NETFLIX MOVIES AND TV SHOWS CLUSTERING

Raja Chowdhury

Data science trainee,

AlmaBetter, Bangalore

Abstract

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

Introduction

Netflix is an American company that revolutioniz ed the watching of movies and television shows. Established in 1997 in California by Reed Hasting s (the current CEO) and Marc Randolph. Netflix offers film and television series library through distribution deals as well as its own productions, known as **Netflix Originals**. In 2007 Netflix started subscription based streaming service that allows its members to watch TV shows and movies on an internet connected device. Depend ing on our plan, we can also download TV shows and movies to our iOS, Android, or Windows 10 device and watch without an internet connection The movie and television library, and recommen dation system make accurate suggestion movie to the customers and the convenience of access and relatively low price brought Netflix dominate d position in the market. They have over 8000 movies or tv shows available on their platform, as of now, they have over 200M subscribers globally and in 2021 they made $5.1 billion of net profit.

Netflix users value this platform for a wide offer, high quality subtitles, and, yes, accurate, personalized recommendations. Netflix exploits an advanced machine learning based recommendation system that analyzes users' cho ices to suggest them new movies and TV shows. I n essence, it's a set of algorithms using machine l earning to analyze user data and movie ratings. T o make it more effective, Netflix has set up 1,300 recommendation clusters based on users viewin g preferences. As a result, whenever you turn on Netflix, you see a list of movies and TV shows tail ored to your interests and user profile. The goal i s to help each user find a movie or a TV show the y will enjoy and do so as quickly as possible (esti mations are they have just 90 seconds for that). The recommendation system itself remains in th e back end and is invisible to the user.

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.

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 increasingly focused on TV rather than movies in recent years?
4. Clustering similar content by matching text-based features.

Objective

NetflixRecommender recommends Netflix movies and TV shows 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 Description

Attribute Information:

The dataset provided contains 7787 rows and 12 columns.

The following are the columns in the dataset:

- **Show id:** Unique identifier of the record in the dataset
- **Type**: Whether it is a TV show or movie
- **Title:** Title of the show or movie
- **Director:** Director of the TV show or movie
- **Cast:** The cast of the movie or TV show
- **Country:** The list of the country in which a show/ movie is released or watched
- **Date added:** The date on which the content was onboarded on the Netflix platform
- **Release year:** Year of the release of the show/ movie
- **Rating:** The rating informs about the suitability of the content for a specific age group
- **Duration:** Duration is specified in terms of minutes for movies and in terms of the number of seasons in the case of TV shows
- **Listed in:** This columns species the category/ genre of the content
- **Description:** A short summary about the storyline of the content

Approach

As the problem statement says, understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that we used Affinity Propagation, Agglomerative Clustering, and K-means Clustering.

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.

Table1. *The above table shows the dataset in the form of Pandas Data Frame*

Natural Language Processing (NLP) Model:

For the NLP portion of this project, I will first convert all plot descriptions to word vectors so they can be processed by the NLP model. Then, the similarity between all word vectors will be calculated using cosine similarity (measures the angle between two vectors, resulting in a score between -1 and 1, corresponding to complete opposites or perfectly similar vectors). Finally, I will extract the 5 movies or TV shows with the most similar plot description to a given movie or TV show.

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

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 dose not contains any duplicate values.

4. 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:

I. Proportion of type of content
II. Country-wise count of content
III. Total release for last 10 years.
IV. Type and Rating-wise content count
V. Top 10 genres in movie content
VI. Length distribution of movies.
VII. Season-wise distribution of TV shows.
VIII. Count of content appropriate for different ages.
IX. Age-appropriate content count in top 10 countries with maximum content.
X. Proportion of movies and TV shows content appropriate for different ages.
XI. Season wise distribution of TV shows.
XII. Longest TV shows.
XIII. Top 10 topics on Netflix.
XIV. Extracting the features and creating the document term matrix.
XV. Topic modeling using LDA and LSA.
XVI. Most important features of topic.

5. Missing or Null value treatment:

In datasets, missing values arise due to numerous reasons such as errors, or handling errors in data.

We checked for null values present in our data and the dataset contains a null value.

In order to handle the null values, some columns and some of the null values are dropped.

6. 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.

7. Tfidf vectorization:

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit a machine learning algorithm for prediction.

We have also utilized the PCA because it can help us improve performance at a very low cost of model accuracy. Other benefits of PCA include reduction of noise in the data, feature selection (to a certain extent), and the ability to produce independent, uncorrelated features of the data.

So, it's essential to transform our text into tfidf vectorizer, then convert it into an array so that we can fit into our model.

- **Finding number of clusters**

  The goal is to separate groups with similar characteristics and assign them to clusters.

  We used the Elbow method and the Silhouette score to do so, and we have determined that 28 clusters should be an optimal number of clusters.

- **Fitting into model**

  In this task, we have implemented a K means clustering algorithm. K-means is a technique for

data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities (k).

8. Data Preprocessing:

Removing Punctuation: Punctuations does not carry any meaning in clustering, so removing punctuations helps to get rid of unhelpful parts of the data, or noise.

Removing stop-words: Stop-words are basically a set of commonly used words in any language, not just in English. If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

Stemming: Stemming is the process of removing a part of a word, or reducing a word to its stem or root. Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

9. 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.

for example:

- **Data analysis:** often a huge dataset contains several large clusters, analyzing which separately, you can come to interesting insights.
- **Anomaly detection:** as we saw before, data points located in the regions of low density can be considered as anomalies
- **Semi-supervised learning:** clustering approaches often helps you to automatically label partially labeled data for classification tasks.
- **Indirectly clustering tasks (tasks where clustering helps to gain good results):** recommender systems, search engines, etc.
- **Directly clustering tasks**: customer segmentation, image segmentation, etc.

Building a clustering model:

Clustering models allow you to categorize records into a certain number of clusters. This can help you identify natural groups in your data.

Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. In fact, you may not even know exactly how many groups to look for.

This is what distinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict.

These models are often referred to as **unsupervised learning** models, since there is no

external standard by which to judge the model's classification performance.

10. Topic Modeling:

• Latent Semantic Analysis (LSA)

LSA, which stands for Latent Semantic Analysis, is one of the foundational techniques used in topic modeling. The core idea is to take a matrix of documents and terms and try to decompose it into separate two matrices –

• A document-topic matrix

• A topic-term matrix.

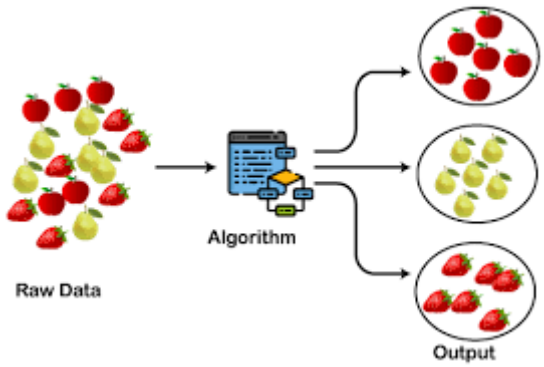
Therefore, the learning of LSA for latent topics includes matrix decomposition on the document-term matrix using Singular value decomposition. It is typically used as a dimension reduction or noise reducing technique.

• Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

11. Clusters Model Implementation

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



1. Silhouette's Coefficient-

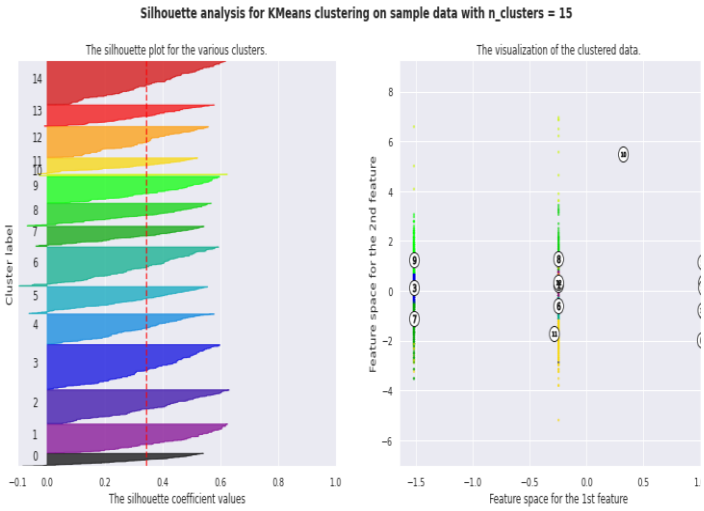
If the ground truth labels are not known, the evaluation must be performed utilizing the model itself. The Silhouette Coefficient is an example of such an evaluation, where a more increased Silhouette Coefficient score correlates to a model with better-defined clusters. The Silhouette Coefficient is 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.

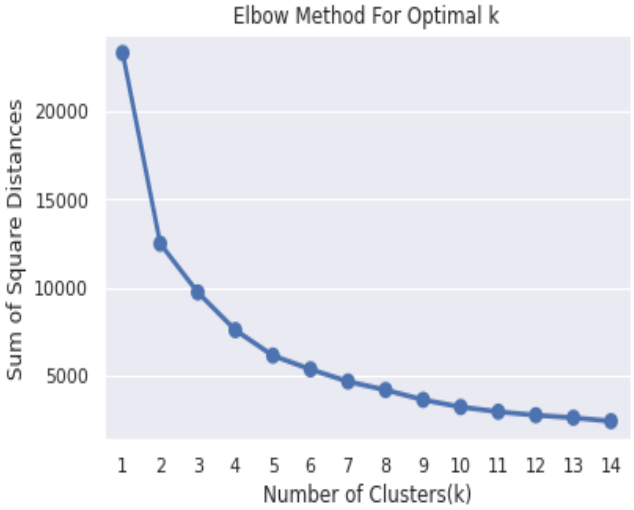
For n_clusters = 2, silhouette score is 0.42843328899854627
For n_clusters = 3, silhouette score is 0.3832991558393045
For n_clusters = 4, silhouette score is 0.37431547662296216
For n_clusters = 5, silhouette score is 0.372084391833681 6
For n_clusters = 6, silhouette score is 0.3681694950084983
For n_clusters = 7, silhouette score is 0.3760261702686321 6
For n_clusters = 8, silhouette score is 0.3707733418425295
For n_clusters = 9, silhouette score is 0.374851392272243
For n_clusters = 10, silhouette score is 0.36498814933847673
For n_clusters = 11, silhouette score is 0.3550141388753564
For n_clusters = 12, silhouette score is 0.3560344374158886 4
For n_clusters = 13, silhouette score is 0.3497366048331119
For n_clusters = 14, silhouette score is 0.3437866138686444
For n_clusters = 15, silhouette score is 0.3427424627239503



2. Elbow Curve:

The Elbow Curve is one of the most popular methods to determine this optimal value of k. The Elbow Method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.



3. DBSCAN:

DBSCAN stands for density based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.

There are two key parameters of DBSCAN:

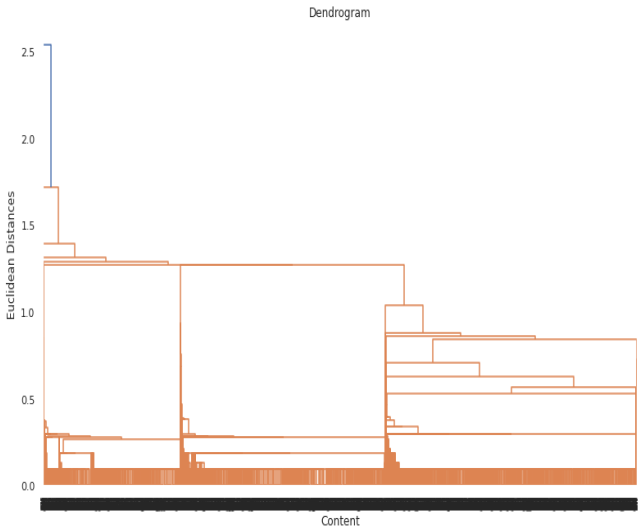
eps: The distance that specifies the neighborhoods. Two points are considered to be neighbors if the distance between them are less than or equal to eps.

minPts: Minimum number of data points to define a cluster.

4. Dendrogram

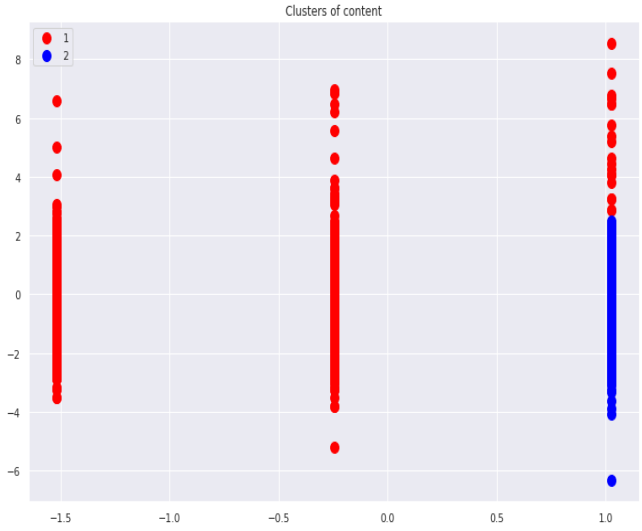
A dendrogram is a diagram that shows the hierar chical relationship between objects. It is most co mmonly created as an output from hierarchical cl ustering. The main use of a dendrogram is to wor k out the best way to allocate objects to clusters.

The dendrogram below shows the hierarchical cl ustering of six observations shown on the scatter plot to the left.



5. 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.



12. K-means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

K-means algorithm works:

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non overlapping subgroups where each data point belongs to only one group. 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.

Conclusion

Tailored recommendations can be made based on information about movies and TV shows. In addition, similar models can be developed to provide valuable recommendations to consumers in other domains.

1) Director and cast contain a large number of null values so we will drop these 2 columns.
2) In this dataset there are two types of con tents where 30.86% includes TV shows a nd the remaining 69.14% carries Movies.
3) We have reached a conclusion from our analysis from the content added over years that Netflix is focusing movies and TV shows (Fom 2016 data we get to know that Movies is increased by 80% and TV shows is increased by 73% compare)
4) From the dataset insights we can conclud e that the most number of TV Shows rele ased in 2017 and for Movies it is 2020
5) On Netflix USA has the largest number of contents. And most of the countries pref erred to produce movies more than TV s hows.
6) Most of the movies are belonging to 3 ca tegories
7) TOP 3 content categories are Internation al movies, dramas, comedies.
8) In text analysis (NLP) I used stop words, r emoved punctuations, stemming & TF-IDF vectorizer and other functions of NLP
9) Applied different clustering models like K means, hierarchical, Agglomerative clust ering, DBSCAN on data we got the best cl uster arrangements.
10) By applying different clustering algorithm s to our dataset,we get the optimal num ber of cluster is equal to 3.