**NETFLIX MOVIES & TV SHOWS CLUSTERING**

**Puneet Subhanji**

**Data science trainee,**

**Alma Better, Bangalore**

**Abstract:**

The goal was to anticipate a large amount of comparable content by matching text-based components.

Exploratory Data Analysis is performed on the dataset to extract information from it, while the main invalid features are treated. Moreover, some hypothesis testing was also performed from the encounters from EDA. After that our actual variable should be highlighted in the description part where NLP exercises are performed on it and after that vectorized by using TFIDF.

Starting then and for the foreseeable future, all that remained was to discover the groupings and suit our models by knowing various bunches, and further, the metrics are used to evaluate the model.

Initially, in the 1st step imported the data set in order to do analysis on the data set in order to understand the features of data obtained and also checked for Null values and treated them.

Analyzing all of the variables in the data set and determining the best solution for the specified tasks. Performed the Exploratory data analysis. Performed hypothesis testing to get insights into the duration of movies and content in relation to various variables.

After doing feature engineering and finding the number of clusters, using a model and evaluating it.

**1.Problem Statement**

This dataset contains TV shows and movies that are currently available on Netflix as of 2019. The data was gathered through Flixable, a third-party Netflix search engine.

In 2018, they presented an intriguing research revealing that the number of TV series available on Netflix has nearly quadrupled since 2010. Since 2010, the amount of movies available on the streaming service has reduced by around 2,000 films, while its number of TV shows has nearly tripled.

It will be fascinating to see what additional insights can be extracted from the same information.

Combining this information with additional external datasets like IMDB ratings, rotten tomatoes can also provide many interesting findings.

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

**2. Introduction**

Netflix is one of the world's most popular diversion administrations, with 222 million paying subscribers in more than 190 countries enjoying TV shows, narratives, include movies and portable games across a wide range of genres and languages. Individuals can gaze as much as they want, anytime and wherever they want, on any web-connected screen. Individuals can play, take a break, and then resume viewing without being interrupted by advertising or responsibilities. The goal of our job is to forecast clusters with similar content by comparing text-based features such as column descriptions in a brief visual overview of contents.

**3. Steps involved:**

The following steps are involved in the project

1. **Exploratory Data Analysis**:

We performed the Exploratory Data Analysis after mounting our drive and retrieving and reading the provided dataset.

To gain a knowledge of the data and how the material in the dataset is dispersed, this stage has examined its nature and specifics such as which nations are viewing more and which types of material are in demand.

1. **Missing or Null value treatment:**

In datasets, missing values occur for a variety of causes, including mistakes or data handling issues.

We verified our data for null values and discovered that it contained null values.

Certain columns and some of the null values are omitted in order to manage the null values.

1. **Hypothesis from the data visualized:**

Using sample data, we test hypotheses to corroborate our observations about the population, within the desired error level. Through hypothesis testing, we can examine if we have enough statistical data to infer whether or not the population hypothesis is correct.

We used hypothesis testing to get insight into the duration of movies and content in relation to various variables.

1. **Feature Engineering:**

Initially, To convert a description column to a list, we utilised the .tolist() method, then prior to doing text clustering on the data, we performed several NLP operations on the text columns.

We converted the text to lower case using the lower technique, then we eliminated the stopword and punctuation such as URLs, @handles, etc then tokenized the text for further process.

1. **Tfidf vectorization**

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a typical approach for converting text into a meaningful representation of numbers that may be used to fit a machine learning algorithm for prediction.

We also used PCA since it can assist us boost performance at a very minimal cost in terms of model correctness. Other advantages of PCA include data noise reduction, feature selection (to a certain extent), and the capacity to generate data characteristics that are independent and uncorrelated.

So it's essential to transform our text into tfidf vectorizer, then convert it into an array so that we can fit into our model.

1. **PCA:**

We also used PCA since it can assist us boost performance at a very minimal cost in terms of model correctness. Other advantages of PCA include data noise reduction, feature selection (to a certain extent), and the capacity to generate data characteristics that are independent and uncorrelated.

So it's essential to transform our text into tfidf vectorizer, then convert it into an array so that we can fit into our model.

1. **Finding number of clusters :**

The idea is to distinguish groups with similar features and assign them to clusters.

We used the Elbow method and the Silhouette score to do so, and We determined that 28 clusters are the best number of clusters.

1. **Fitting into model**

In this task, we have implemented a K means clustering algorithm. K-means is a data clustering approach that may be used for unsupervised machine learning. It can sort unlabeled data into a preset number of clusters based on similarities (k).

1. **Model Evaluation**

In this step, we completed our models are evaluated here, and I utilized metrics like: -

1. Silhouette’s coefficient,
2. Calinski-Harabasz Index,
3. Davies-Bouldin Index

**4. Algorithms:**

1. **K Means Clustering:**

k-means is a data clustering approach 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).

Unsupervised algorithms create conclusions from datasets using just input vectors and no prior knowledge, or labeled, outcomes. A cluster refers to a collection of data items that have been aggregated together due to certain commonalities. You’ll determine a target number *k*, which directs to the number of centroids you require in the dataset. A centroid is the imaginary or real location representing the center of the cluster. By lowering the in-cluster sum of squares, each data point is assigned to one of the clusters. In other words, the K-means algorithm determines *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The *‘means’* in the K-means refers to averaging of the data; that is, discovering the centroid.

**5. Model performance:**

Model can be evaluated by various metrics such as:

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 calculated for each sample and is made up of two scores:

* The Silhouette Coefficient is calculated for each sample and is made up of two scores.
* The mean distance between a sample and all other points in the *next nearest cluster*.

The Silhouette Coefficient *s* for a single sample is then given as:

.

1. **Calinski-Harabasz score**

If the ground truth labels are not known, to evaluate the model, utilize the Variance Ratio Criterion, where a higher Calinski-Harabasz score relates to a model with better-defined clusters. The index is the ratio of total dispersion between clusters to total dispersion within clusters for all clusters (where dispersion is defined as the sum of distances squared).

Formula given;

**CH(k)=[B(k)/W(k)]×[(n−k)/(k−1)]**, where

n = data point

k = clusters

W(k) = within cluster variation

B(k) = between cluster variation.

1. **Davies-Bouldin index**

If the ground truth labels are not known, the Davies-Bouldin index can be used to assess the model, where a lower Davies-Bouldin index correlates to a model with better separation between the clusters.

This index represents the average ‘similarity’ between clusters, where the similarity is a metric that relates the distance between clusters to the size of the clusters themselves.

Zero is the lowest possible score. Greater relative to zero values imply a better division.

**8. Conclusion:**

* Since importing the dataset, we have done null value treatment, feature engineering, and EDA, then fulfilled given tasks.
* Anupam Kher has appeared in more Indian films than anybody else, although he was born in the United States
* The two most popular nations for Netflix are the United States and India.
* It has been determined that Netflix is increasingly focused on movies rather than TV series, particularly after 2014.
* Among different types of content available in different countries, content TV-MA is available in the majority of countries. This might be due to the fact that it indicates that it is just for adult audiences, and the Netflix audience appreciates stuff like this.
* Also explained different clusters based on their content; The K-means clustering technique was used to define 28 clusters, with cluster number nine having the most clusters;
* Finally, create a scatter plot in which we may interact with material related to that cluster.

**REFERENCES:**

<https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c#:~:text=Silhouette%20Coefficient%20or%20silhouette%20score%20is%20a%20metric%20used%20to,each%20other%20and%20clearly%20distinguished.&text=a%3D%20average%20intra%2Dcluster%20distance,each%20point%20within%20a%20cluster>.

<https://machinelearningmastery.com/clustering-algorithms-with-python>

[https://towardsdatascience.com/introduction-to-machine-learning-algorith](https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a)