Netflix Movies & TV Shows Clustering

Ankit Rai
Data Science Trainees
AlmaBetter, Bangalore

ABSTRACT

Netflix has become dominant company in the on-demand media industry, with 167 million paying subscribers around the world. By creating compelling original programming, analyzing its user data to serve subscribers better, and above all by letting people consume content in the ways they prefer, Netflix disrupted the television industry and forced cable companies to change the way they do business. Netflix is essentially a storehouse of content, including movies, documentaries, and television series, both pre-existing and its own. For a flat monthly fee, subscribers can consume any program at any time on whatever device they prefer.

Our job is to perform data analysis and to develop an unsupervised model which can perform clustering similar content by matching text-based features.

PROBLEM STATEMENT

We have been provided a dataset 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 data set.

Following will be our objectives:

* To perform Exploratory Data Analysis

* Understanding what type of content is available in different countries

* Is Netflix has increasingly focused on TV rather than movies in recent years.

* Clustering similar content by matching text-based features

ABOUT THE DATA

We have been provided a dataset which contain details of the contents available on Netflix. There are 12 number variables, which are as below:

* 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 Movie

* 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 Releaseyear of the movie / show

* rating : TV Rating of the movie / show

* duration : Total Duration - in minutes or number of seasons

* listed_in : Genere

* description: The Summary description

STEPS INVOLVED

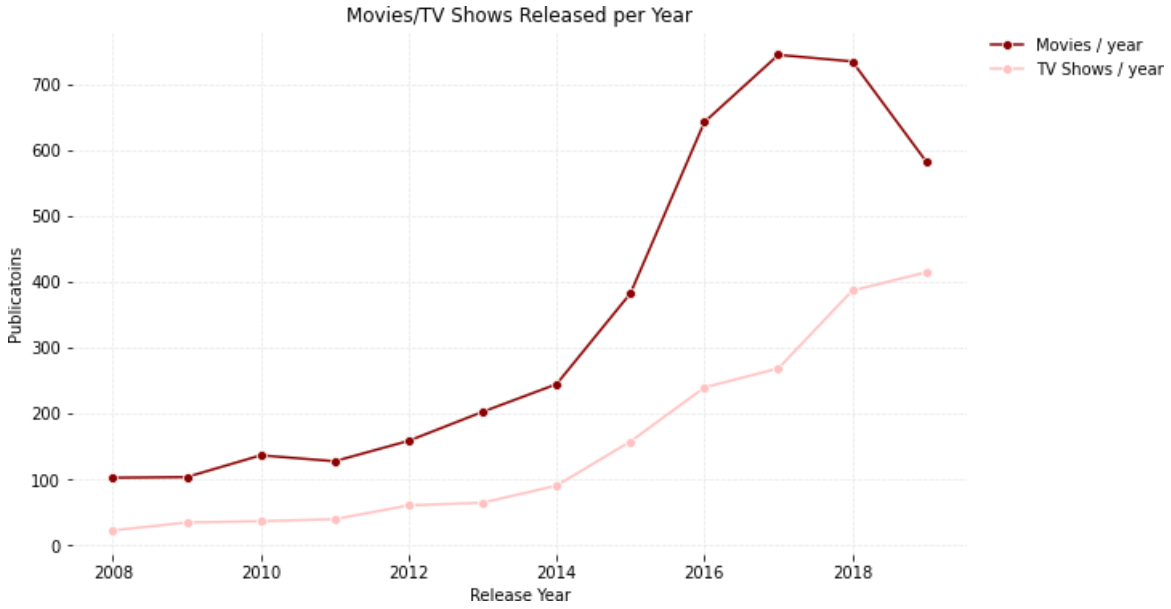
1. **Data Pre-processing**

   We started with checking for null values and found that five variables -"director", "cast", "country", "date_added" and "rating" had null values in them. Dropping whole rows with null values would lead to loss of information. Also, since all these columns have object d-type therefore we cannot simply replace their null values with the mean or median. Better approach would be to replace null values with a clear text indicating "No Data". But before that, we have to get rid of null values in "date_added" column by dropping them as there are only 10 numbers and it would be convenient for us in future while extracting Month and Year from the column.

2. Discovering Information – EDA

Once our data was cleaned, we started analyzing different variables. This process helped us figuring out various aspects and trends among contents on Netflix. We used Bar plots and line graphs to represent the trends. Some of the discovered aspects were:

- ✓ We found that most of the content on Netflix are of TV-MA and TV-14 rating.
- ✓ USA and India are two countries producing the maximum number of contents.
- ✓ Documentaries and Stand-up are top genre in terms of number of contents they have on platform. Further we found number of movies on Netflix outnumbers TV-shows
- ✓ As can be seen in below line chart, there is a decline in publications of Movies in the recent years (After 2016). Netflix is focusing more on TV-shows compared to movies since recent years.



3. Data/Text Preparation for Modelling –

- **Feature Selection:** After extracting useful insights from our data, our next objective was to apply unsupervised algorithm for clustering of similar contents based on text features. Now, since the basis of clusters are texts features, we selected two important variables- “listed_in” and “description”, for the modelling part.

- **Text Cleaning:** We cleaned our text data by following processes:

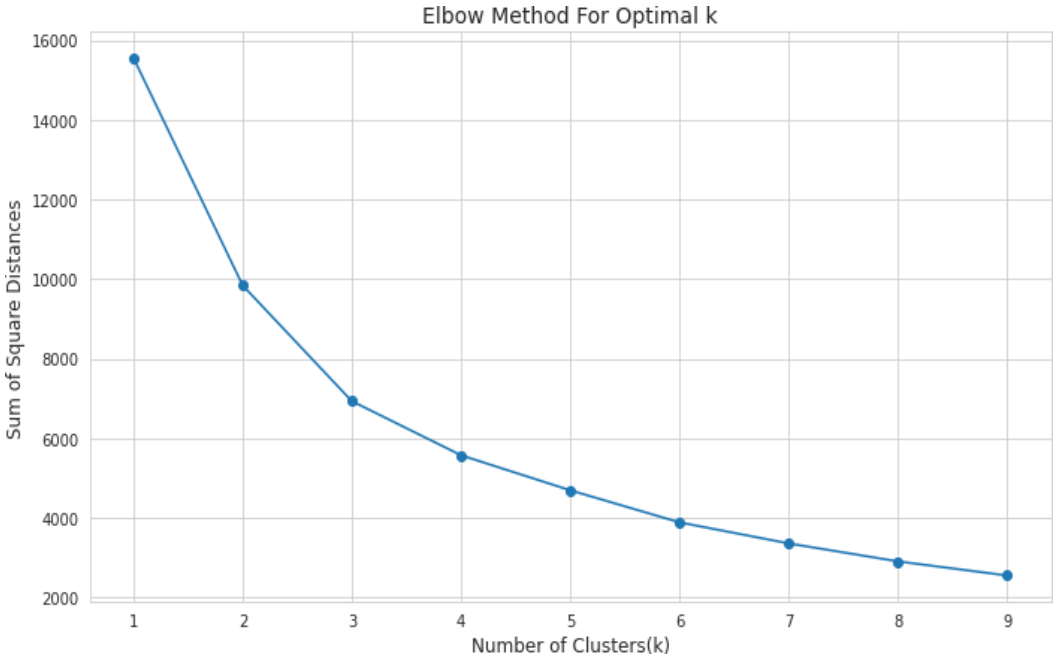
  We processed our text by removing useless characters like - stop words, punctuation and did stemming. After getting the length for each text feature we rescaled them for generalization and started applying algorithms.

CLUSTERING

After getting our text data ready, we started using unsupervised algorithms for clustering.

a) 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 centers or cluster centroid), serving as a prototype of the cluster.



From this chart we need to check, which would be the best number of clusters from 2,3,4,5, and 7.

SILHOUETTE SCORE

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 like 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.

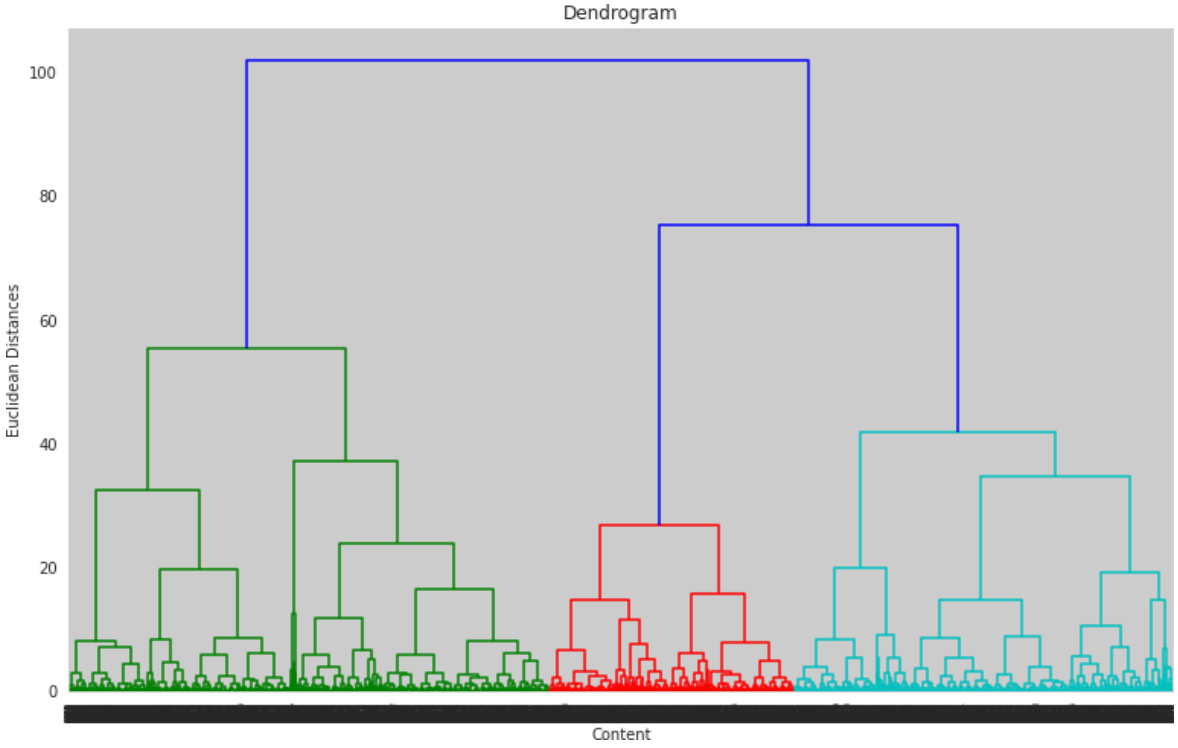
For n_clusters = 2, silhouette score is 0.3551415129065328
For n_clusters = 3, silhouette score is 0.35586172779109915
For n_clusters = 4, silhouette score is 0.32858920525532515
For n_clusters = 5, silhouette score is 0.3348785202036102

So, in this model the 3 clusters are giving best result.
So that we will consider 3 clusters as optimum clusters

We have also done an interactive clustering visualization in the notebook.

b) Hierarchical 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.



The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold.

So we consider, no. of Cluster = 3

Future Scopes

- More Post Cluster Analysis
- Integrate the Netflix dataset with other datasets and present more insights and clusters.
- We could have done some more research on the recommendation system. (Based on TFIDF, rather than cosine similarity)

CONCLUSION

We have reached to the conclusion of our exercise!! We started this project with the intention to identify the useful insights and trends on the Netflix for its contents. For this, we performed exploratory data analysis on our data after cleaning and making it easy to analyze. This analysis gave some useful insights like most of the content on Netflix are of TV-MA and TV-14 rating, USA and India are two countries producing the maximum number of content and Netflix is focusing more on TV Shows compared to the movies in the recent years.
Our next job was to develop a clustering model. For this, we first selected useful variables, and cleaned the text data. After cleaning text data and calculating the remaining text length, we rescaled the data for generalization. We first used K-means clustering method. To find appropriate cluster number, we used elbow method and finally got the best silhouette score of around 0.35 for 3 numbers of clusters. Next, we applied Hierarchal Agglomerative Clustering for which we made dendrogram. We also obtained silhouette score of around 0.32.

With this we achieved our objectives of the project.