Capstone Project

On

Netflix Movies and TV Shows Clustering

by

Premanand Gaikwad

AlmaBetter

CONTENT

1. Introduction
2. Problem Statement
3. Data Summary
4. Exploratory Data Analysis (EDA)
5. Data Cleaning and Text Pre-processing
6. Topic Modelling/Clustering
7. Models Performance Metrics
8. Conclusion

Introduction

- ✓ Netflix, Inc. is an American subscription streaming service and production company based in Los Gatos, California. Founded on August 29, 1997, Netflix had 220.7 million subscribers worldwide

- ✓ Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device. You can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection.

Problem Statement

- ✓ Netflix content varies by region and may change over time. You can watch from a wide variety of award-winning Netflix Originals, TV shows, movies, documentaries, and more. The more you watch, the better Netflix gets at recommending TV shows and movies we think you'll enjoy.

- ✓ we know that, Users of the such a platform have experienced it ourselves that endless scrolling of selecting the show what to watch and users spend more time deciding what to watch than watching their movie.

- ✓ The goal of this project is to cluster the movies or TV shows on the Netflix and build a movie recommendation system for users. Initially we are going to analyze a available data from the Netflix to get the insights to know statistics of the data by different feature wise.

Data Summary

- ✓ 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 the given datasets, 7787 records and 12 features are available. 11 features contain text data and 1 feature contain numerical data.

- ✓ Attribute Information

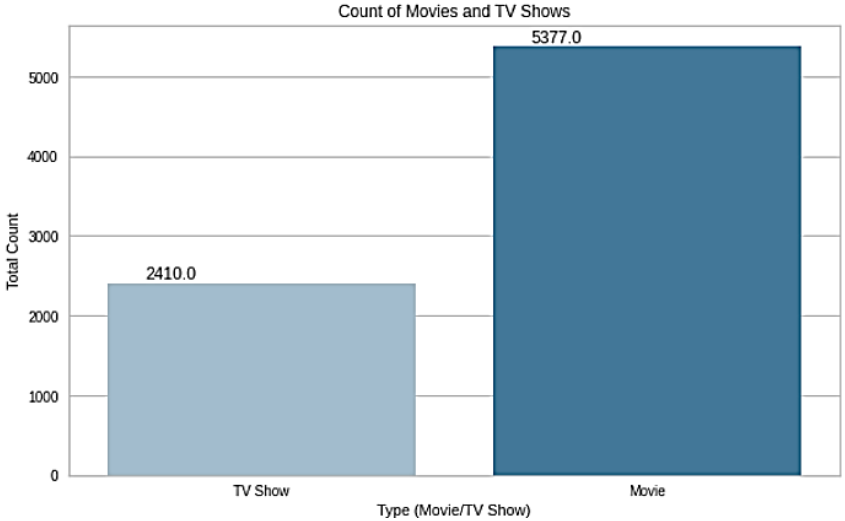
1. show_id : Unique ID for every movie / TV show
2. type : Identifier - a movie or TV show
3. title : Title of the movie / TV show
4. director : Director of the movie
5. cast : Actors involved in the movie / show

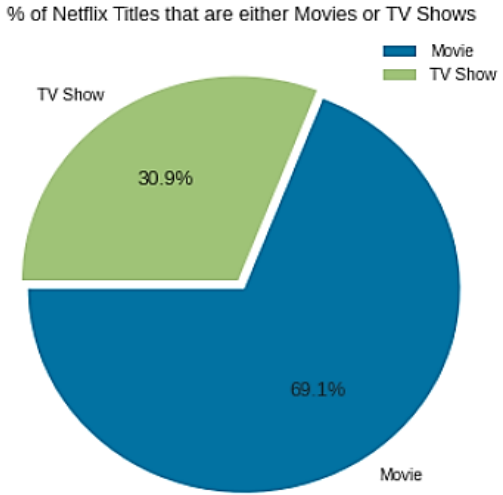
6. country : Country where the movie / show was produced
7. date_added : Date it was added on Netflix
8. release_year : Actual release year of the movie / show
9. rating : TV Rating of the movie / show
10. duration : Total duration - in minutes or number of seasons
11. listed_in : Genre
12. description: The summary description

Exploratory Data Analysis (EDA)

Netflix Film Types :

Comparison between the total number of movies and shows in this dataset just to get an idea of which one is the majority.

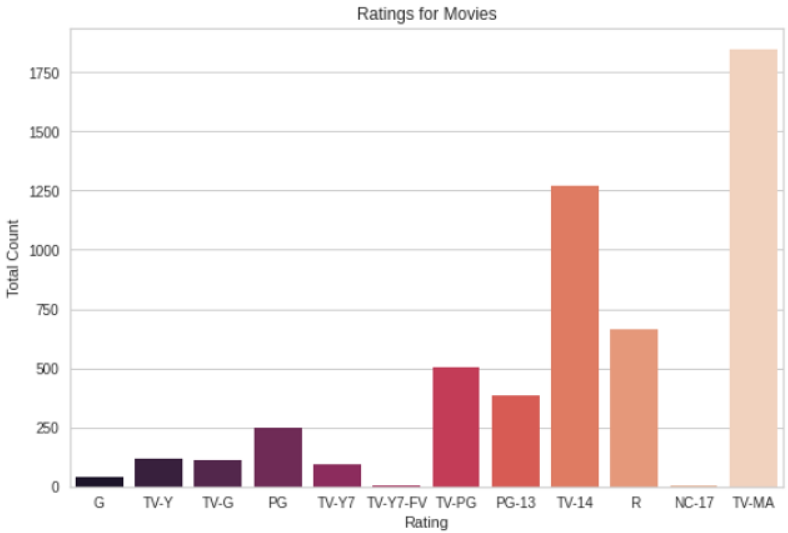


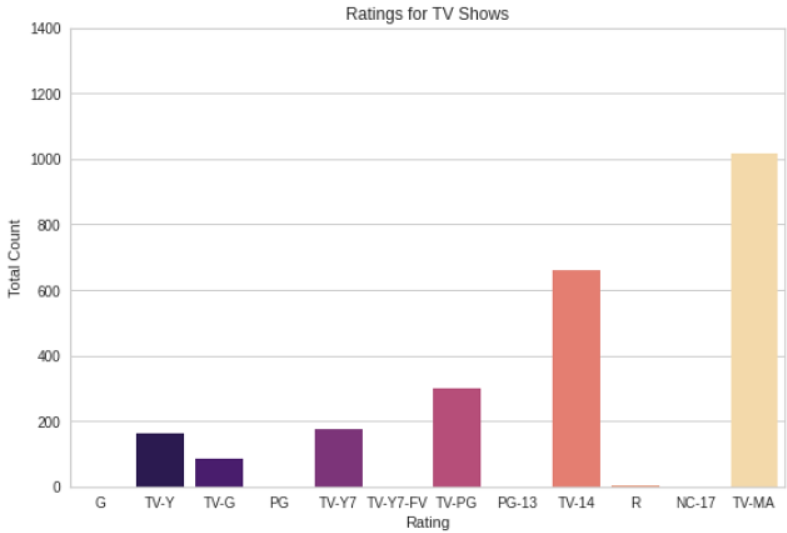


So there are roughly 5377 movies and almost 2410 shows, with movies being the majority. This makes sense since shows are always an ongoing thing and have episodes. However, in terms of titles, there are far more movie titles (69.1 %) than TV show titles (30.9 %).

Netflix Film Ratings :

Now, we will explore the ratings which are based on the film rating system.

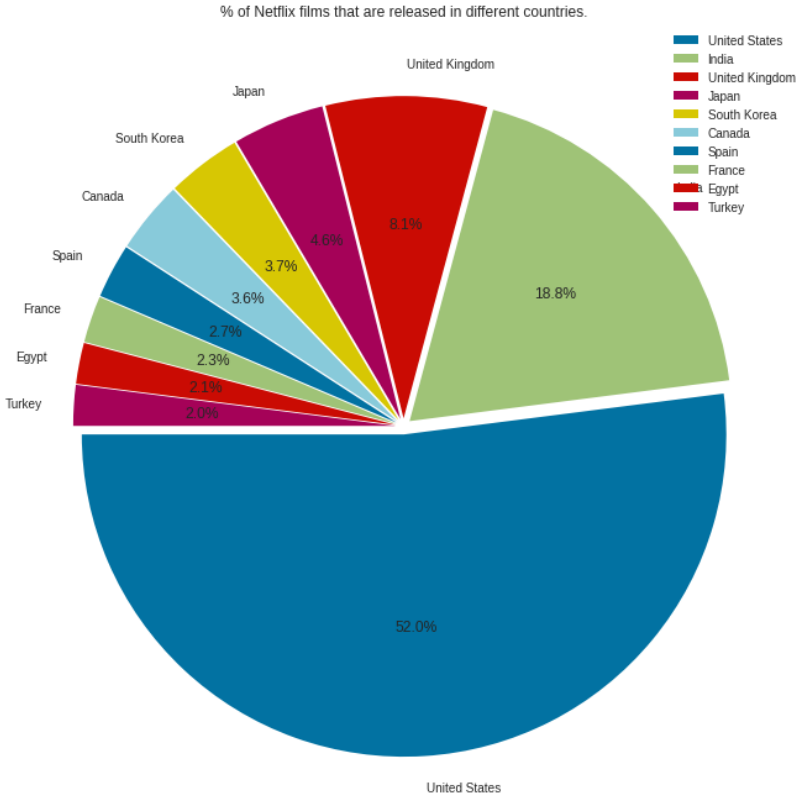




There is much more content for a more mature audience(MA). For the mature audience, there is much more movie content than the TV shows. Also, for the younger audience (under the age of 17), it is the similar, there are more movies than TV shows.

Netflix Films Country :

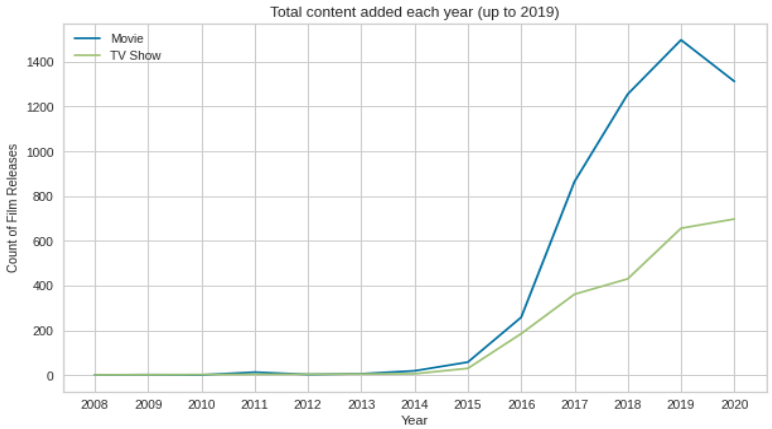
Now we will check the density of the films releases in different country.



- ✓ Overall, there is much more content that comes from the United States (52 %) and India (18 %).
- ✓ In the US, Hollywood spends a lot of money on its movies. The US has historically been the largest market for films, so American studios have amassed a lot of wealth and resources. That's why most of the movies come from the US.

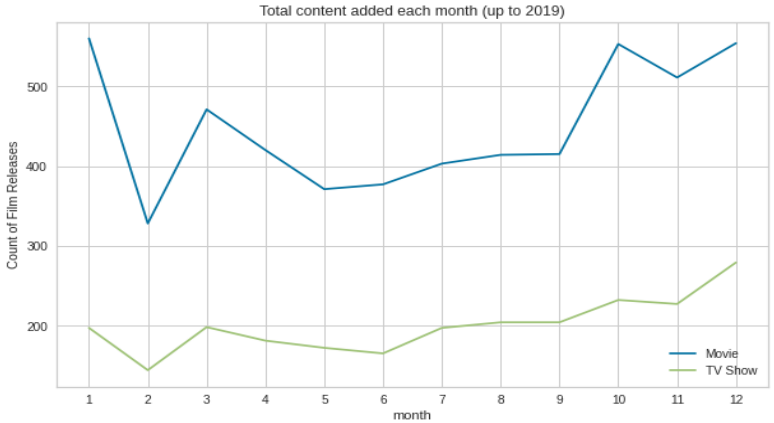
Content added on Netflix :

Now we will take a look at the amount content, Netflix has added throughout the previous years.



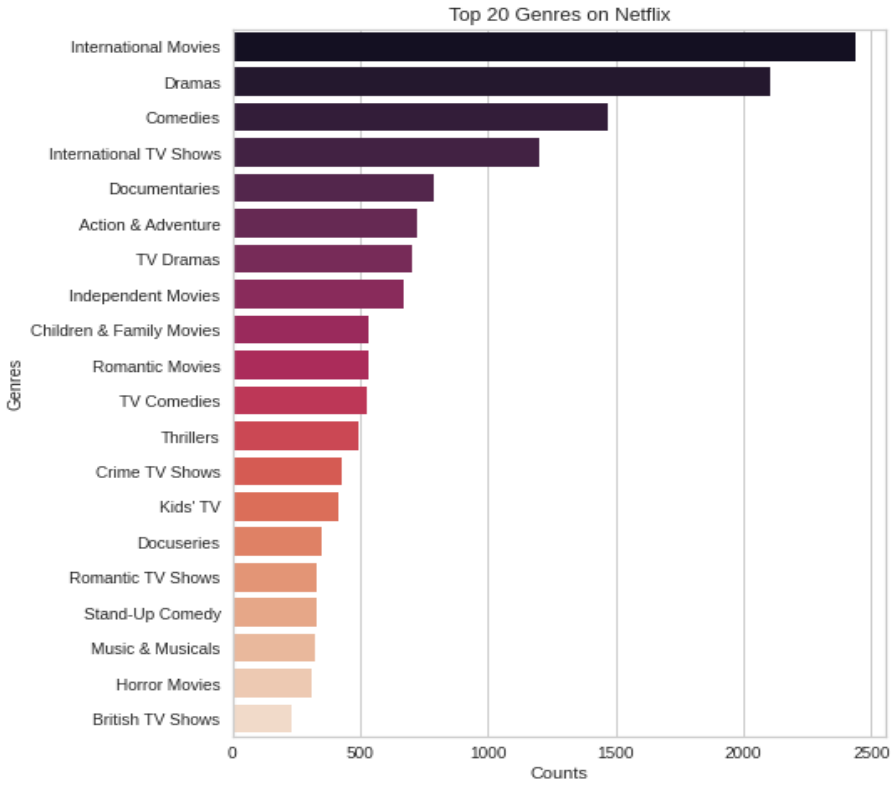
Consistent growth in the number of movies on Netflix compared to TV shows. Based on the above timeline, Netflix started gaining traction after 2014. Since then, the amount of content added has been tremendous.

By observing above plot it is seen that most of movies are release in holidays month i.e. January, October, November, December.



Popular Genres :

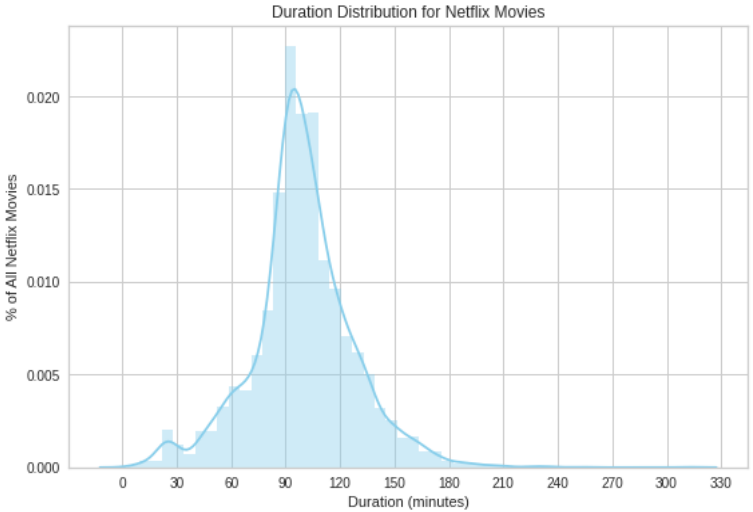
Now we will check the type of movie or TV shows audience prefer.

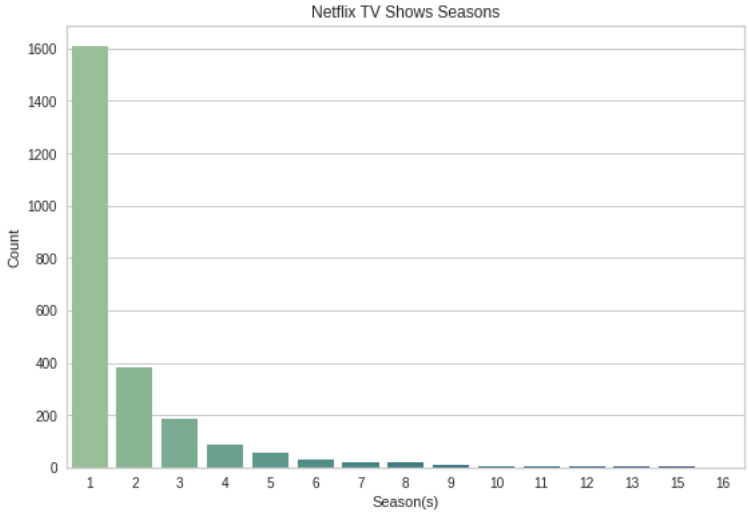


International movies take a peak and are followed by dramas and comedies. As we can see from the above pie chart, the United States has the most content available. It looks like Netflix has decided to release a tone of international movies. The reason for this could be that most Netflix subscribers aren't actually in the United States, but rather the majority of viewers are international subscribers.

Netflix Film Duration :

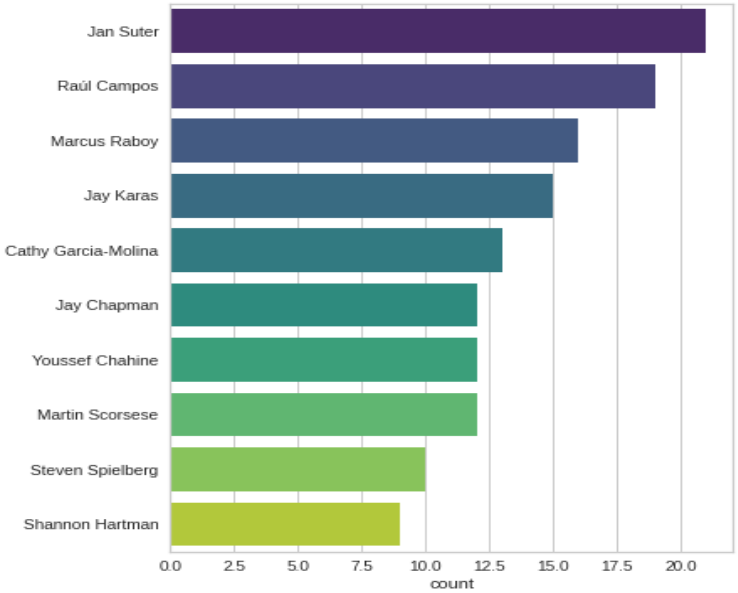
In this we will observe the distribution of the duration of the movies on normal plot and distribution of the seasons for TV shows on bar plot.





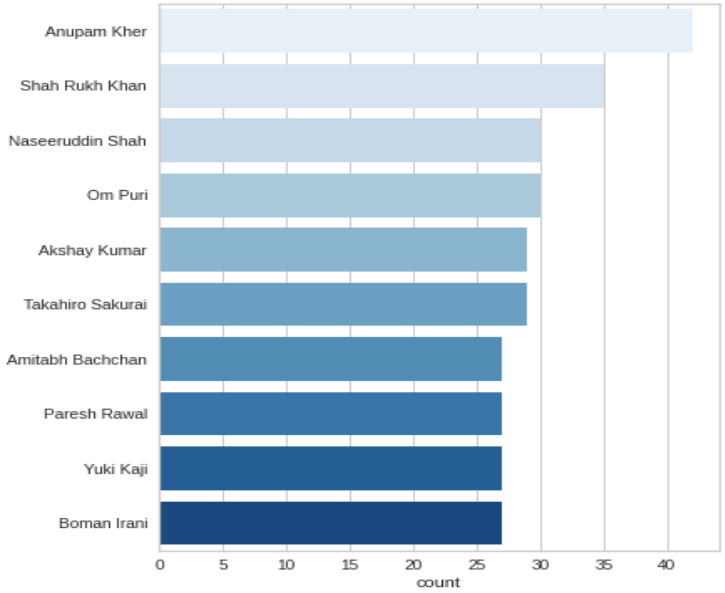
We know that movies are measured in time and shows are measured by seasons, so we have split the duration feature into two parts. Above, on the left side, we can see that most of the movies are about 90 minutes long, which seems to make sense. On the right hand side, distribution is skewed toward the right because most TV shows have one season.

Top 10 Director on the Netflix



'Jan Suter' directed the most of the movies, as stated previously regarding the top genres, it's no surprise that the most popular directors on Netflix with the most titles are mainly international as well.

Top 10 actors on Netflix



'Anupam Kher' casted in the most of the films. as above most popular actors on Netflix based on the number of titles are all international as well.

Data Cleaning and Text Pre-processing

Text pre-processing of the text data is an essential step as it makes the raw text ready for mining and making it suitable for a machine learning model. The objective of this step is to clean noise those are less relevant to cluster:

- **Stop words** are those words in natural language that have very little meaning, such as “is”, “an”, “the”, etc. To remove stop words from a sentence, divide your text into words and then remove the word if it exists in the list of stop words provided by NLTK.

- **Lemmatization** is the process of converting a word to its base form. lemmatization considers the context and converts the word to its meaningful base form. lemmatization is that it is more accurate than stemming.

- **Label Encoding** refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated.

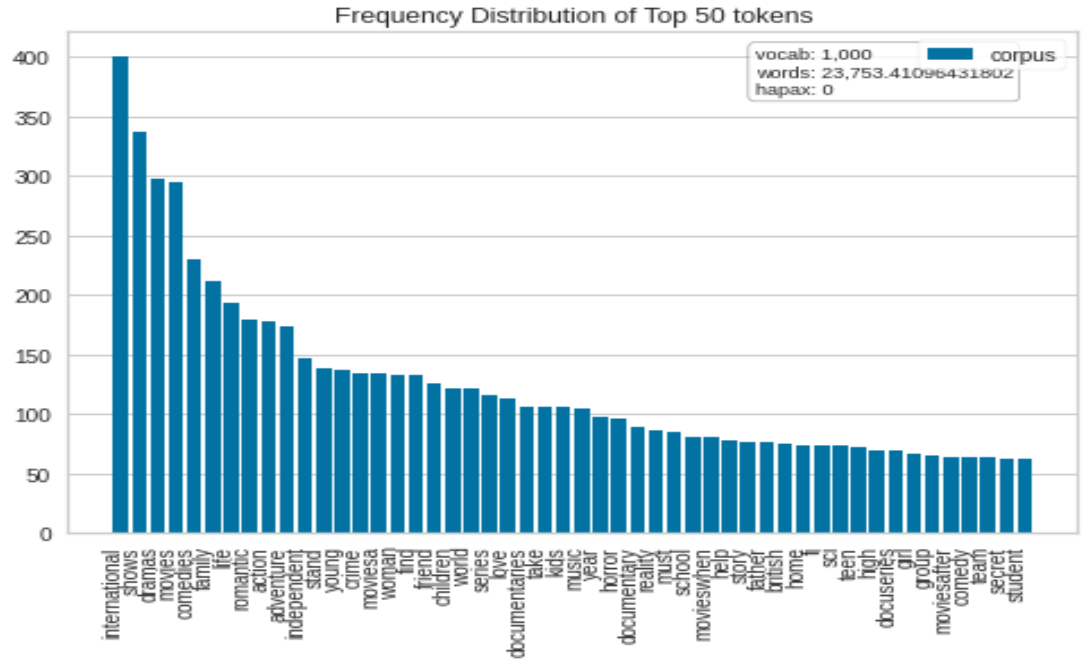
- **StandardScaler** removes the mean and scales each feature/variable to unit variance. useful for the features that follow a Normal distribution.

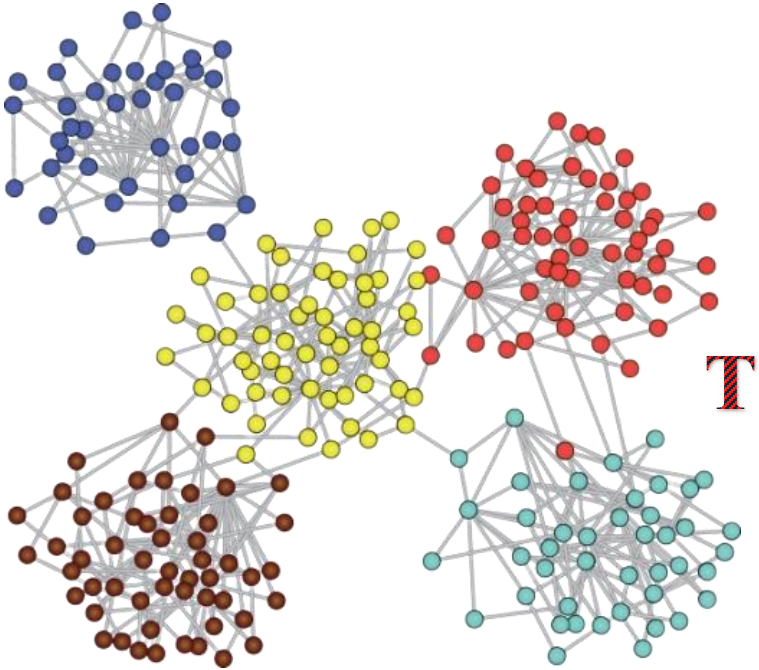
Vectorization TFIDF, short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. Vectorization is a step in feature extraction, the idea is to get some distinct features out of the text for the model to train on, by converting text to numerical vectors

	fish	rabbit	raccoon	turtle
0	0.889003	0.457901	0.000000	0.000000
1	0.920187	0.236982	0.000000	0.311602
2	0.701512	0.000000	0.712658	0.000000

Frequency of the Words :

A frequency distribution tells us the frequency of each vocabulary item in the text. It tells us how the total number of word tokens in the text are distributed across the vocabulary item

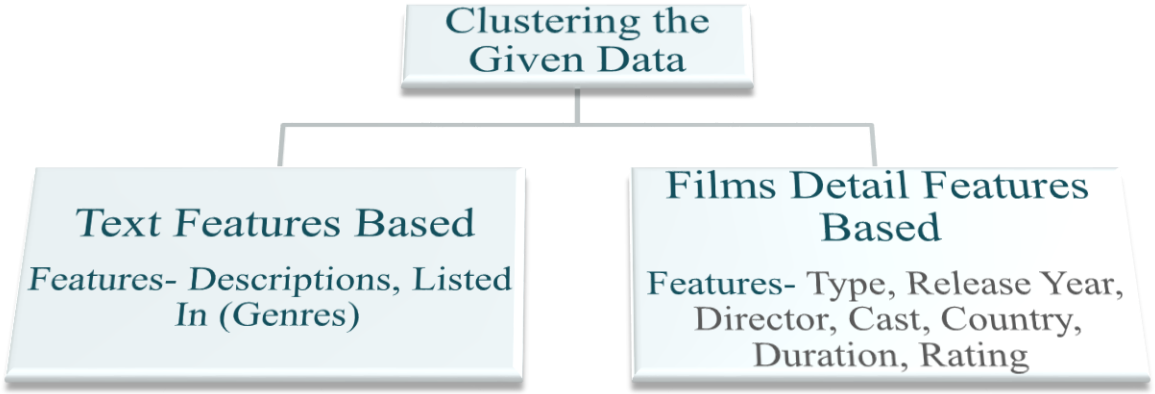




Topic Modelling / Clustering

Clustering the Movies and TV Shows

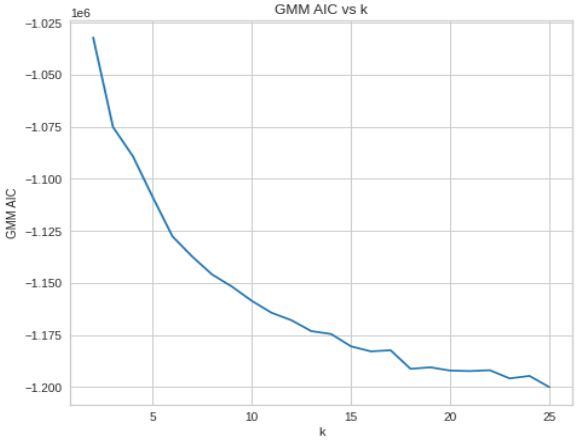
We are going to cluster the films data based on the description, genres of the film data and films other detail feature based (type, release year, director, cast, country, duration, rating) because most of the audience choose the film based on the description and type (genres) of the movie or either cast of the film or availability of the time. So we have grouped them into two parts to cluster the data for a better recommendation.

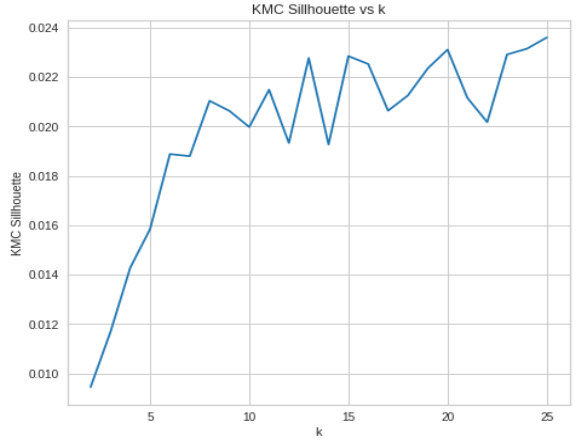


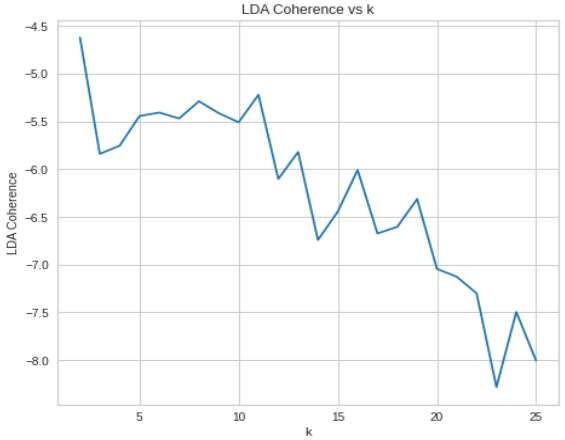
Clustering Models :

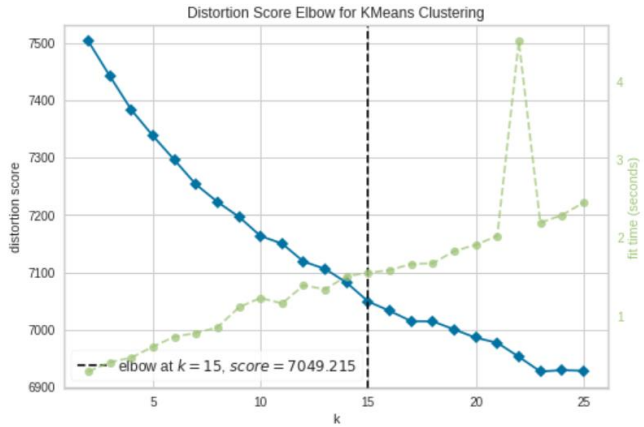
1. Gaussian Mixture Model (GMMs)
2. K-Means Clustering Model (KMC)
3. Elbow Method
4. Latent Dirichlet Allocation Model (LDA)
5. Affinity Propagation Model (APM)
6. Agglomerative Clustering

Model Performance Metric







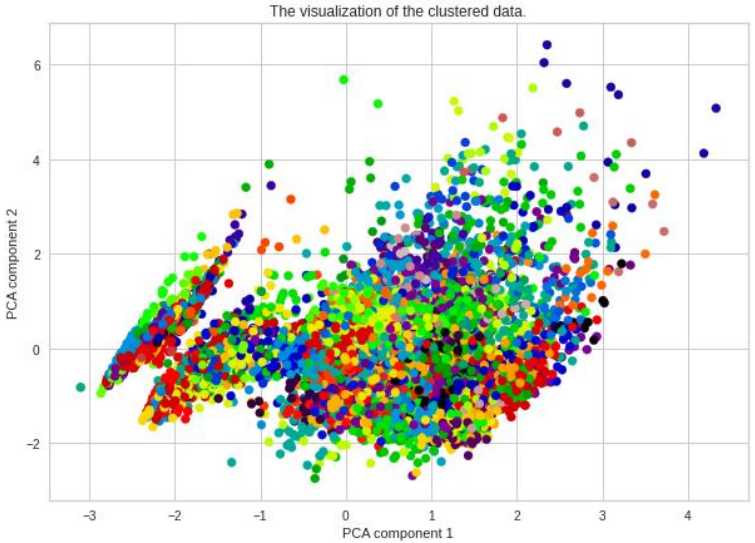


Significance of the score

- For AIC, the lower the score is better.
- For silhouette, the higher the score is better.
- For coherence, the higher the score is better.

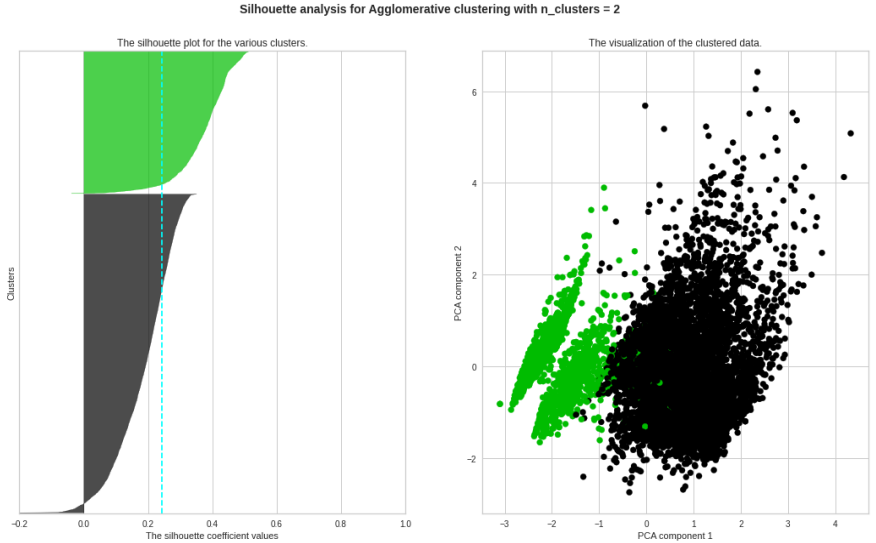
It seems that the K-mean cluster's silhouette score gives the highest score at 25 topics or cluster and elbow methods give 15 topics at best. Gaussian mixture models-AIC seems to hint at the 25 topics. If we rely on LDA-coherence, k=12 is the highest. We will simulate the data from 25 latent/hidden topics by considering all scores.

Affinity Propagation Model (APM)



Silhouette score is o.328 which is not that well and estimated number of clusters are 226 which is not possible to account.

Agglomerative Clustering



Silhouette score is 0.2441 it is not good and number of clusters are 2 which is also not well clustered.

It seems that both the above model not doing well as silhouette score is very less.

Dash App for Recommendation

Netflix Movies and TV-Show Recommender

Select Movie/TV Show:

Avengers: Infinity War

Recommended Movies

Show Name	Cast	Release Year	Duration/Seasons	Genre	Description
9	Elijah Wood, John C. Reilly	2009	80 min	Action & Adventure, Independent Movies, Sci-Fi & Fantasy	In a postapocalyptic world, rag-doll robots hide in fear from dangerous machines out to exterminate them, until a brave newcomer joins the group.
Goli Soda 2	Samuthirakani, Bharath Seeni	2018	128 min	Action & Adventure, Dramas, International Movies	A taxi driver, a gangster and an athlete struggle to better their lives despite obstacles like crooked politicians, evil dons and caste barriers.
Æon Flux	Charlize Theron, Marton Csokas	2005	93 min	Action & Adventure, Sci-Fi & Fantasy	Aiming to hasten an uprising, the leader of an underground rebellion dispatches acrobatic assassin Aeon Flux to eliminate the government's top leader.
10,000 B.C.	Steven Strait, Camilla Belle	2008	109 min	Action & Adventure	Fierce mammoth hunter D'Leh sets out on an impossible journey to rescue the woman he loves from a vicious warlord and save the people of his village.
14 Blades	Donnie Yen, Zhao Wei	2010	113 min	Action & Adventure, International Movies	In the age of the Ming Dynasty, Quinglong is the best of the Jinyiwei, an elite assassin squad made up of highly trained former street urchins. When evil [illegible]h Jia unseats the emperor, Quinglong is called to action but is quickly bet
16 Blocks	Bruce Willis, Mos Def	2006	102 min	Action & Adventure	Tasked with escorting a prosecution witness to court, an aging cop gears up for the

Conclusion

- We can see from the data that the company took certain approaches in their marketing strategy to break into new markets around the world. As of June 30, 2022, Netflix had 220.7 million subscribers worldwide, including 73.3 million in the United States and almost 147 million internationally. A large part of its success was due to the decision to expand to international markets, and we can see that a good number of international movies and TV shows were added over the years as part of Netflix's global expansion.
- It seems that the K-mean cluster's silhouette score gives the highest score For 25 topics or cluster and elbow methods give 15 topics at best.
- Gaussian mixture models-AIC score seems to hint at the 25 topics.
- If we rely on the LDA-coherence score, k=12 is the highest.
- As we cluster the categorical data using the Affinity Propagation Model and Agglomerative Model they didn't perform well and gave the worse clustering result.
- So that, We will simulate the data from 25 latent/hidden topics by considering all scores.


NETFLIX
Thank You