NETFLIX MOVIES & TV SHOWS CLUSTERING

Anurag Taiskar

Data science trainee,
Alma Better, Bangalore

Abstract:

The objective was to anticipate bunches of comparable substance by matching text-based elements.
Exploratory Data Analysis is done on the dataset to get the insights from the information however the principal invalid qualities are taken care of. Likewise, some hypothesis testing was additionally performed from the experiences from EDA. After that description segment is our objective variable must be highlighted where NLP activities are performed on it and after that vectorized by utilizing TFIDF. From that point forward, all that was left was to track down the clusters and fit our models by knowing various clusters, and further, the model is assessed utilizing the metrics.

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.

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 a subscription streaming service and production company.

Netflix is very popular across the countries and it's one of the most demanding OTT platforms for a variety of content to watch for entertainment consisting of different genres from different countries.
The Aim of our task is to predict clusters with similar content by matching the text based features such as description of columns in a small plot summary of contents.

3. Steps involved:

The following steps are involved in the project

1. **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.

2. **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 values.

   In order to handle the null values, some columns and some of the null values are dropped.

3. **Hypothesis from the data visualized:**
   Hypothesis testing is done to confirm our observation about the population using sample data, within the desired error level. Through hypothesis testing, we can determine whether we have enough statistical evidence to conclude if the hypothesis about the population is true or not.
   We have performed hypothesis testing to get the insights on duration of movies and content with respect to different variables.

4. **Feature Engineering:**
   Initially, We used the ‘tolist()function to convert a description column to a list, then before conducting the text clustering of the data, we did some NLP operations on the text columns.
   We used the lower method to convert the text into lower case then we removed the stopword as well punctuation such as URLs, @handles, etc then tokenized the text for further process.

5. **Tfidf vectorization**
   TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit a machine learning algorithm for prediction.
   We have also utilized the PCA because it can help us improve performance at a very low cost of

model accuracy. Other benefits of PCA include reduction of noise in the data, feature selection (to a certain extent), and the ability to produce independent, uncorrelated features of the data.

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.

6. **Finding number of clusters :**

The goal is to separate groups with similar characteristics and assign them to clusters.

We used the Elbow method and the Silhouette score to do so, and we have determined that 28 clusters should be an optimal number of clusters.

7. Fitting into model

In this task, we have implemented a K means clustering algorithm. K-means is a technique for data clustering 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).

8. Model Evaluation

In this step, we have done Evaluation of our models is performed here where I used metrics like: -

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

4. Algorithms:

1. K Means Clustering:

k-means is a technique for data clustering 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 make inferences from datasets using only input vectors without referring to known, or labeled, outcomes. A cluster directs to a collection of data points aggregated jointly because of specific similarities. 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. Every data point is allocated to each of the clusters by reducing the in-cluster sum of squares. In other words, the K-means algorithm determines *k* number of centroids, and then allocates every data point to the nearest cluster, while maintaining the centroids as small as achievable. 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:

91. **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:

- The mean distance between a sample and all other points in the same class.
- 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:

.

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

92. Calinski-Harabasz score

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

Formula given;

$CH(k)=[B(k)/W(k)]\times[(n-k)/(k-1)]$, where

n = data point

k = clusters

W(k) = within cluster variation

B(k) = between cluster variation.

93. 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 resemblance is a measure that compares the length between clusters with the size of the clusters themselves.

Zero is the lowest possible score. Values more relative to zero indicate a better partition.

8. Conclusion:

- We've done null value treatment, feature engineering, and EDA since loading the dataset then completed assigned tasks.
- In this context, we've noticed that Netflix is increasingly focusing on movies rather than TV shows, especially after 2014.

- We found that different types of content are available in different countries, but TV-MA is the content that is available in the majority of countries. This could be because it shows that it is just for adult audiences, and the Netflix audience enjoys content like this.
- We've also explained different clusters based on their content; we've defined 28 clusters and enforced the K-means clustering algorithm. And then we determined that cluster number nine has the most clusters; we've also plotted a scatter plot in which we may interact with similar content about 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