UNSUPERVISED MACHINE LEARNING

Netflix Movies and TV Shows Clustering

Divyesh Dhanani, Mrityunjay Singh Chandel, Nimisha Jadhav,

Vishal Chakrabarty, Sagar Tikmani

(DATA SCIENCE TRAINEES)

Almabetter, Bangalore

Abstract

Netflix is an OTT Platform that contains a large collection of TV Shows and Movies that we can access anytime online with a suitable internet connection and a subscription plan. Users can cancel their subscriptions anytime. So, the Company needs to provide content according to the users, so that the users stay hooked and don't cancel the subscription. With the help of Recommender Systems, it provides suggestions suitable to the users.

Introduction

Netflix is a company that offers online subscription service to watch TV shows and movies by streaming the media online. According to the Market Line (2017) Netflix has a subscriber more than 93 million members across 190 countries in the world and their business has three segments which are: domestic streaming, international streaming and domestic DVD. Netflix is an interesting company to study in because Netflix makes a disruptive innovation which changes how people rent a movie into streaming the movie online (Richardson, 2011). Netflix can see the opportunity of DVD subscription business will change into stream subscription more than 10 years ago that other company cannot predict. Netflix become very popular with their unofficial slogan "Netflix & Chill" that become the pop culture in recent years.

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 2000 titles since 2010, while its number of TV shows has nearly tripled.

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.

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. Understanding what type of content is available in different countries and Is Netflix increasingly focuses on TV Shows rather than movies in recent years.

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 release year 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.

Netflix Workflow

- Content request made by subscriber.
- Resolvers pass the request to the Netflix domain's authoritative server.
- Requested content is retrieved from an index stored on a database.
- Content is pushed out from a storage location or accelerated service.
- Edge locations determine where to route the request to optimize content delivery.
- The content is streamed to the subscriber.

Steps Involved

The following steps are involved in the project:

1. **Inspection of Data**: It is the process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision making.
2. **Exploratory Data Analysis**: Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
   Explorations and Visualization are as follows:
   - Netflix Content Analysis
   - Growth of Content over years
   - Analysis based on a country
   - Genre vs Country
   - Creating the word cloud to see which appear the most in the titles of the movies and the TV shows.
   - Creating the word cloud to see which appear words appear the most in the description of the movies and TV shows.
   - Duration
   - Top 25 Directors whose content is available on Netflix.
   - Top 10 actors whose content is available in Netflix.
   - Top TV Show ratings.

3. **Data Preprocessing:** Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.
   - Create clusters for our data now using text columns.
   - Create a function to remove the punctuation.
   - Removing the stop words.
   - Creating the cluster.
   - Dimensionality reduction.

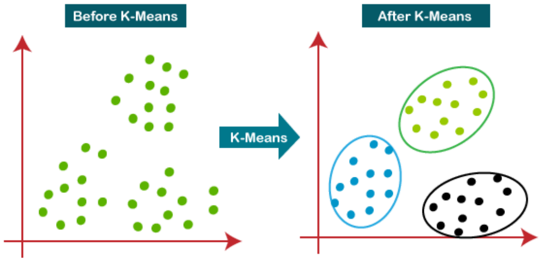
4. **K-Means Clustering:** It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. It allows us to

cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

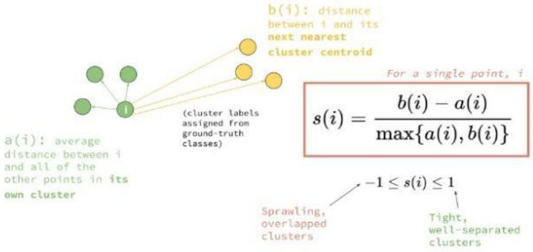
The k-means algorithm mainly performs two tasks:

- Determines the best value for K centre points or centroids by an iterative process.
- Assigns each data points to its closest k-centre. Those data points which are near to the particular k-centre, create a cluster.

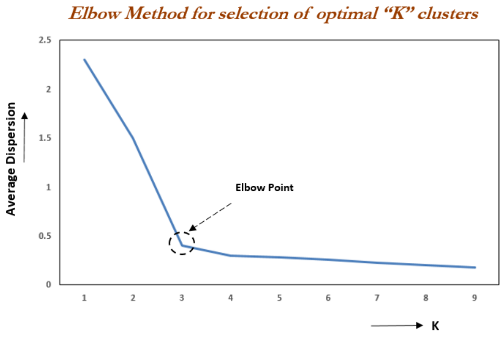


5. **Methods to select Optimum Value**

- **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: Means clusters are well apart from each other and clearly distinguished.


b(i): distance between i and its next nearest cluster centroid
i
(cluster labels assigned from ground-truth classes)
a(i): average distance between i and all of the other points in its own cluster
For a single point, i
$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$
$-1 \leq s(i) \leq 1$
Sprawling, overlapped clusters
Tight, well-separated clusters

- **Elbow Method** : Elbow method helps data scientists to select the optimal number of clusters for KNN clustering. It is one of the most popular methods to determine this optimal value of K**.**



<u>Conclusion</u>:

- Information about our dataset which comprises of 7787 rows and 12 columns, where columns like director, cast, country, date_added had some null values which were treated accordingly.

- Insights from our Exploratory Data Analysis:

  - 2015 was the year where the spike of growth began, 2019 and 2020 were the peak years where highest numbers of movies and TV shows were added on Netflix.

  - It is noticed that US, India and UK majorly create movies on this platform.

  - Jan suter is the most popular director on Netflix.

  - Anupam Kher and Shahrukh Khan are the most popular actors on Netflix.

- By applying Silhouette Score Method, we found the optimum value of K = 10.

- Using the given data set a simple recommender system was also created using the cosine similarity and recommendations for TV Shows and movies were obtained.

  - 68.1 % of the content available on Netflix are movies and 30.9 % of the content are TV shows.

  - Relative growth is observed here in the number of movies on Netflix than Tv shows.