**NETFLIX MOVIES & TV SHOWS CLUSTERING**

Biswajit gochhayat

DATA SCIENCE TRAINEE

ALMABATTER, BANGLORE

ABSTRACT:

Netflix is a company that manages a large collection of TV shows and movies, streaming them anytime 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.

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 (a total viewer pool of -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 recommendations.

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.

**IN THIS PROJECT, REQUIRED TO DO THE:**

1. Exploratory Data Analysis
2. Understanding what type of content is available in different countries
3. Is Netflix increasingly focused on TV rather than movies in recent years?
4. Clustering similar content by matching text-based features.

OBJECTIVE:

NetflixRecommender recommends Netflix movies and TV shows based on a user's favorite 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 creates algorithms to tailor content based on user interests and behavior.

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
* **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 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 summary of the storyline of the content

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 analyzing the data and visualizing it and building 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.
* Nltk: It is a toolkit built 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.
* Wordcloud: Visual representation of text data.
* Sklearn: For analysis and prediction.
* **Table1*.*** *The above table shows the dataset in the form of Pandas Data Frame*

STEPS OF PROJECT:

The following steps are involved in the project

1. STEPS OF PROJECT:

We will need to replace blank countries with the mode (most common) country. It would be better to keep the 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.

1. DUPLICATE VALUES TREATMENT:

Duplicate values do not contribute anything to the accuracy of results.

****Our dataset does not contain any duplicate values

1. EXPLORATORYORY 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, detected outliers, and anomalous events, and 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 an 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 have been analyzed in this step.

Explorations and visualizations are as follows:

1. Proportion of type of content
2. Country-wise count of content
3. Total release for the last 10 years.
4. Type and Rating-wise content count
5. Top 10 genres in movie content
6. Top 20 Actors on Netflix.
7. Length distribution of movies.
8. Season-wise distribution of TV shows.
9. Count of content appropriate for different ages.
10. Age-appropriate content counts in the top 10 countries with maximum content.
11. Proportion of movies and TV shows content appropriate for different ages.
12. Season-disease distribution of TV shows.
13. Longest TV shows.
14. Top 10 topics on Netflix.
15. Extracting the features and creating the document term matrix.
16. Topic modeling using LDA and LSA.
17. Most important features of the topic.
18. MISSING OR NULL VALUES 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. To handle the null values, some columns and some of the null values are dropped.

1. HYPOTHESIS FROM THE DATA VISUALIZED:

Hypothesis testing is done to confirm our observation about the population using sample data, within the desired error level. Through hypothesis testing, we can determine whether we have enough statistical evidence to conclude if the hypothesis about the population is true or not.

We have performed hypothesis testing to get insights into the duration of movies and content concerning different variables.

6.TFIDF VECTORIZATION:

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform the 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 a TF-IDF vectorizer, then convert it into an array so that we can fit it into our model.

* **Finding the 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 the 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 unlabeled data into a predetermined number of clusters based on similarities (k).

7. DATA PREPROCESSING:

* **Removing Punctuation**: Punctuations do 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 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.

8. CLUSTERING:

Clustering (also called cluster analysis) is a task of grouping similar instances into clusters. More formally, clustering is the task of grouping the population of unlabeled 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, analyzing 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 anomalies

**• Semi-supervised learning:** clustering approaches often helps you to automatically label partially labeled 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 labeling 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. 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.

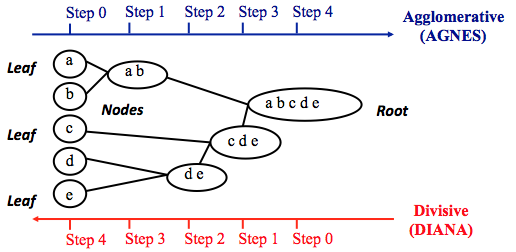
These models are often referred to as **unsupervised learning** models since there is no external standard by which to judge the model's classification performance.

9. CLUSTER MODEL IMPLEMENTATION:

1. AGGLOMERATIVE CLUSTERING
2. K-MEAN CLUSTERING.
3. Hierarchical clustering
4. AGGLOMERATIVE CLUSTERING:

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). This procedure is iterated until all points are member of just one single big cluster (root) (see figure below).

The inverse of agglomerative clustering is divisive clustering, which is also known as DIANA (Divise Analysis) and it works in a “top-down” manner. It begins with the root, in which all objects are included in a single cluster. At each step of iteration, the most heterogeneous cluster is divided into two. The process is iterated until all objects are in their own cluster (see figure below).



## Steps to agglomerative hierarchical clustering

## We’ll follow the steps below to perform agglomerative hierarchical clustering using R software:

1. Preparing the data
2. Computing (dis)similarity information between every pair of objects in the data set.
3. Using linkage function to group objects into hierarchical cluster tree, based on the distance information generated at step 1. Objects/clusters that are in close proximity are linked together using the linkage function.
4. Determining where to cut the hierarchical tree into clusters. This creates a partition of the data.

We’ll describe each of these steps in the next section.

1. K-mean 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. ***Figure3.* Ideal clustering**

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

We created the sample data using build blobs and used range n\_clusters to

specify the number of clusters we wanted to utilize in k means.

Silhouette score and visualization

For clusters = 2 The average silhouette score is : 0.7049787496083262

For clusters = 3 The average silhouette score is : 0.5882004012129721

For clusters = 4 The average silhouette score is : 0.6505186632729437

For clusters = 5 The average silhouette score is : 0.56376469026194

For clusters = 6 The average silhouette score is : 0.4504666294372765

1. HIERERCHICAL CLUSTERING

Also called **Hierarchical cluster analysis** or **HCA**is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom.

For e.g: All files and folders on our hard disk are organized in a hierarchy.

The algorithm groups similar objects into groups called ***clusters***. The endpoint is a set of clusters or groups*,*where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

This clustering technique is divided into two types:

1. Agglomerative Hierarchical Clustering
2. Divisive Hierarchical Clustering

### **Agglomerative Hierarchical Clustering**

The Agglomerative Hierarchical Clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It’s also known as AGNES (Agglomerative Nesting). It's a “[bottom-up](https://en.wikipedia.org/wiki/Top-down_and_bottom-up_design)” approach: **each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.**

**How does it work?**

1. Make each data point a single-point cluster → forms N clusters
2. Take the two closest data points and make them one cluster → forms N-1 clusters
3. Take the two closest clusters and make them one cluster → Forms N-2 clusters.
4. Repeat step-3 until you are left with only one cluster.

Have a look at the visual representation of Agglomerative Hierarchical Clustering for better understanding:

There are several ways to measure the distance between clusters in order to decide the rules for clustering, and they are often called Linkage Methods. Some of the common linkage methods are:

* **Complete-linkage**: the distance between two clusters is defined as the *longest* distance between two points in each cluster.
* **Single-linkage**: the distance between two clusters is defined as the *shortest* distance between two points in each cluster. This linkage may be used to detect high values in your dataset which may be outliers as they will be merged at the end.
* **Average-linkage**: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.
* **Centroid-linkage:** finds the centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.

The choice of linkage method entirely depends on you and there is no hard and fast method that will always give you good results. Different linkage methods lead to different clusters.

**What is a Dendrogram?**

A Dendrogram is a type of tree diagram showing hierarchical relationships between different sets of data.

As already said a Dendrogram contains the memory of hierarchical clustering algorithm, so just by looking at the Dendrgram you can tell how the cluster is formed.

**Conclusion:**

Tailored recommendations can be made based on information about movies and TV shows. In addition, similar models can be developed to provide valuable recommendations to consumers in other domains.

* We've done null value treatment, feature engineering, and EDA since loading the dataset then completed assigned tasks.
* Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it and we have 10 features for the further implementation.
* We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies)
* Most films were released in the years 2018, 2019, and 2020 and and united nation have the maximum content on Netflix.
* The months of October, November, December and January had the largest number of films and television series released
* On Netflix, Dramas genre contains the maximum content among all of the genres and the most of the content added in December month and less content in February.
* By applying the silhouette score method for n range clusters on dataset we got best score which is 0.244 for 3 clusters it means content explained well on their own clusters, by using elbow method after k = 3 curve gets linear it means k = 3 will be the best cluster
* By applying different clustering algorithms to our dataset. We get the optimal number of clusters is equal to 4.