**Capstone Project-4 Submission**

**NETFLIX MOVIES & 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.

**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.

**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.

**In this project, you 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?
4. 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 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 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

**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 to build the model to predict the Netflix clustring

* Pandas: Extensively used to load and wrangle with the dataset.
* Matplotlib: Used for visualization.
* Seaborn: Used for visualization.
* Nltk: 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.
* Wordcloud: Visual representation of text data.
* ****Sklearn: For the purpose of analysis and prediction.

**Steps Involved**

The following steps are involved in the project

**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.

**3. 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. 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 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 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 count in top 10 countries with maximum content.
11. Proportion of movies and TV shows content appropriate for different ages.
12. Season wise distribution of TV shows.
13. Longest TV shows.
14. Top 10 topics on Netflix.
15. Extracting the features and creating the document term metrix.
16. Topic modeling using LDA and LSA.
17. Most important features of topic.

**4. 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.

**5. Data Preprocessing:**

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.

Removing non-ASCII characters-

Remove stop words and lower case:-

Remove punctuations-

Lemmatization:-

Tokenization:-

Vectorization:-

Dimensionality reduction using PCA:-

We find that 100% of the variance is explained by about ~7500 components.

Also, more than 80% of the variance is explained just by 4000 components.

we can take the top 4000 components, which will still be able to capture more than 80% of variance.

**6. 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 as 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. 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.

**7. Topic Modeling:**

1. Select the attributes based on which you want to cluster the shows
2. Text preprocessing: Remove all non-ascii characters, stopwords and punctuation marks, convert all textual data to lowercase.
3. Lemmatization to generate a meaningful word out of corpus of words
4. Tokenization of corpus
5. Word vectorization
6. Dimensionality reduction
7. Use different algorithms to cluster the movies, obtain the optimal number of clusters using different techniques

8.Build optimal number of clusters and visualize the contents of each cluster using wordclouds.

We will cluster the shows on Netflix based on the following attributes:

* Director
* Cast
* Country
* Listed in (genres)

# **8. Clusters Model Implementation**

1. K-means Clustering
2. Hierarchical Clustering

**8.1 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.

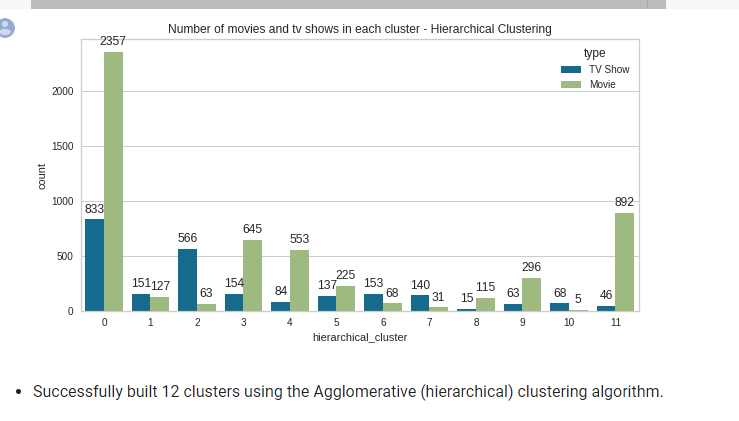
k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

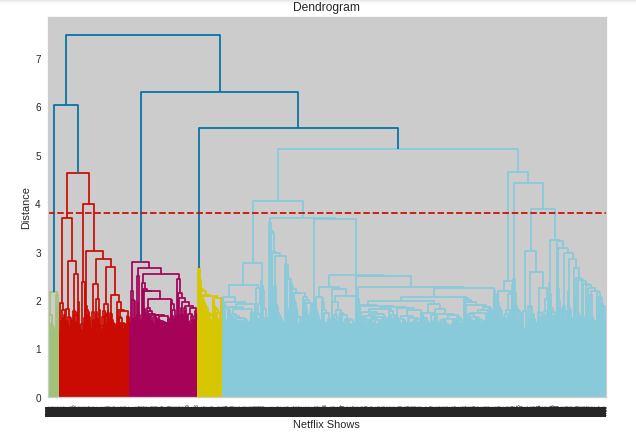
We created the sample data using build blobs and used range n\_clusters to

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

Silhouette score and visualization

**8.2. Hierarchical Clustering:**

* Clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters were built after visualizing the dendogram.
* From the dendogram, at an Euclidean distance of 3.8 units, 12 clusters can be built. Hence the number of clusters were taken as 12.
* Algorithm: Agglomerative clustering
* Distance: Euclidean
* Linkage: Ward



**9. Silhouette Coefficient or silhouette score(meaning)**

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. 1: Means clusters are well apart from each other and clearly distinguished. ... a= average intra-cluster distance i.e., the average distance between each point within a cluster.

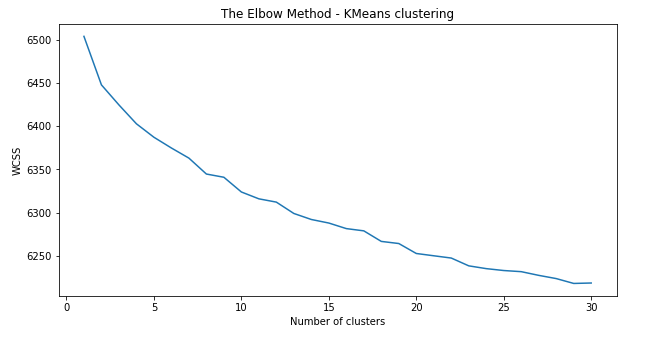
**1. Silhouette’s Coefficient**-

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

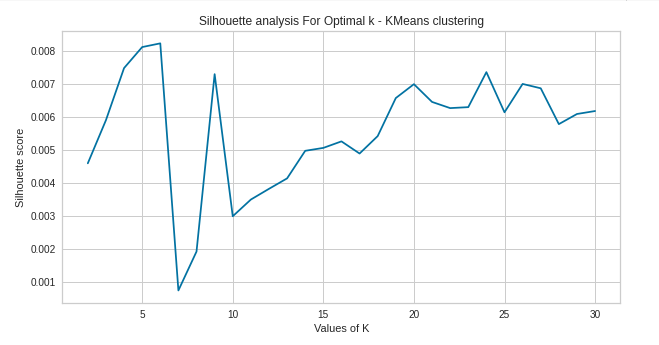
* Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.
* Mean distance between the observation and all other data points of the next nearest cluster. This distance can also 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:

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. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observation belonging to all the clusters:



1. **Elbow Curve:**

****

The 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.

**10.Conclusion**

* 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.
* The dataset contained about 7787 records, and 11 attributes.
* We began by dealing with the dataset's missing values and doing exploratory data analysis (EDA).
* 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.
* It was decided to cluster the data based on the attributes: director, cast, country, genre, and description. The values in these attributes were tokenized, preprocessed, and then vectorized using TFIDF vectorizer.
* 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.
* 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.
* 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.