Netflix Movies and TV Shows Clustering

Potdar Vinayak
Data science trainee
AlmaBetter, Bangalore

Abstract:

Netflix is one of the most loved OTT platforms for watching tv shows and movies. It has been known to be the archive for some of the best content online. The shows and movies are recommended to used based on their previous watches. We will be trying to replicate some part of it.

Our experiment will be trying to cluster the movies and tv shows according to their respective features. We will be focusing on the text based features for the clustering, like the description of the movie or show along with the genre.

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.

In this project, you are required to do

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix has increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features

Attribute Information

1. show_id : Unique ID for every Movie / Tv Show
2. type : Identifier - A Movie or TV Show
3. title : Title of the Movie / Tv Show
4. director : Director of the Movie
5. cast : Actors involved in the movie / show
6. country : Country where the movie / show was produced
7. date_added : Date it was added on Netflix
8. release_year : Actual Releaseyear of the movie / show
9. rating : TV Rating of the movie / show
10. duration : Total Duration - in minutes or number of seasons
11. listed_in : Genre

12. description: The Summary description

2. Introduction

The platform has a wide variety of tv shows and movies from almost every possible genre. Now, all the tv shows and movies cannot be recommended to the users, as it is not possible to show everything on the landing page. We have to cater to the likings of the user and recommend shows and movies that he/she might like. In order to do that we need to understand what are the metrics that will help us in differentiating and clustering them.

Our goal here is to build a clustering model that will be able to cluster similar shows and movies. These clusters can be used to build a recommender system for the platform.

3. Platform Working Dynamics

Nowadays, all the platforms such as Spotify, Prime Music, Prime Video, Netflix start out by asking the newly signed up user his/her interests beforehand. Once the basic interests of the users are recorded, they proceed with associating the user with the content liked by similar users. This will be working in the background where every recommended piece of content will be tracked. Based on the actions of the user, they will be clustered in the similar type of user base. This clustering is for both users and content.

We will be trying to cluster the tv-shows and movies based on the description and genre details.

4. Things in Favor

The dataset was in a fairly good shape to start working with. There was not much content missing from the data.

We did not have to do much cleaning to get the working dataset ready for the models. Since we only required the text-based data to work with, we used the remaining data for the exploratory data analysis.

5. Challenges Faced

For the most part the dataset had almost everything that was required to form clusters. Since this is a unsupervised learning model, we had to check the clusters of the data that we had formed.

The model will cluster the shows and movies into some cluster, but the ambiguity of that show/movie being eligible to join another cluster also exists.

For example, a horror comedy genre movie is by theory a part of horror and comedy at the same time. The model will place this in either one of the clusters which will reduce the exposure of that movie to the other cluster.

6. Steps involved:

- **Null Value Analysis and Treatment.**
  Right after loading the dataset, we went ahead and checked the descriptive information about the data using some basic head(), tail(), describe() functions. Later, we checked for the missing values from the dataset and figured, that four features had some missing values. Only one feature 'director' has the greatest number of missing values. I dropped this column as it wasn't going to help in the clustering model. The rest of the features like the country were replaced with the mode value of the features.

- **Exploratory Data Analysis**
  I had some direction to do the Exploratory Data Analysis, provided in the problem statement. I held the analysis of the data revolving around those points.

  I explored the publishing trends like the highest publishing done in a year, distribution of movies and tv shows on the platform, country wise activity etc. The problem statement wanted to know if the platform has been focusing more on the TV shows lately, which was analyzed comprehensively.

- **Data Preprocessing / Feature Engineering**
  The features that I was supposed to work on were text-based features like description and genre of the tv show/ movie. Here, I got rid of punctuations, stop words and did stemming on the text data.
  I further divided the dataset into two parts, one was only text data and the other had the rest of the data.
  This text data was for the training of the clustering model and the rest of the data was for validation of the models and checking the cluster formations.
  Post this process, I vectorized the text data.

- **Topic Modelling**
  In this process I performed LDA and LSA on the vectorized text data, in order to extract 10 main topics from the data set. This did not mean that the clusters were going to be 10, but it just gave an idea to start with.

- **Clustering**
  Now that I had some direction with the data, I proceeded to find the optimal number of clusters for the model.
  This was done by three methods, viz. Silhouette Analysis, Elbow Method and dendrogram.
  Once I had the optimal number of clusters, it was time to actually form the clusters by fitting a model on the data, which was KMeans model.

- **Cluster Themes Extraction**
  In this step I already had all the three clusters formed. Now it was time to extract the keywords from the clusters which defined the theme of the clusters. I extracted the centroid info and the associated words with every cluster to understand the theme they represent.

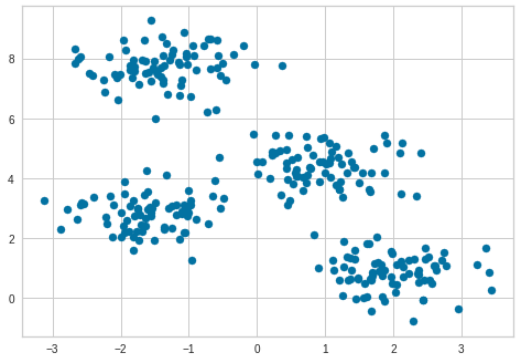
7.1. Algorithm:

1. KMeans Clustering:

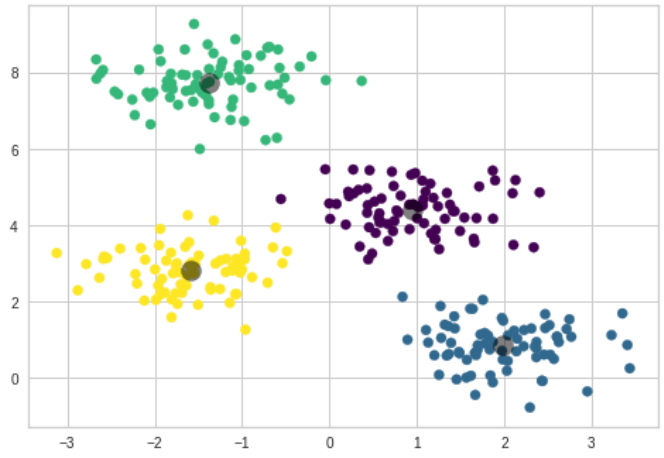
k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:
The "cluster center" is the arithmetic mean of all the points belonging to the cluster.
Each point is closer to its own cluster center than to other cluster centers.

Those two assumptions are the basis of the k-means model.



The data points are clustered into the number of specified clusters by the developer. This sets n number of centroids throughout the space and changes according to the data. These centroids are changed until the mean distances becomes constant.



There are some caveats of this model-

- This may or may not be the global optimal solution.
- Kmeans can be a slow model for large data sets.
- The developer has to select the number of clusters beforehand using some other methods, like silhouette analysis, dendrogram, elbow method etc.

8. Conclusion:

The major 4 tasks that were given have been accomplished.

1- We completed the EDA and found the trends and distribution of the content.
2- US, India and UK are the major content publishers of the platform with more emphasis on the movies than tv-shows.
3- The hypothesis of movies generating more engagement was successfully rejected and a few major countries like UK, Japan and South Korea have been publishing more TV-Shows than Movies. In the overall bigger picture, we can see that, in the recent times the publishing of movies has gone down and tv shows has increased. Here, I can safely state that the platform is increasingly focusing on the TV- Shows than movies.
4- I clustered the data using text-based features like description and genre. The optimal number of clusters turned out to be **3.** The model used for clustering is Kmeans Model.

References-

1. **Almabetter**
2. **GeeksforGeeks**
3. **Analytics Vidhya**