NETFLIX MOVIES AND TV SHOWS CLUSTERING

Adil Khan, Sunil Panigrahi, Shubham Kumar, Sharaffin B, Vivek Singh
Data Science Trainees, Almabetter

ABSTRACT

Netflix is an American subscription streaming service and production company. It is the one of the largest Platform which provides the collection of TV shows and movies, streaming via online means. The monthly subscription by user makes Netflix a profitable business and the flexibility in subscription users can cancel it anytime. So to engage customers to this platform Netflix must keep their content interesting that can hook users on the platform.

.

PROBLEM STATEMENT

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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. Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings. In this project, we were required to Exploratory Data Analysis, understanding what type content is available in different countries, Is Netflix has increasingly focusing on TV rather than movies in recent years.

INTRODUCTION

Netflix Recommender models are based on a user's favorite movie or TV show. It uses a Natural Language Processing (NLP) model and a K-Means Clustering model to make these recommendations. These models use information about movies and TV shows such as their plot descriptions and genres to make suggestions. The motivation behind this project is to develop a deeper understanding of recommender systems and create a model that can perform Clustering on comparable material by matching text-based attributes. Specifically, thinking about how Netflix create algorithms to tailor content based on user interests and behavior.

DATA SUMMARY

The provided data set has following different columns. They are given below

a) **show_id** : Unique ID for every Movie / Tv Show
b) **type** : A Movie or TV Show
c) **title** : Title of the Movie / Tv Show
d) **director** : Director of the Movie
e) **cast** : Actors involved in the movie / show
f) **country** : Country where the movie / show was produced
g) **date_added** : Date it was added on Netflix
h) **release_year** : Actual Release year of the movie / show
i) **rating** : TV Rating of the movie / show
j) **duration** : Total Duration - in minutes or number of seasons
k) **listed_in :** Genres
l) **description**: The Summary description

TOOLS USED

The whole project was done using python, in google Collaboratory. Following libraries were used for analyzing the data and visualizing it and to build the model to predict the Netflix clustering

- Pandas: Extensively used to load and wrangle with the dataset.
- Matplotlib: Used for visualization.
- Seaborn: Used for visualization.
- Nltk: It is a toolkit build for working with NLP.
- Datetime: Used for analyzing the date variable.
- Warnings: For filtering and ignoring the warnings.
- NumPy: For some math operations in predictions.
- Wordcloud: Visual representation of text data.
- Sklearn: For the purpose of analysis and prediction.

Steps Involved

The following steps are involved in the project

1. Handling missing values:

We will need to replace blank countries with the mode (most common) country. It would be better to keep 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 dose not contribute anything to accuracy of results.

Our dataset does not contains any duplicate values.

DATA PREPROCESSING

- Label Encoding-refers to converting the labels into a numeric form so as to convert them into the machine-readable form.

- Lemmatization- Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. ... For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.

- Removing Stop words - To remove stop words from a sentence, you can divide your text into words and then remove the word if it exits in the list of stop words provided by NLTK.

- Tf - idf Vectorization - TF-IDF stands for “Term Frequency — Inverse Document Frequency”. This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining.

- Min-max Scaling - For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. It preserves shape of original distribution.

EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis (EDA) as the name suggests, is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. It also helps to understand the relationship between the variables (if any) and it will be useful for feature engineering. It helps to understand data well before making any assumptions, to identify obvious errors, as well as better understand patterns within data, detect outliers, anomalous events, find interesting relations among the variables.

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.

Explorations and visualizations are as follows:

a) total releases for last 10 years
b) analysis on release year of tv show
c) analysis on release year of movies
d) analysis based on top 10 country
e) top ten directors
f) top 10 actors
g) top movie ratings based on rating system
h) top tv show ratings based on rating system
i) number of movies and tv shows based on content
j) top ten genre in movies
k) month wise analysis of movie releases
l) most occurred word in titles
m) relation between audience type and rating

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.

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 like Data Analysis, Anamoly detection, Semi-Supervised learning etc

Clusters Model Implementation

1. **Silhouette score**
2. **Elbow Method**
3. **DBSCAN**
4. **Dendrogram**
5. **Agglomerative Clustering**

13. Silhouette Coefficient or silhouette score

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1..

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 determined for each sample and comprised of two scores

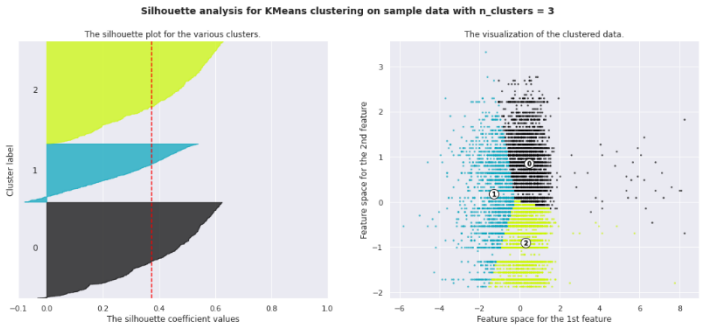
- Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.

- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.

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

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

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. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observation belonging to all the clusters. By this method we have K=3 as the best cluster

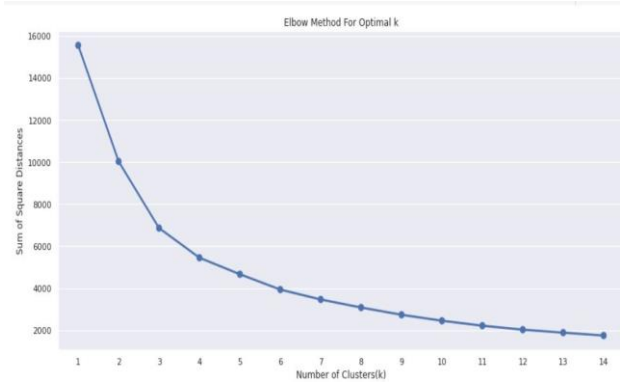


2. Elbow Curve:

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

To determine the optimal number of clusters, we have to select the value of k at the "elbow" ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal

number of clusters for the data is **3**.

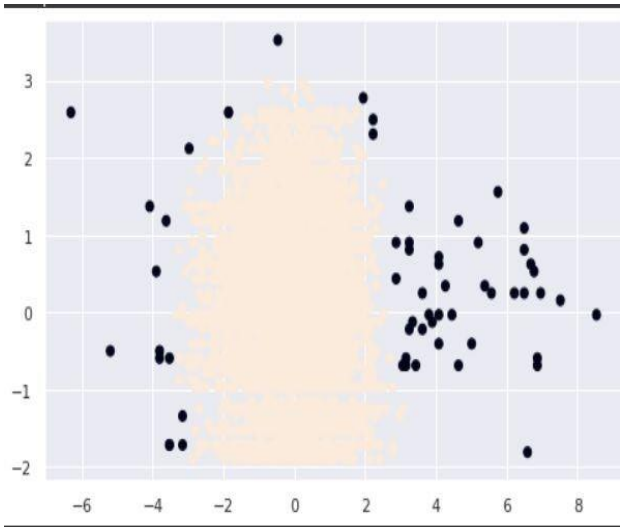


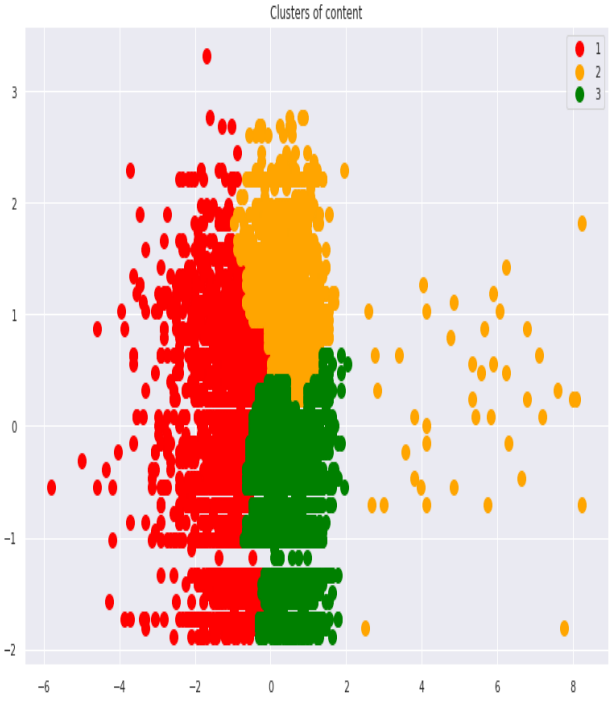
AGGLOMERATIVE CLUSTERING

The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. ... Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.

DBSCAN

The **DBSCAN algorithm** is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points





OBSERVATIONS

1. In this dataset there are two types of contents where 30.86% includes TV shows and the remaining 69.14% carries Movies

2. The popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been increasing significantly

3. The United States produced the highest amount of content followed by India

4. The most popular director on Netflix, with the most titles, is Jan Suter followed by Raul Campos

5. TOP 3 content categories are International movies, dramas, comedies.

6. It seems like words like "Love", "Man", "World", "Story", "Christmas" are very common in titles. The word "Christmas" occurred so many times. The reason maybe those movies released on the month of December. Thus, by creating a monthly column we can safely conclude,

7. More number of movies get released with Christmas in it.

8. The audience prefers TV-MA (adult)and TV-14(Young Adults) shows more and the least preferred rating shows are Nc-17(kids).

CONCLUSION

In text analysis (NLP) we used stop-words, removed punctuations, stemming & TF-IDF vectorizer and other functions of NLP. Applied different clustering models like K-means, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements. By applying the silhouette score method for n range clusters on dataset we got best score which is 0.3746 for 3 clusters it means content explained well on their own clusters, by using elbow method after k = 3 is the best cluster.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Stackoverflow