Un-Supervised ML (Clustering) on Netflix Movie and TV Shows.

Soma Pavan Kumar
Data science trainee,
AlmaBetter, Bangalore

Abstract:

In this project, I used a dataset containing all the movies and TV shows available on Netflix as of 2019. I applied two different unsupervised learning algorithms. These clusters can be successfully interpreted, and they appear to be really useful. Not only do we gain new knowledge about the data using them, but they also show that a user recommendation model could be built using this clustering.

Keywords: Data Wrangling, Pandas, Matplotlib, Seaborn, Machine Learning, Natural Language Processing, K-means Clustering.

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 services 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.
satisfying accuracies long before the actual event

2. Introduction

Netflix, Inc. is an American over-the-top content platform and production company headquartered in Los Gatos, California. The company's primary business is a subscription-based streaming service offering online streaming from a library of films and television series, including those produced in-house. The streaming platform has increased his catalogue substantially in his last 10 years of existence. Netflix has way more films than all his competitors, such as HBO or Amazon video, which are following Netflix`s steps in order to obtain the same success.

An element that is closely related to Netflix is IMDb. IMDb (an acronym for Internet Movie Database) is an online database of information related to films, television

programs, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. These ratings are created by the users, and will be used during this project study along with the Netflix catalogue.

3. Data Preprocessing

Data preprocessing is a data mining technique which consists in transforming the data in order to make it understandable. It could be changing the type, the format, splitting the data, verifying that there are no missing values but also creating new columns thanks to columns we initially have. In machine learning, the data processing step is critical because it involves cleaning, integration, transformation, scaling, standardizing data and many other tasks, in order to have a good preparation for the application of models.

To begin we first did some data exploration by checking types, missing values and also duplicate values.

Pandas:

It has functions for analysing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas allows us to analyse big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science. Pandas are also able to delete rows that are not relevant, or contain wrong values, like empty or NULL values. This is called cleaning the data.

Pandas library also provides us with functions such as describe () which provides us with statistics of the data including mean, mode, median and also Interquartile ranges.

Pandas library also gives us some functions like shape, which gives us the number of rows and columns in the whole data. Also, some functions like shape which gives us all the number of cells or elements in the whole of data.

Pandas also has some features like drop () which drops the whole column or a row as specified by the user.

4. Data Visualization

Data Visualization is an interdisciplinary field that deals with graphic representation of data. It is a particular efficient way of communication when the data is numerous as for example a time series

The first plot is maybe one of the most important because it shows all the correlations of the features. To complete this visualization, we created a ranking of the features which are the most correlated to the target. So, it gives an idea of which features we have to focus on.

Moreover, in the notebook, you can see that there are significant red squares in the plot

indicating that there are significant high correlations.

All of the Data Visualizations In this project are done using Matplotlib.pyplot, which is an inbuilt library in python that deals with plotting graphs including, bar graphs, histogram, all the way till pie charts and scatter plots.

Matplotlib:

Matplotlib is a low-level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter.

Matplotlib is open source and we can use it freely. Matplotlib is mostly written in python, a few segments are written in C, Objective-C and JavaScript for Platform compatibility.

Plotting in Matplotlib:

The plot () function is used to draw points (markers) in a diagram. By default, the plot () function draws a line from point to point.

The function takes parameters for specifying points in the diagram. Parameter 1 is an array containing the points on the **x-axis**. parameter 2 is an array containing the points on the **y-axis.**

In this project we've had 8 visualizations:

1. The First one is to find the distributions of the variable type (movie or TV Show).

2. The Second is to find the frequency of different kinds of rating.

3. The Third is to find the relation between Type and Rating for the type movie and TV Shows.

4. The fourth is to plot the distribution of data after dropping the null values.

5. Also, plotting the frequency of the number of titles released per year.

6. Also plotting the frequency of different Genres in titles released by Netflix.

7. Plotting a word Cloud of different Countries.

8. Plotting a Word Cloud for different Directors.

Groupby Function:

A groupby operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts of data and compute operations on these groups.

Any groupby operation involves one of the following operations on the original object. They are −

- Splitting the Object
- Applying a function
- Combining the results

5. Data modelling:

We have a clustering problem because we don't have any target variable. So, the goal

of this part is to apply many algorithms in order to find the algorithm with the best indicator. Let's apply some algorithms to our problem

5.1 Natural Language Processing (NLP):

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyse large amounts of natural language data.

The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.

The NLP (Natural Language Processing) gives us recommendations for the TV Show 'Breaking Bad' are: "Unicorn Store", "Summer Night", "Maniac", "Mission Istanbul" and "Kaalia".

The NLP (Natural Language Processing) gives us the recommendation for the movie "6 underground", are: "machali Jal ki rani Hai", "aaviri", "Summer Night", "History of Joy", "Woody Woodpecker"

The Recommendations provided by the model NLP are not up to the mark and let's proceed to our next model "K-means" Clustering.

5.2 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 (cluster centres or cluster centroid), serving as a prototype of the cluster.

k-means clustering minimizes within-cluster variances (squared Euclidean distances).

The model is tested for the same movie and TV show.

For the TV Show 'breaking bad' the recommendations were: '13 reasons why', 'Another Life', 'Battle Creek'.

For the movie '6 Underground', the recommendations were: 'A Haunted House', 'A Night at the Roxbury'.

6. Conclusion:

The Experiment I chose is to Cluster the movie recommendations from Netflix Movie and TV shows dataset.

The dataset has 12 features on offer.

We applied 2 Machine Learning Algorithms namely:

1. Natural Language Processing (NLP)
2. K-means Clustering

According to the recommendations seen from both the models, the results from the K-means model were very close to the films in terms of the description of the movie.

References:

- Geeks for geeks
- Stack overflow