NETFLIX MOVIES AND TV SHOWS CLUSTERING

**Ashfaque sayyed**

**Data science trainee,**

**AlmaBetter, 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**

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. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

**2. Introduction**

Netflix, Inc. is an American technology and media services provider and production company headquartered in Los Gatos, California. Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. The company’s primary business is its subscription-based streaming service, which offers online streaming of a library of films and television series, including those produced in-house.

Netflix is a popular entertainment service used by people around the world. This EDA will explore the Netflix dataset through visualizations and graphs using python libraries, matplotlib, and seaborn.Netflix is known for its strong recommendation engines. They use a mix of content-based and collaborative filtering models to recommend tv shows and movies. In this task, one can create a recommendation engine based on text/description similarity techniques.

I will proceed with reading the data, and then perform data analysis. The practice of examining data using analytical or statistical methods in order to identify meaningful information is known as data analysis. After data analysis, we will find out the data distribution and data types. We will train 3 Clustering ML algorithms to predict the output. We will also compare the outputs.

Let us get started with the project implementation.

**3.Data descriptions**

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

**4. 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 a 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 do not contribute anything to accuracy of results.
* Our dataset does not contains any duplicate values.

**3.Exploratory data analysis :**

* + 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.

1. **Data Pre-processing :**

* 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 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

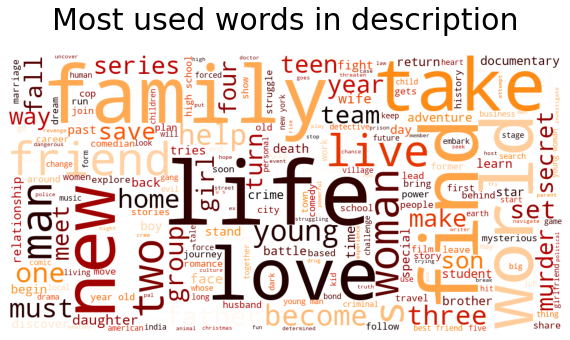
reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

**Word cloud:**

Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it’s mentioned within a given text and the more important it is.



****

**5. 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 help 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 .

**6. K- means Clustering**

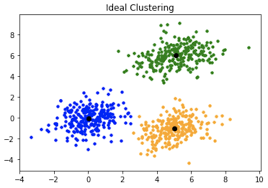
Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

**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 clustering is one of the simplest and popular unsupervised machine learning algorithms.

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. Silhouette score :**

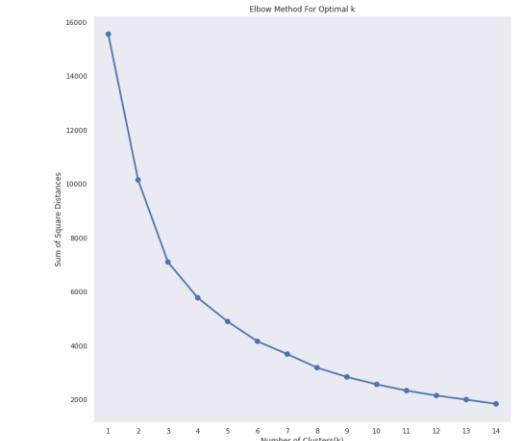
Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each other.

Coefficient s for a single sample is then given as:

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.

1. **Elbow method**

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



The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k. As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.

**6. Conclusion:**

1. In our dataset there is around 69% content as movies and remaining 31% as TV shows. Netflix is releasing more movies than TV shows.
2. we can clearly see that on January October and in December there is more content added on netflix And in February very less amount of content added.
3. Documentaries, stand-up comedy, Dramas and international movies are the top most genres in Netflix.
4. United states has the highest number of content on the netflix, followed by India.
5. US, india, UK, canada, spain etc netflix content is mostly adults.
6. Netflix content is more for teens in india and egypt, followed by south korea.
7. from 2017 number of Movies added increased tremendously, but at the same time TV shows added from 2017 are also increased but as comparison to Movies they are very less in numbers.
8. In cumulative explained variance graph we got 80% of variance captured by 3000 components only, thats why we selected no. of components as 3000.
9. We selected no. of clusters as 15 from Elbow method.
10. Calculated silhouette score for 15 no. of clusters which was showing 0.008.
11. 6. Then we plotted average silhouette score for clusters ranging from 2 to 16, and in that we get silhouette score 0.00708 for cluster=15 which is pretty close to earlier we calculated.

**References-**

1. Geeks for Geeks
2. Analytics Vidhya
3. scikitlearn