Author : Zeeshan Ahmed

Course Title: Netflix Movies and TV Shows Clustering

Keywords: Machine Learning,Clustering, Natural Language Processing, Clustering

Abstract:

Netflix is an online platform consisting of streaming services for entertainment purposes. It is a subscription based service with a wide range of content. Most of its content is divided among two types namely Movies and TVShows. Now in recent years it's the most popular OTT platform for people all around the world. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not losing their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

1. Problem Statement

In this project we'll be working with Netflix data to interpret the latest trends and gain insights on the content listed, 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 was about time that a recommended system was created. To deliver it, we are going to analyze the data and Cluster similar content by matching text-based features by building a recommendation system.

In our Data we've following columns:

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 Release Year of the movie / show
rating : TV Rating of the movie / show
duration : Total Duration - in minutes or number of seasons
listed_in : Genre
description: The Summary description of the movie

2. Steps Involved

1) Exploratory Data Analysis :-

EDA helps us to understand data in a much easier way. It gives us some valuable information that is helpful in model building as well as useful outside-model-work. In our data we plot bar plots, pie chart in order to gain insights that might be useful to understand data/people

2) Null Value Treatment :-

It is said that data with more than 40% null values should be dropped as they only drag down the model if used. If treated then won't have much effect either and so getting any column at the beginning and to clear it out is very helpful. In our case when checked for null values we didn't have any. Checked for null values and there are null values in director, cast, country, release year, rating columns. Treated the null values in the column country by filling it by mode,treated the null values in the cast column by replacing the null values with 'No Cast'.

3) Adding new features :-

1. Created a new feature Audience_AgeGroup which has three values namely 'Millennials', 'GenZ', 'Kids' which has values with respect to ratings of content on Netflix.

2. Added a Month column which is extracted from data_added column.

3. Converted the duration values of TV Shows which had values in seasons into minutes and added these values into a new column.

3. Conclusions from EDA :

1. Majority of the content on Netflix is movies
2. There weren't any TVShows until 2013 on Netflix
3. The rate of TVShow content has been greater than that of movies
4. and now in the year 2020, there are equal number of movies and TVShows on Netflix
5. Most of the content on Netflix is from United States
6. There are a wide range of movies with respect to ratings on Netflix. The highest number of content with ratings are TV-MA and TV-14 from Movie and TV Show.
7. Netflix has the highest content count for individuals of the age group Millennials and lowest content for Kids.
8. Highest Number of movies and TV shows were produced in the years 2015-2019
9. December was the month were the most amount of content was added on Netflix followed by October

   Features selection

4. Features selection :-

● Here we are going to do Clustering similar content by matching text-based features so column description is one of the important feature. ● Convert the text to lowercase.

● Tokenize the text.

● Removed all the stop words and punctuation.

● WE use 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 algorithm for prediction.

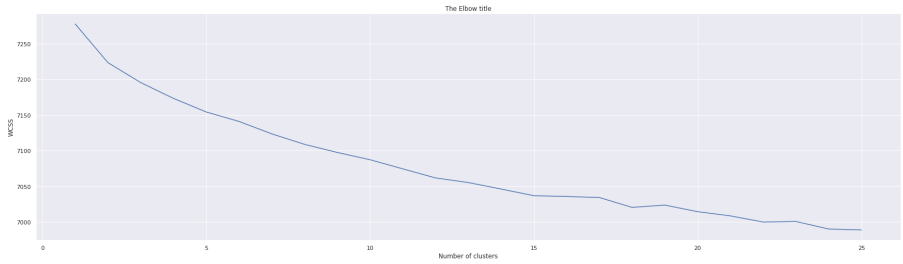
5.ML algorithms(unsupervised)

1. K-mean
2.agglomerative clustering

1. K-Means:

● K-Means Clustering is an Unsupervised Learning algorithm which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

Elbow Method:



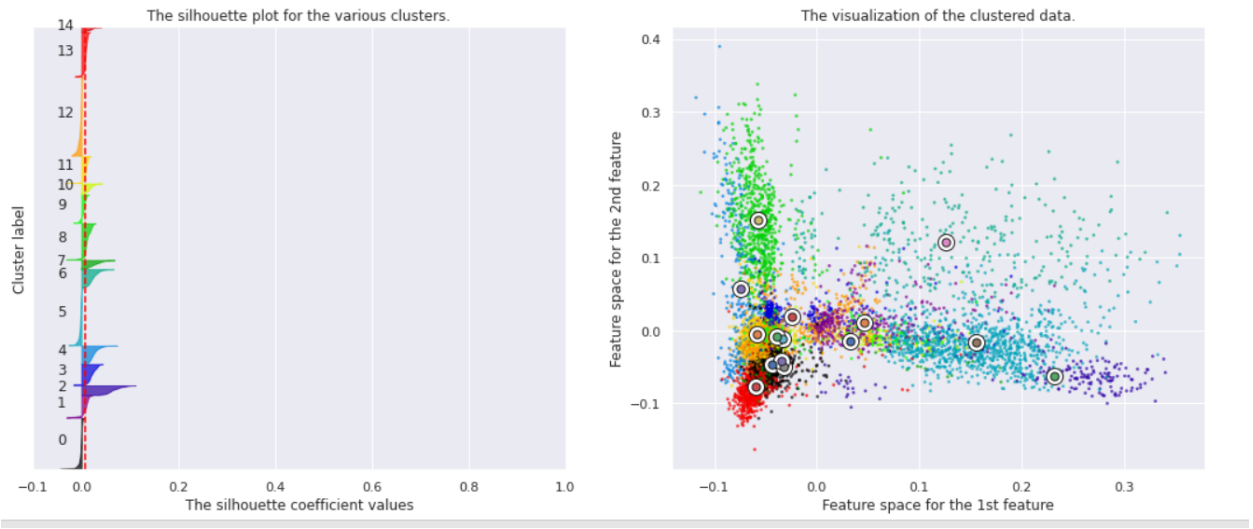
From the above graph, the Elbow method generates 18 clusters.

Evaluation

1.Silhouette Score : is a metric to evaluate the performance of a clustering algorithm. It uses compactness of individual clusters(intra cluster distance) and separation amongst clusters (inter cluster distance) to measure an overall representative score of how well our clustering algorithm has performed

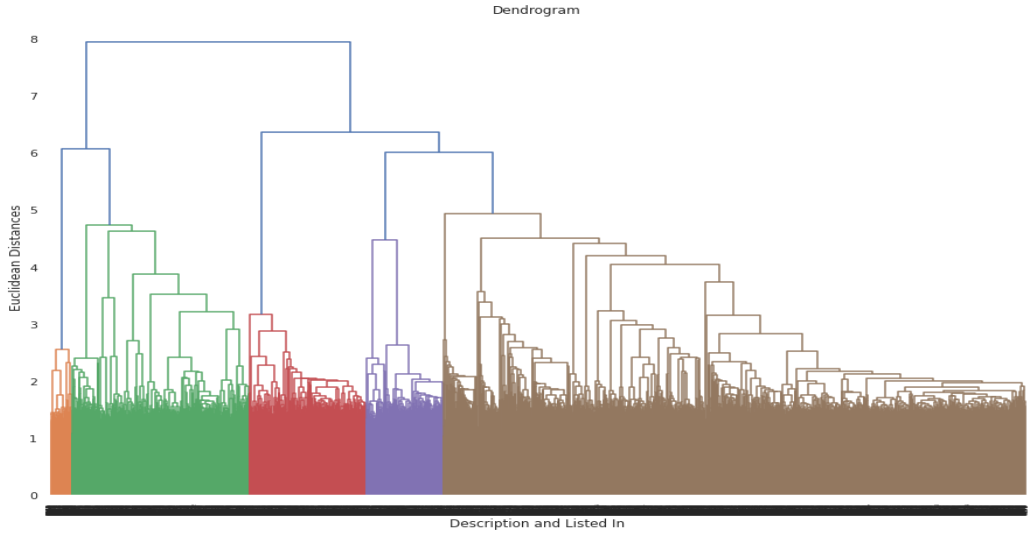
● Silhouette scores would always lie between -1 to 1. 1 representing better clustering

● Silhouette score is 0.0077226308384750935 so model is performing well



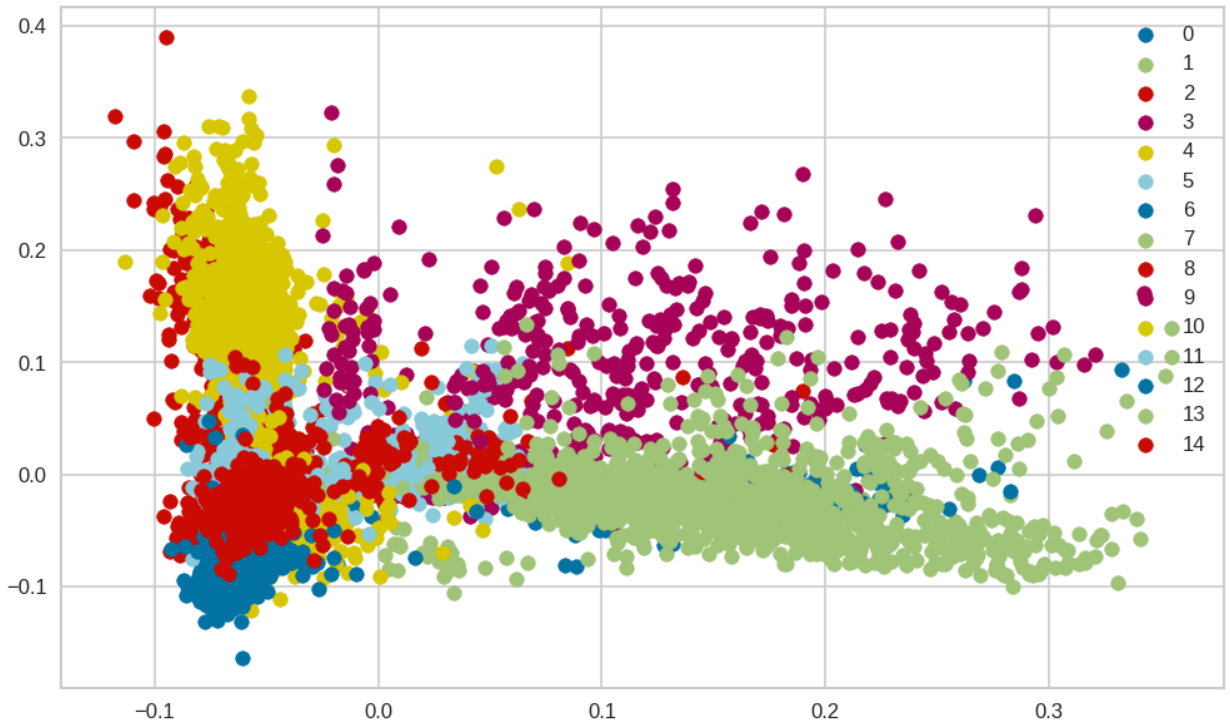
2.Agglomerative Clustering :-

- In agglomerative clustering no need to give the value of k beforehand
- The agglomerative hierarchical clustering algorithm is a popular example of HCA
- Here I used ward linkage
- The optimal number of clusters is 13 using the Dendrogram

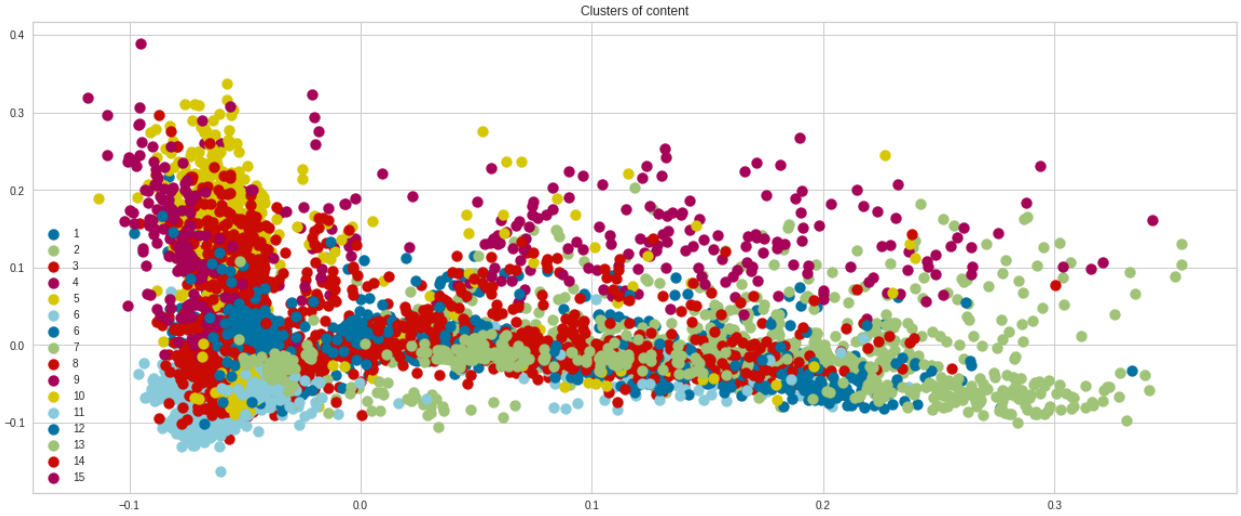


Visualization of Clusters:

K-means Clustering with 15 Clusters:



AgglomerativeClustering with 15 Clusters:



Recommendation System:

Built a recommendation system using cosine similarity taking the clusters data points as input and the output would be the data points from the same cluster. The optimal numbers of clusters was determined to be 15 after the average of the k values from elbow, dendogram algorithms.

1. Getting the recommendation for the movie ‘American Psycho’ on Netflix;

	Recommendations
0	Shine On with Reese
1	Love Is Blind
2	Death Note
3	My Scientology Movie
4	How to Make an American Quilt
5	The Good Cop
6	Zodiac
7	Rain Man
8	A Family Man
9	Lucas Brothers: On Drugs

2. Getting the recommendation for TV Show ‘The Stranger Things’ on Netflix:

	Recommendations
0	Beyond Stranger Things
1	Prank Encounters
2	The Umbrella Academy
3	Reckoning
4	Sleepless Society: Nyctophobia
5	Anjaan: Special Crimes Unit
6	The OA
7	Kiss Me First
8	The 4400
9	Eli

Conclusion:

● The use of a combination of topic models to process text data has aided in clustering movies and TV shows on Netflix.

● The best performing models, K-Means and Hierarchical Clustering, grouped data into 15 clusters.

● In addition to helping build recommendation engines, this labeled content can be studied and explored to determine the type of content on demand, potentially providing intuition to content creators about the content Netflix would be interested in signing

Thank you