



UnSupervised ML(Clustering) - Netflix Movies and TV Shows Clustering

Technical documentation

Mohammed Akifuddin Kashif
akifkashif007@gmail.com



Introduction

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 lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

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.



Overview of the data

Attribute Information:

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




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.

Steps involved

1. Extracting the head and tail to get an idea of the dataset
2. Extracting the description of the dataset to get the mean, min, max values and data types of the columns
3. Extracting the info of the data to show the non-null count of the column values
4. Getting number of unique values for each column
5. Extracting the shape of the i.e, number of rows and columns
6. Checked for null values and there are null values in director, cast, country, release year, rating columns.


- 
7. 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'
 8. Plotting relevant graphs to extract information from them.

Plots Used:

- BarPlot
- CountPlot
- PieChart
- ViolinPlot
- Squarify(TreeMap)
- StripPlot
- Histplot

Feature Engineering:

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.

Drawing conclusions from EDA:

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



Text Processing

The steps involved in text preprocessing are :

A) Tokenization:

Involves breaking of natural language text into chunks of information that can be considered as discrete elements. The token occurrences in a document can be used directly as a vector representing that document.

B) Punctuation Removal:

All the punctuations from the text are removed.

C) Stopword Removal:

Common words that add very little or no significant insight to the text being processed are removed beforehand. This reduces time and computational complexity.

D) Stemming Words:

Stemming is the process of reducing inflected words to their word stem, base or root form—generally a written word form. This reduces different forms of the same word carrying the same base meaning. It should be noted that stemming does not remove synonyms.



Feature Selection

- Relevant non-text attributes describing the content's maturity ratings, duration, year of release and type of content are taken.
- The attributes exempted are 'show id', 'title' and 'added date' as they add little to no substance in the qualitative and quantitative characterization of the video itself.
- To feed in information about the video content's plot, cast and genres, we will be using the preprocessed and topic modeled version of text attributes 'Description', 'Listed in' and 'Cast'.


Performance Metrics

Silhouette Score: is used to measure the separation distance between clusters. It displays a measure of how close each point in a cluster is to points in the neighboring clusters.

Methods to Choose Optimal Cluster Numbers

Elbow Method: The elbow method plots the value of the cost function produced by different values of clusters, k , in K-means clustering.

The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.



Dendrogram Method: Dendrograms are a diagrammatic representation of the hierarchical relationship between the data points. These are used to observe the output of hierarchical agglomerative clustering. The number of clusters is determined by slicing the dendrogram horizontally. All the resulting child branches formed below the horizontal cut represent an individual cluster at the highest level in the system.

Models Used

- K-means
- Hierarchical Clustering

K-means:

It is an iterative algorithm that divides the unlabeled dataset into K unique clusters where each dataset belongs to only one group having similar properties.

This algorithm aims to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabeled dataset as input, divides the dataset into K number of clusters, and repeats the process until it can't find better clusters.

For K-Means clustering the elbow and optimal silhouette score were found at 8 clusters with a silhouette score of 0.4686, Davies-Bouldin Index of 0.887 and Calinski-Harbaz Score of 2901.84.



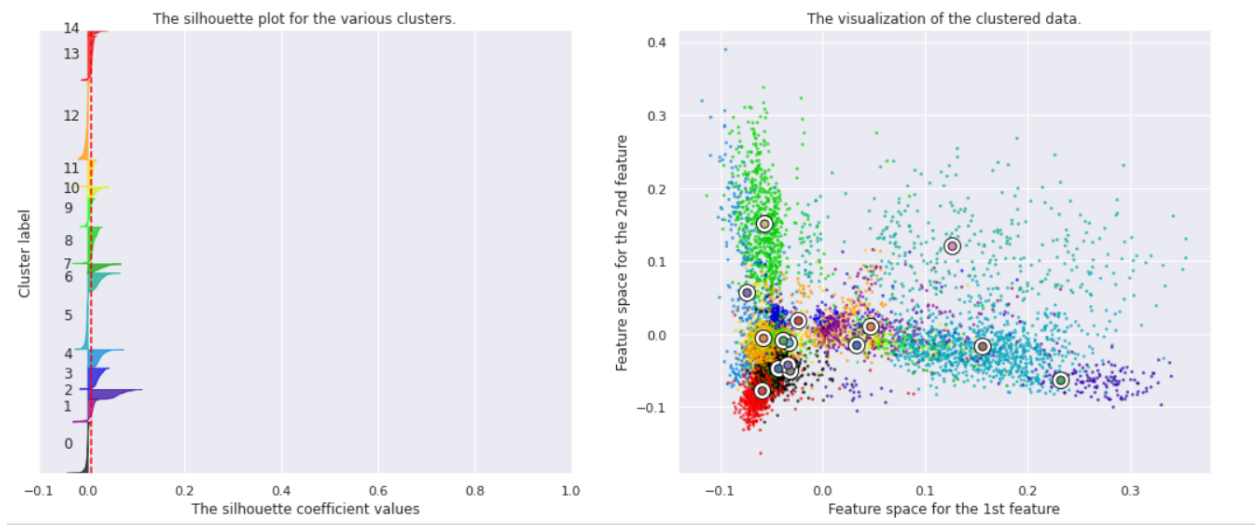
Hierarchical Clustering:

A Hierarchical clustering method works via grouping data into a tree of clusters. The aim is to produce a hierarchical series of nested clusters.

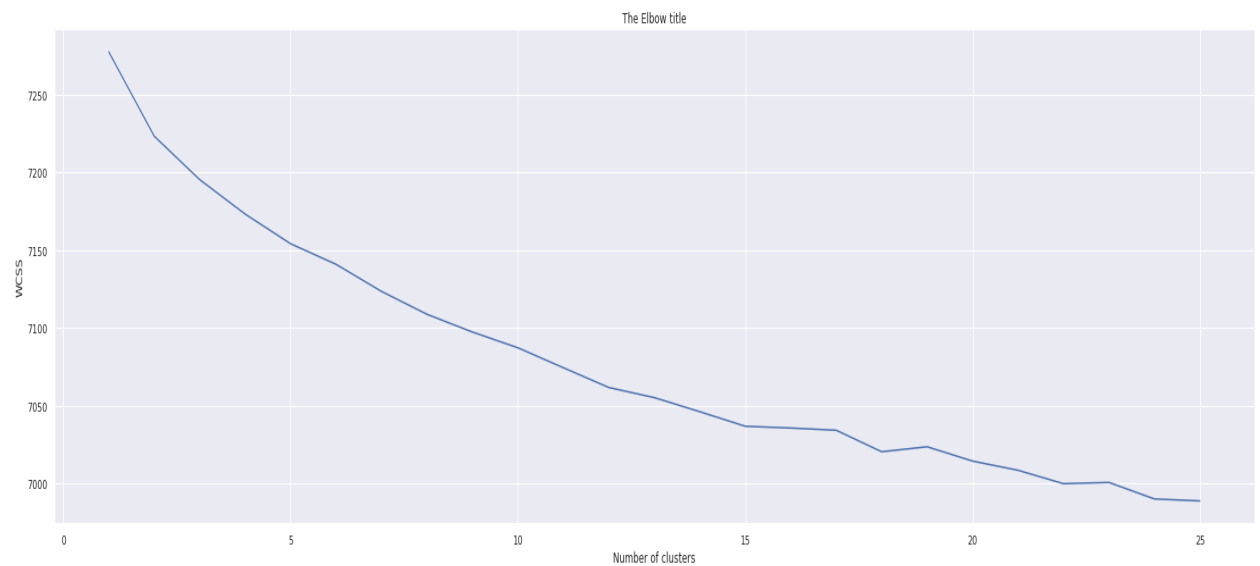
The agglomerative algorithm starts by considering every data point as an individual cluster and calculates the similarity of one cluster with all the other clusters.

The highly similar or close clusters are merged and the proximity matrix for each cluster is recalculated. These steps are repeated until only a single cluster is made.

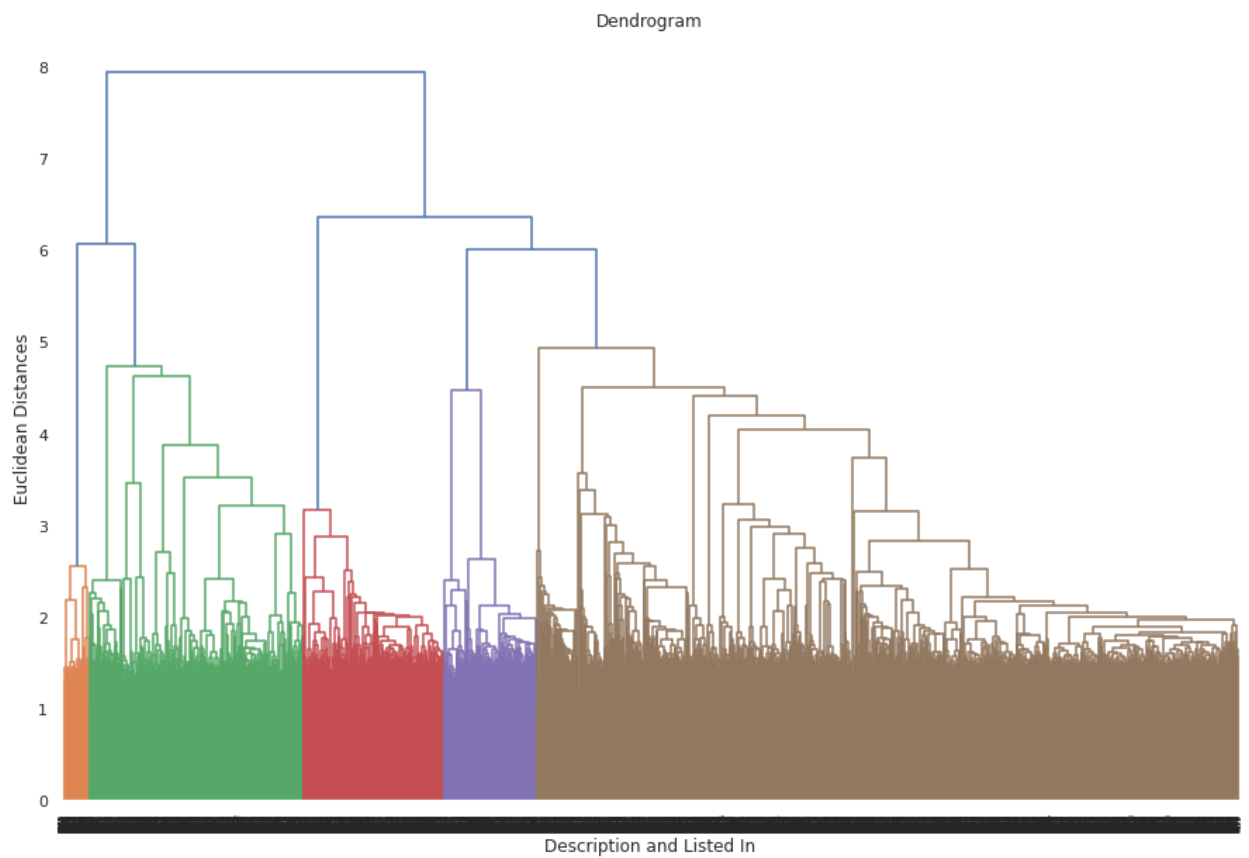
Silhouette Score:



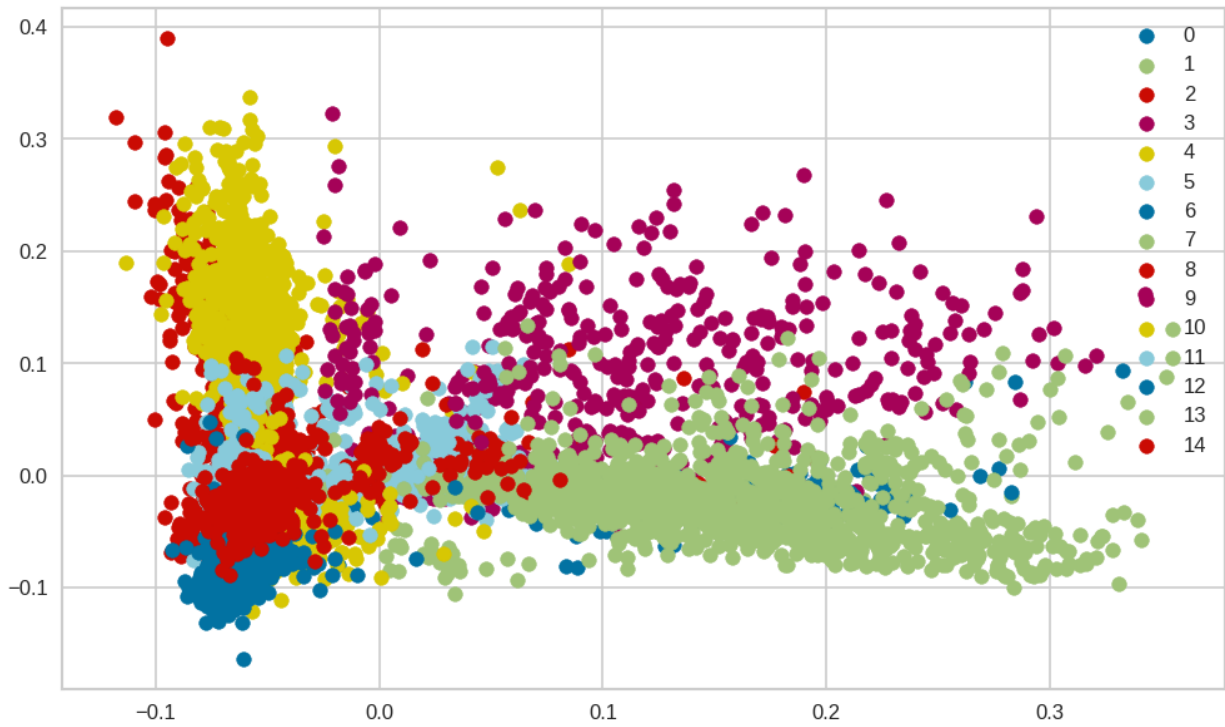
Elbow Visualization:

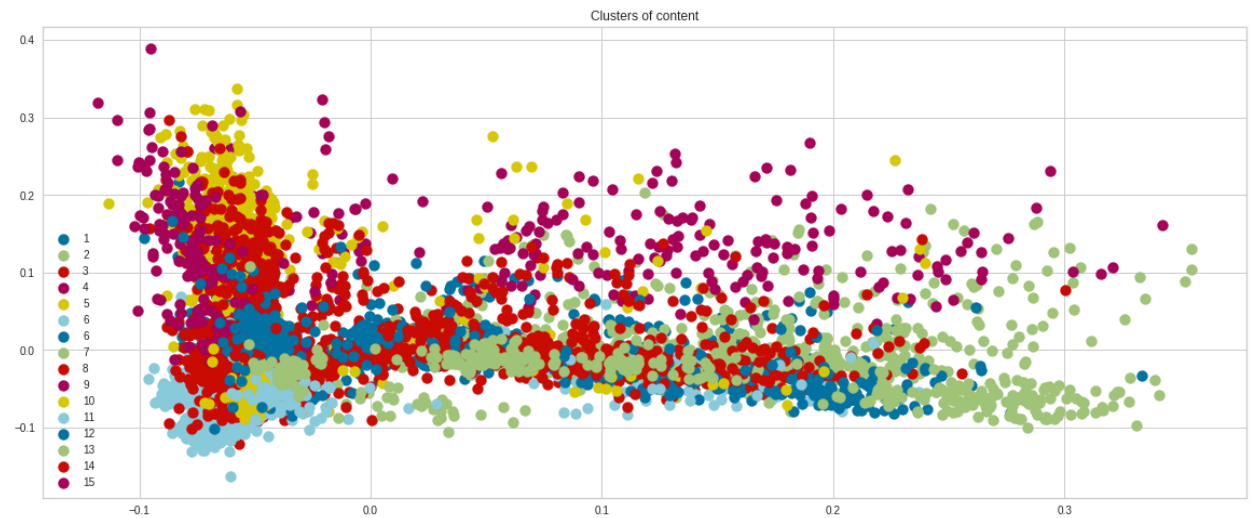


Dendrogram:



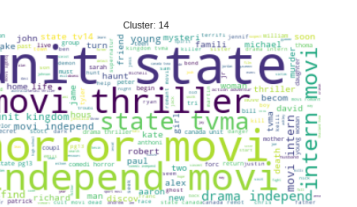
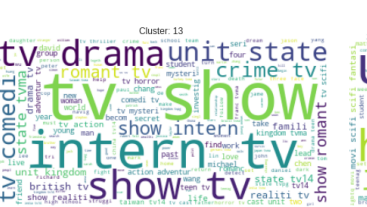
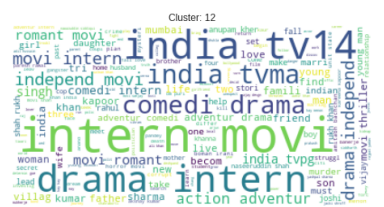
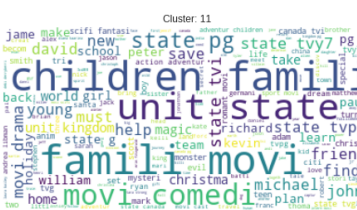
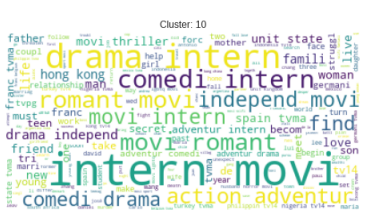
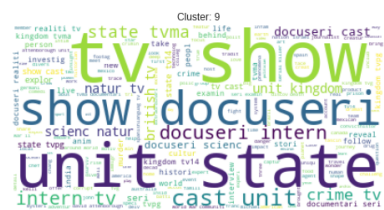
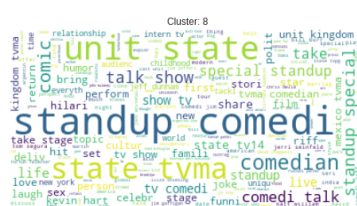
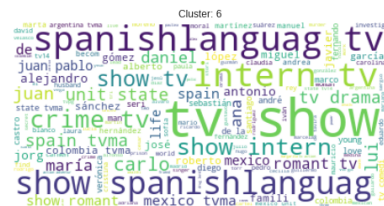
K-means Clustering with 15 Clusters:





Word Cloud for 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, dendrogram algorithms.

1. Getting the recommendation for 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.