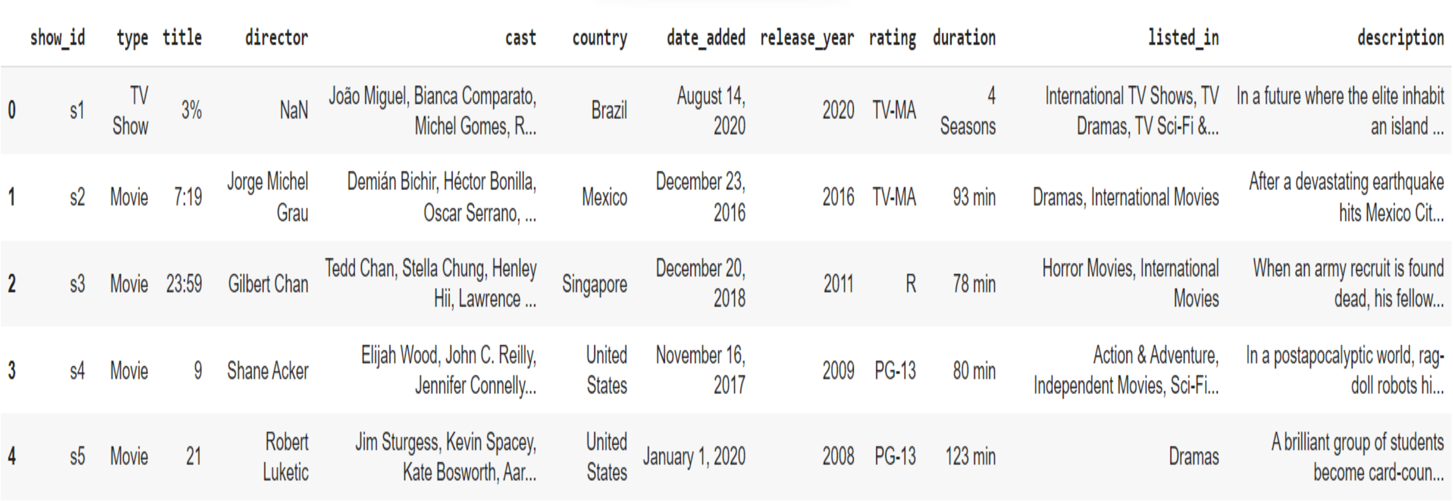
We will need to replace blank countries with the mode (most common) country. It would be better to keep director because it can be fascinating to look at a specific filmmaker's movie. As a result, we substitute the null values with the word 'unknown' for further analysis.

There are very few null entries in the date added fields thus we delete them.

2. Duplicate Values Treatment:

Duplicate values dose not contribute anything to accuracy of results.

Our dataset dose not contains any duplicate values.



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

Table1. The above table shows the dataset in the form of Pandas Data Frame

3. Natural Language Processing (NLP) Model:

For the NLP portion of this project, I will first convert all plot descriptions to word vectors so they can be processed by the NLP model. Then, the similarity between all word vectors will be calculated using cosine similarity (measures the angle between two vectors, resulting in a score between -1 and 1, corresponding to complete opposites or perfectly similar vectors). Finally, I will extract the 5 movies or TV shows with the most similar plot description to a given movie or TV show.

4. Exploratory Data Analysis:

Exploratory Data Analysis (EDA) as the name suggests, is used to analyse 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.

After mounting our drive and fetching and reading the dataset given, we performed the Exploratory Data Analysis for it.

To get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc has been analysed in this step.

Explorations and visualizations are as follows:

I. Proportion of type of content

II.Country-wise count of content

III.Total release for last 10 years.

IV.Type and Rating-wise content count

V.Top 10 genres in movie content

VI. Top 10 Actors on Netflix.
VII. Length distribution of movies.
VIII. Season-wise distribution of TV shows.
IX. Count of content appropriate for different ages.
X. Age-appropriate content count in top 10 countries with maximum content.
XI. Proportion of movies and TV shows content appropriate for different ages.
XII. Season wise distribution of TV shows.
XIII. Longest TV shows.
XIV. Top 10 topics on Netflix.
XV. Extracting the features and creating the document term Metrix.
XVI. Topic modelling using LDA.
XVII. Most important features of topic.

5. Missing or Null value treatment:

In datasets, missing values arise due to numerous reasons such as errors, or handling errors in data.

We checked for null values present in our data and the dataset contains a null value.

In order to handle the null values, some columns and some of the null values are dropped.

7. Tf-idf vectorization:

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit a machine learning algorithm for prediction.

We have also utilized the PCA because it can help us improve performance at a very low cost of model accuracy. Other benefits of PCA include reduction of noise in the data, feature selection (to a certain extent), and the ability to produce independent, uncorrelated features of the data.

So, it's essential to transform our text into tf-idf vectorizer, then convert it into an array so that we can fit into our model.

- **Finding number of clusters**

The goal is to separate groups with similar characteristics and assign them to clusters.

We used the Elbow method and the Silhouette score to do so, and we have determined that 28 clusters should be an optimal number of clusters.

- **Fitting into model**

In this task, we have implemented a K means clustering algorithm. K-means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabelled data into a predetermined number of clusters based on similarities (k).

8. Data Pre-processing:

Removing Punctuation: Punctuations does not carry any meaning in clustering, so removing punctuations helps to get rid of unhelpful parts of the data, or noise.

Removing stop-words: Stop-words are basically a set of commonly used words in any language, not just in English. If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

Stemming: Stemming is the process of removing a part of a word, or reducing a word to its stem or root. Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

9. Clustering:

Cluster formally, clustering is the task of grouping the population of unlabelled data points into clusters in a way that data points in the same cluster are more similar to each other than to data points in other clusters. The clustering task is probably the most important in unsupervised learning, since it has many applications.

for example:

- **Data analysis:** often a huge dataset contains several large clusters, analysing which separately, you can come to interesting insights.
- **Anomaly detection:** as we saw before, data points located in the regions of low density can be considered as anomalies
- **Semi-supervised learning:** clustering approaches often helps you to automatically label partially labelled data for classification tasks.
- **Indirectly clustering tasks (tasks where clustering helps to gain good results):** recommender systems, search engines, etc.
- **Directly clustering tasks**: customer segmentation, image segmentation, etc.

Building a clustering model

Clustering models allow you to categorize records into a certain number of clusters. This can help you identify natural groups in your data.

Clustering models focus on identifying groups of similar records and labelling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. In fact, you may not even know exactly how many groups to look for.

This is what distinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict.

These models are often referred to as **unsupervised learning** models, since there is no external standard by which to judge the model's classification performance.

10. Topic Modelling:

- **Latent Dirichlet Allocation (LDA)**

LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

11. Clusters Model Implementation

1. K-means Clustering

2. Hierarchical clustering

1. K-means Clustering

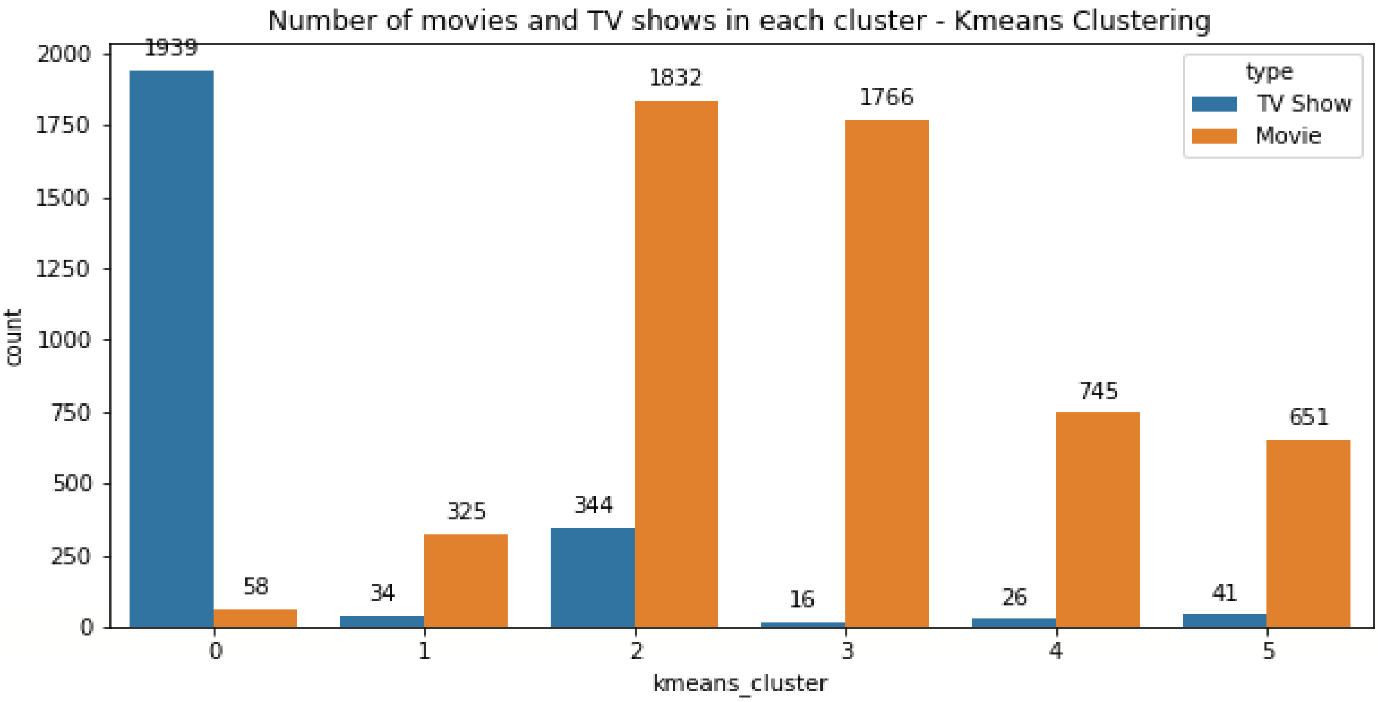
K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

K-means algorithm works:

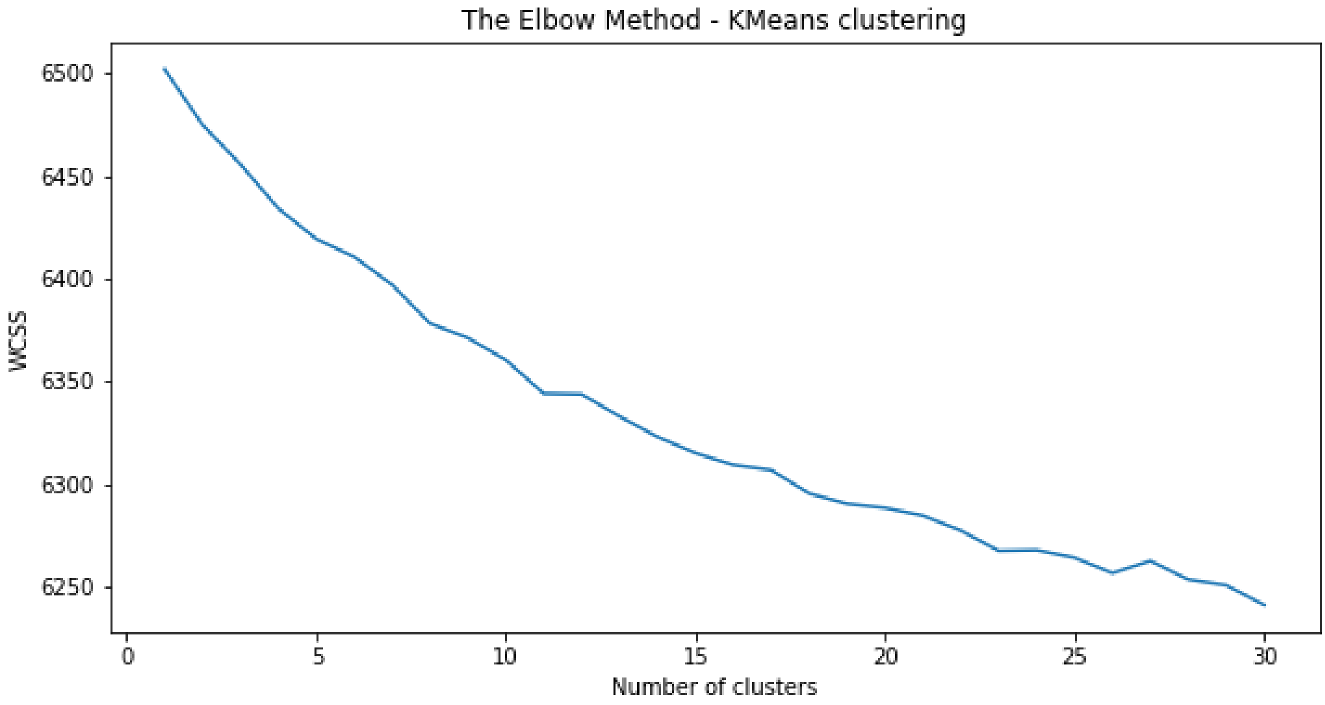
To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

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



Elbow Curve:



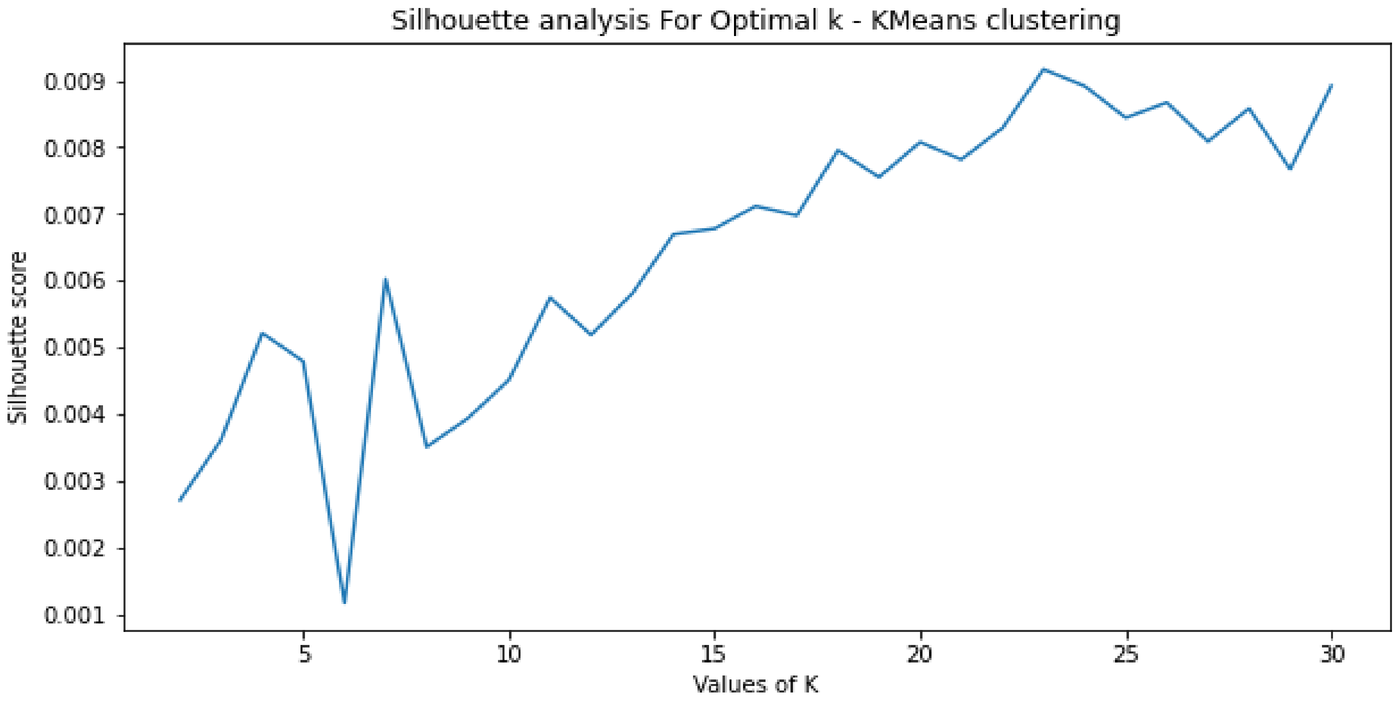
Elbow Curve is one of the most popular methods to determine this optimal value of k.

The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

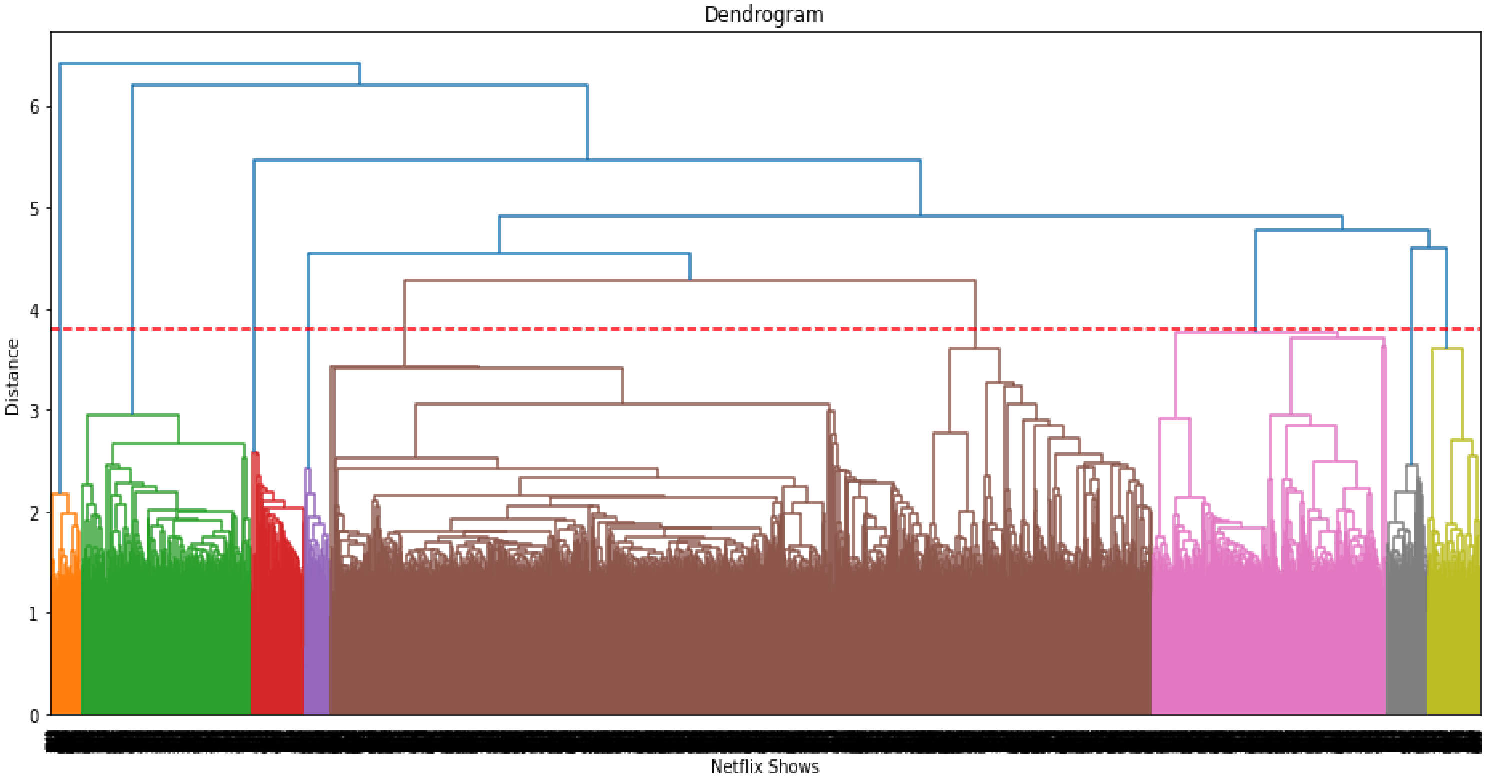
➢ Silhouette's Coefficient-

If the ground truth labels are not known, the evaluation must be performed utilizing the model itself. The Silhouette Coefficient is an example of such an evaluation, where a more increased Silhouette Coefficient score correlates to a model with better-defined clusters. The Silhouette Coefficient is determined for each sample and comprised of two scores

- Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b. The Silhouette Coefficient s for a single sample is then given as:$s = \frac{b-a}{max(a,b)}$



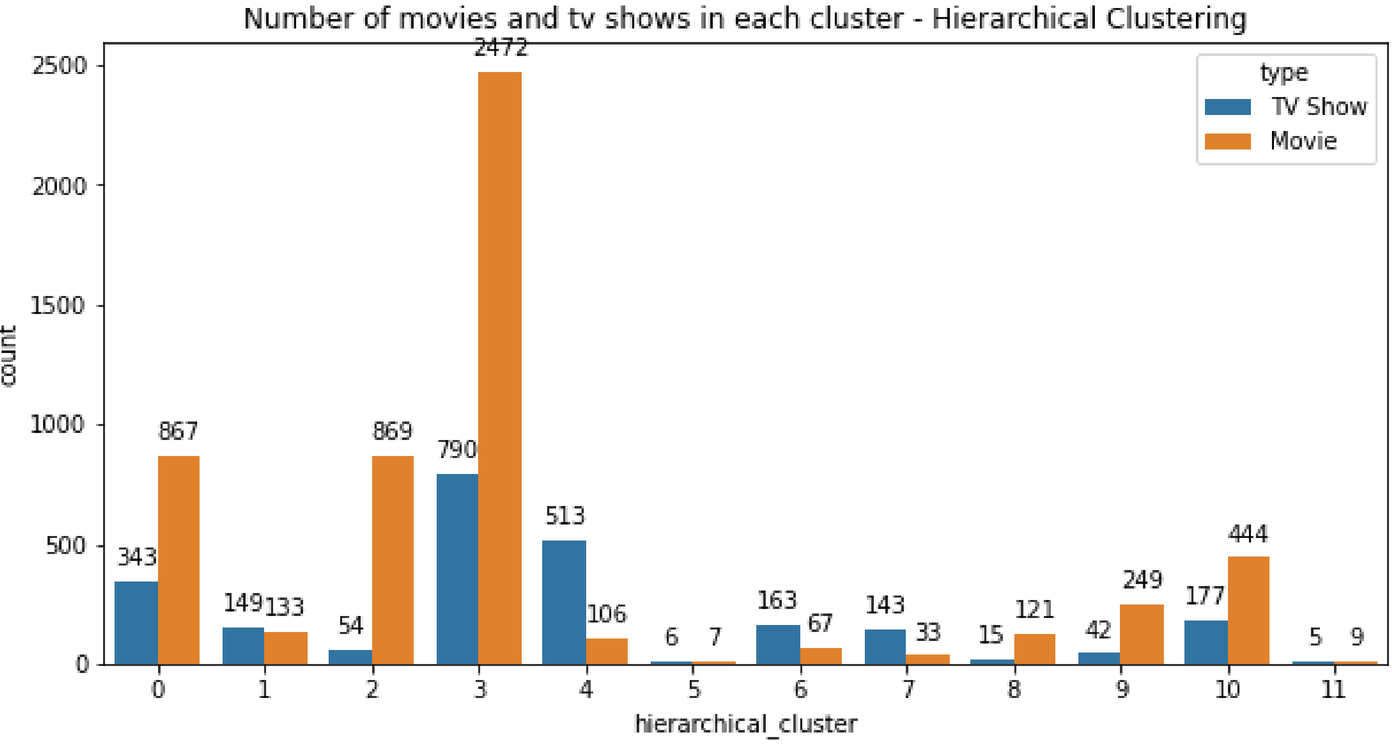
2.Hierarchical-clustering



Hierarchical Clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Hierarchical Clustering can be performed with either a distance matrix or raw data. When raw data is provided, the software will automatically compute a distance matrix in the background.

The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. ... Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.



Conclusion: -

1.In this project, we worked on a text clustering problem wherein we had to classify/group the Netflix shows into certain clusters such that the shows within a cluster are similar to each other and the shows in different clusters are dissimilar to each other.

2.The dataset contained about 7787 records, and 12 attributes. We began by dealing with the dataset's missing values and doing exploratory data analysis (EDA).

3.It was found that Netflix hosts more movies than TV shows on its platform, and the total number of shows added on Netflix is growing exponentially. Also, majority of the shows were produced in the United States, and the majority of the shows on Netflix were created for adults and young adults age group.

4.Once obtained the required insights from the EDA, we start with Pre-processing the text data by removing the punctuation, and, stop words. This filtered data is passed through TF - IDF Vectorizer since we are conducting a text-based clustering and the model needs the data to be vectorized in order to predict the desired results.

5.It was decided to cluster the data based on the attributes: director, cast, country, genre, and description. The values in these attributes were tokenized, pre-processed, and then vectorized using TFIDF vectorizer.

6.Through TFIDF Vectorization, we created a total of 20000 attributes. We used Principal Component Analysis (PCA) to handle the curse of dimensionality. 4000 components were able to capture more than 80% of variance, and hence, the number of

components were restricted to 4000. We first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to be 6. This was obtained through the elbow method and Silhouette score analysis.

7.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.

8.A content-based recommender system was built using the similarity matrix obtained after using cosine similarity. This recommender system will make 10 recommendations to the user based on the type of show they watched.