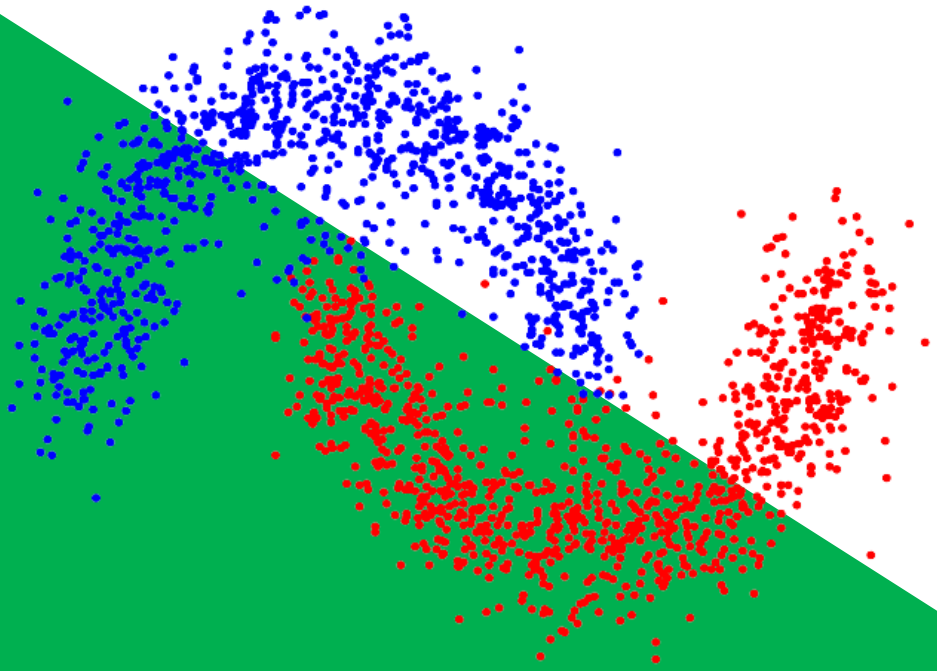
Technical Documentation

NETFLIX

Movies & TV Shows

CLUSTERNING MODEL



AI

BY : SHUBHAM SARTAPE

Abstract

OTT platforms have been a buzz now days. It is a solid source of entertainment which has surpassed almost all the limits. Netflix is one the most well-known OTT platforms in today's era. It contains several movies, TV shows and web series released across the world. By analysing the content of Netflix we can know a lot about choices and tastes which public nurture. It helps us to know where the trend is going.

Problem Statement

We are required to :

- Exploratory Data Analysis
- Understanding what type content is available in different countries
- Has Netflix increasingly focused on TV as compared to movies in recent years?
- Clustering similar content by matching text-based features

Introduction

Netflix releases multiple movies, TV shows and web series internationally across the globe in different genres. We have a dataset which consist of several information along with features about the content release as of on 2019. The dataset

has tv shows and movies listed for more than past 70 years. We have other useful datasets like IMDB Ratings, Rotten Tomatoes which we will use alongside our main dataset. It will provide an immense amount of information on analysing about the contents released.

Exploratory Data Analysis

The primarily thing we will do is importing all the necessary libraries like numpy, pandas, sklearn, seaborn, matplotlib etc. We will be importing these libraries so that we can use it while we work along with our given dataset.

Now we will fetch the dataset and do a basic analysis of it by checking its data type, features and some basic statistics about them. Then we will check the null values for each feature of the dataset. With this we came to know that there 5 features in the dataset which contain null values which needs to be rectified. We see that data added and rating column has only single null, so we will drop their null values. The null values in the 'directors' column will be replaced by unknown, 'country' column by its mode value and replacing the null values in 'cast' column as 'no cast'. On looking at the dataset 'show id' column seems quite unimportant so dropping it will be a wise decision. That is how we made data null values free which subsequently made the dataset much cleaner and crisper.

We should get a count of the number of movies and TV shows respectively and present it in a pie chart. It shows that in the

dataset we have 69% of movies and 31% of TV shows. We will do some more cleaning and wrangling activities and then try to get the number of TV shows and movies each country has released. Now try to find the number of movies and TV shows released by each country separately to get some more clarity on our data. In this way we concluded that United States has the highest number of movies and TV shows released followed by India and United Kingdom respectively.

We will now try to get a list of the number of movies and TV shows released every year and get a graph for it to understand the trend. By observing the trend we came to know that Netflix started releasing movies after 1940s and TV shows after 1980s. Also there seems to be a gradual increase in releases after 2000s where more movies have been released as compared to TV shows.

We will try to create a count plot which will let us know about the count of the number of movies and TV shows in Netflix with respect to months. From the count plot we can conclude clearly that in the month of January, October and December most of the movies and TV shows have been released.

We can here create a dist plot to know about the duration of the movies and it shows that the average duration of movies are 100 minutes. We will also get a list of the number of movies as per their duration in descending order. This shows that the lengthiest movie in the list is ‘Black Mirror:

Bandersnatch'. We can get the same list for TV shows as well which says Grey's Anatomy is the longest TV show with 16 seasons followed by Supernatural with 15 seasons.

It can also be seen here on creating a count plot that most of the movies are in the categories of documentaries followed by Stand-up comedies, dramas and so on. Similar thing can be noticed in case of TV shows as well that most of them are in Kids category followed by International TV shows and so on. One can also see that contents with TV-MA (Matured Audience) rating is highest in number followed by contents with rating TV-14.

Feature Engineering

The data we have needs to be clean which means it should be free of the punctuation marks like comma, semi colon etc. While working with the data those punctuation marks may be a source of concern and annoy for us. Similarly we also need to stop words followed by stemming, lemmatizing and tokenizing the data. This will make the dataset and the data in it, more prominent and informative. We will be creating functions of all these activities respectively and call them one by one to get our required work done. This will provide us the appropriate data as per our need and make our rest of the work comparatively easy and simple. We also will try to get the frequency of the words which will help us later in our project. Then we will apply Principal Component Analysis on our data

to reduce its dimensionality. Out of our selected 2000 components, after dimensionality reduction 1500 components cover 90% of the data.

Clustering

K-Means Clustering

Our first action now will be scaling our data to get it in a proper and convenient form so that we can use it according to our ease. On the dataset now we will apply K=Means Clustering using Elbow method to find considering range of 10 clusters. Plotting it in the form of a graph will give a clear picture of 10 optimal clusters.

For more clarity we will use silhouette to get optimal clusters here. The highest silhouette score will show us the optimal clusters. Here, the highest silhouette score is 0.0510 for 3 clusters. Cluster 4 has the highest text value and it shows that as the clusters increases the silhouette decreases which also means that the quality of clusters decreases.

Hierarchical Clustering:-

Now we will go ahead and try to choose the clusters using dendrogram. We will first create dendrogram using our data to get our required clusters. By observing the data we got 6 clusters with their specific silhouette scores. We will be using DB Scan now to get to know some more aspects of the data. It

shows us that the label of the data is too close to form DB Scan clustering.

On doing all these stuffs along the way we received plenty of information and worked on them to get profound information about it. We can conclude our entire journey throughout in some bullet points as mentioned below.

Tools and Library used

Google Colab - http://colab.research.google.com/

Github : For Version Control and Direct raw file access to csv.

Python Libraries used:

- Pandas
- Numpy
- Plotly Express
- Sklearn
- Matplotlib pyplot
- Seaborn
- Warnings (To remove deprecation warnings)

Conclusion

- There are 69% on movie contents and 31% of TV shows contents in Netflix.
- Contents from top 10 countries covers 69.54% of overall content.
- United States has the highest number of movies released followed by India.
- United States has the highest number of TV shows released followed by United Kingdom.
- Netflix started releasing movies after 1940s and TV shows after 1980s.
- There is relatively a gradual but significance increase in content release after 2000 where movies released seems more than TV shows released.
- In the month of January, October and December maximum number of movies/TV shows are released.
- Black Mirror: Bandersnatch is the longest movie and Grey's Anatomy is the longest TV show released.
- Movies with documentaries and Standard Comedy category and TV shows in Kids category are amongst the highest number of releases.
- Most of the contents have TV-MA (Matured Audience) rating in Netflix.
- After PCA dimensionality reduction to summarize 1500 components to cover 90% of the data.
- In K-Means clustering, Elbow method is used to get the optimal clusters.

- To measure the quality of clusters and how well the clusters are separated silhouette score is being taken into consideration.
- In K-Means clustering highest silhouette score we got was 0.510 for 3 clusters.
- In Hierarchical clustering, dendrogram is used to get the optimal clusters and the number of clusters came out to be 6.
- In Hierarchical clustering, the highest silhouette score we got was 0.505 for 3 clusters.
- We can see that by applying K-Means clustering and Hierarchical clustering we got the best clusters most optimal cluster is equal to 3.