NETFLIX MOVIES & TV SHOWS CLUSTERING

Vinay V Lanjewar
Data science trainee,
Alma Better, Bangalore

Abstract:

Netflix is one of the leading OTT platforms, not only in India but also internationally

Netflix manages a large collection of TV shows and movies, streaming it anytime via online . The success of the OTT platforms depends on two things- the variety of content and appropriate recommendations to the users. This business is profitable because users make a monthly payment to access the platform. Exploratory Data Analysis is done on the dataset to get the insights from the information however the principal invalid qualities are taken care of. There are 12 features and around 7700 observations in the dataset and are mostly textual features. Clustering is a useful technique to achieve the best possible recommendations and increase the viewership of the platform.

1. Problem Statement

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.

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 has increasingly focusing on TV rather than movies in recent years.

4. Clustering similar content by matching text-based features

2. Dataset Description

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.

The dataset contains following columns:

- **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 content
- **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

3. Steps Involved

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. EDA :

- 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.
- The United States is the most prolific generator of Netflix content, with India and the United Kingdom trailing far behind.

4. 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.

4. 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 .

5. K- means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

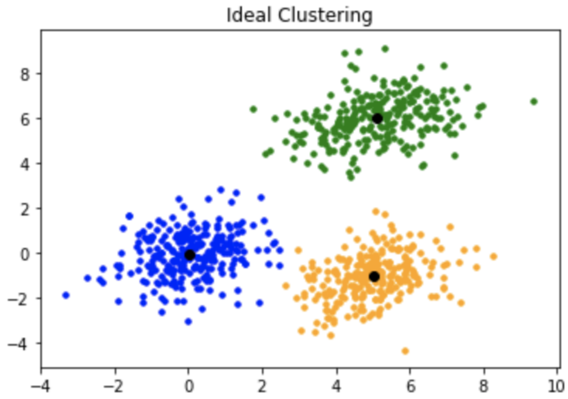
Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

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.

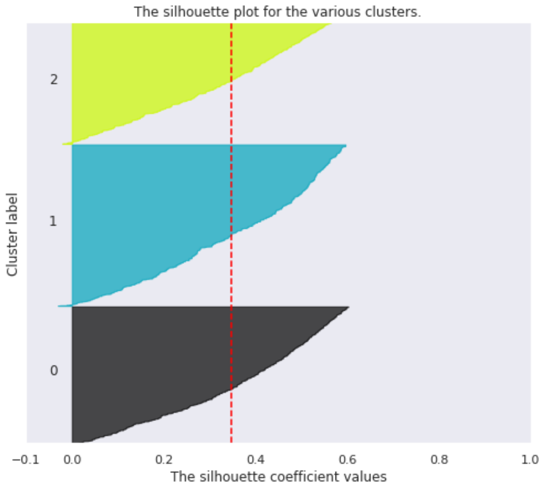
6. Methods to find k value :

1. Sihouette score :

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.

Coefficient s for a single sample is then given as:

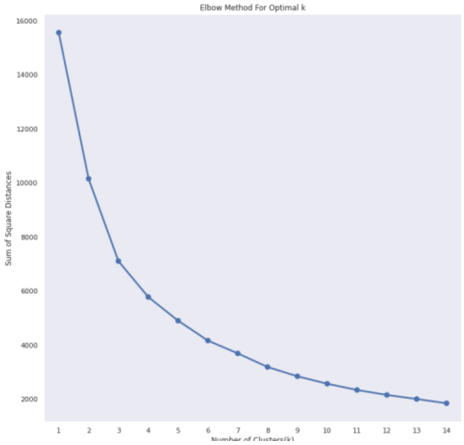
$$s = \frac{b - a}{max(a, b)}$$



2. Elbow Curve :

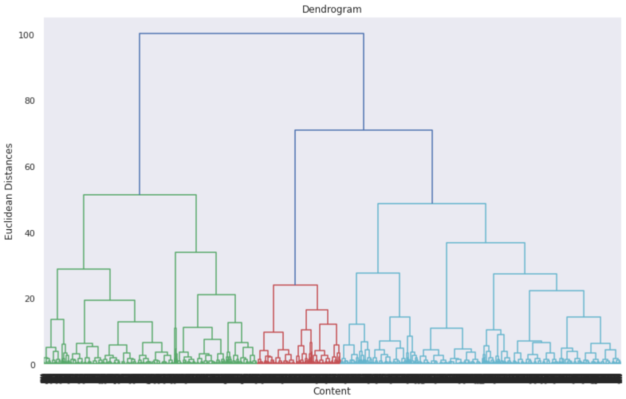
The Elbow Curve is one of the most popular methods to determine this optimal value of k.

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.



3. Dendogram :

The root of the tree (usually the upper or left element) is one large cluster cluster that contains all data points. The leaves (bottom or right elements) are tiny clusters, each of which contains only one data point. According to the generated dendrogram, you can choose the desired separation into any number of clusters.



7. Conclusion :

- 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)
- By analysing the content added over years we get to know that in recent years netflix is focusing movies than TV shows (movies is increased by 80% and TV shows is increased by 73% compare to 2016 data)
- The most number of the movies and TV shows release in 2017 and 2020 respectively and united nation have the maximum content on netflix
- 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 0.348 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
- Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangments
- By applying different clustering algorithms to our dataset ,we get the optimal number of cluster is equal to 3.