**Netflix Movies and TV Shows Clustering**

**Krushna Chaure**

**Data Science Trainee,**

**AlmaBetter, Bangalore**

**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:**

The dataset is collected from Flixable which is a third-party Netflix search engine. Netflix is the world's largest online streaming service provider, with over 220 million subscribers as of 2022-Q2. It is crucial that they effectively cluster the shows that are hosted on their platform in order to enhance the user experience, thereby preventing subscriber churn. We will be able to understand the shows that are similar to and different from one another by creating clusters, which may be leveraged to offer the consumers personalized show suggestions depending on their preferences. The goal of this project is 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.

**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:**

* Exploratory Data Analysis.
* Understanding what type content is available in different countries.
* Is Netflix has increasingly focusing on TV rather than movies in recent years.
* Clustering similar content by matching text-based features.

**Dataset Description:**

**1. show\_id :** Unique ID for every Movie / Tv Show

**2. type :** Identifier - A Movie or TV Show

**3. title :** Title of the Movie / Tv Show

**4. director :** Director of the Movie

**5. cast :** Actors involved in the movie / show

**6. country :** Country where the movie / show was produced

**7. date\_added :** Date it was added on Netflix

**8. release\_year :** Actual Release year of the movie / show

**9. rating :** TV Rating of the movie / show

**10. duration :** Total Duration - in minutes or number of seasons

**11. listed\_in :** Genere

**12. description:** The Summary description

**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.

**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.

**Textual Data Pre-processing:**

**1. Removing Stop words**

• Stop words are common words like “the”, “and” and “but” do not carry much meaning on their own and are often seen as noise in the data

**2. Lowercasing words**

• Lowercasing the words can also reduce the size of the vocabulary, which can make it easier to work with larger texts or texts in languages with a high number of inflected forms.

**3. Removing Punctuation**

• Punctuation marks like periods, commas, and exclamation points can add noise to the data and can sometimes be treated as separate tokens, which can affect the performance of NLP models.

**4. Lemmatization**

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item.So it links words with similar meanings to one word.

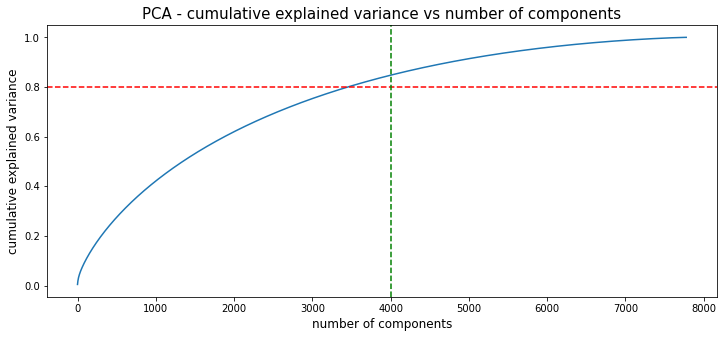
It takes into consideration the context of the word to determine which is the intended meaning the user is looking for. This process allows to decrease noise and speed up the user's task.

**5. Tokenization of corpus and Word vectorization – TFIDF**

• This is important in NLP tasks because most machine learning models expect numerical input and cannot work with raw text data directly. Word vectorization allows you to input the words into a machine learning model in a way that preserves the meaning and context of the words.

**6. Dimensionality reduction – PCA**

• Dimensionality reduction is the process of reducing the number of features or dimensions in a dataset while preserving as much information as possible. As highdimensional datasets can be difficult to work with and can sometimes suffer from the curse of dimensionality.

****

• As you can see that 100% of the variance is explained by about ~7500 components.

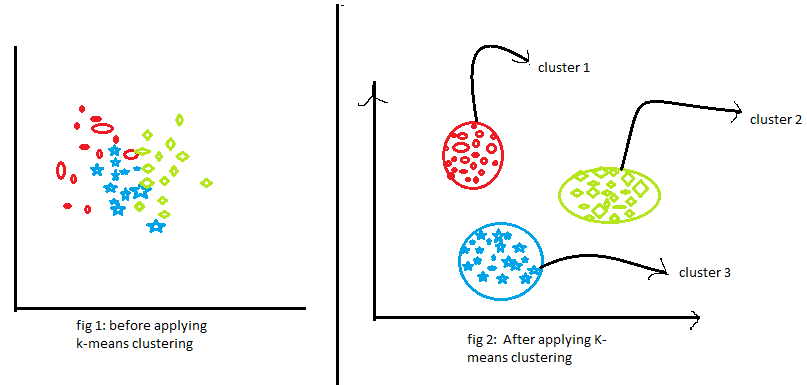
• Also, more than 80% of the variance is explained just by 3000 components.

• Hence to simplify the model, and reduce dimensionality, we take top 3000 components, which will still be able to capture more than 80% of variance.

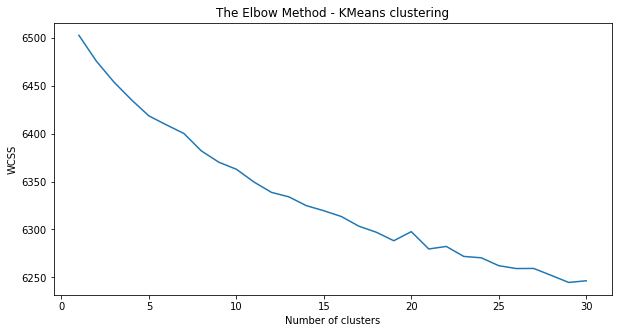
**Clustering Algorithms:**

**K-Means Clustering**

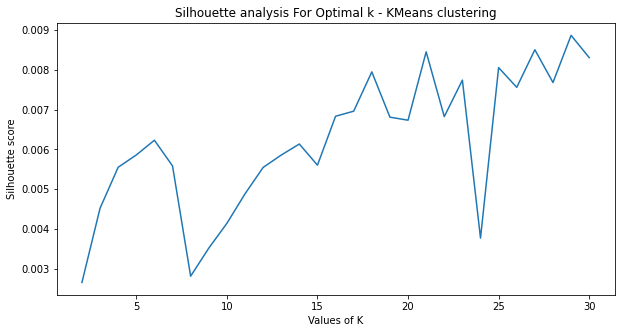
k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



Visualizing the elbow curve and Silhouette score to decide on the optimal number of clusters for Kmeans clustering algorithm

****

• The sum of squared errors between each point and the centroid in a cluster decreases with the increase in the number of clusters.

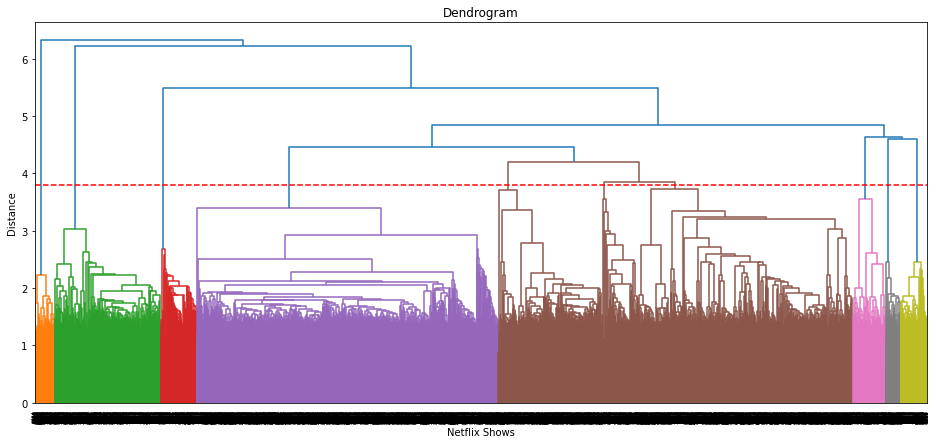
****

• The highest Silhouette score is obtained for 5 clusters.

• Building 5 clusters using the k-means clustering algorithm.

**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



• Visualizing the dendrogram to decide on the optimal number of clusters for the agglomerative (hierarchical) clustering algorithm.

• At a distance of 4 units, 7 clusters can be built using the agglomerative clustering algorithm.

**Recommendation System**

• We can build a simple content-based recommender system based on the similarity of the movie/shows.

• If a person has watched a show on Netflix, the recommender system must be able to recommend a list of similar shows that s/he likes.

• To get the similarity score of the shows, we can use cosine similarity.

• The similarity between two vectors (A and B) is calculated by taking the dot product of the two vectors and dividing it by the magnitude value. We can simply say that the CS score of two vectors increases as the angle between them decreases.

**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.

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.
3. We began by dealing with the dataset's missing values and doing exploratory data analysis (EDA).
4. 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.
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, preprocessed, and then vectorized using TFIDF vectorizer.
6. Through TFIDF Vectorization, we created a total of 20000 attributes.
7. 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.
8. 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.
9. Then clusters were built using the Agglomerative hierarchical clustering algorithm, and the optimal number of clusters came out to be 12. This was obtained after visualizing the dendrogram.
10. 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.

**Acknowledgement**

This project was completed by Krushna Chaure. I am extremely grateful to the celebrated authors whose precious works have been consulted and referred to in this project work. I also wish to convey my appreciation to our peers who provided encouragement and timely support in the hour of need.

**References:**

1. Stackoverflow
2. Medium
3. Kaggle
4. Towardsdatascience
5. Geeksforgeeks