Capstone Project – 4

Netflix Movies and TV Shows Clustering

by:-

Akshit Singh

Agenda

- ➔ Look at the problem statement
- ➔ Study the dataset
- ➔ Looking for null and duplicate values
- ➔ E.D.A and visualization
- ➔ Feature engineering
- ➔ Clustering algorithms
- ➔ Conclusion

Approach

- **Understanding Business problem**
- **Exploratory Data Analysis**
  - Understanding features
  - EDA conclusion
- **Text Pre-processing**
  - Punctuation and stopwords removal
  - Stemming
- **Vectorization**
- **PCA for dimensionality reduction**
- **Modelling**
  - Fitting model to the dataset
- **Model Performance & Evaluation**
- **Conclusion**

Problem Statement

The dataset consists of TV shows and movies available on Netflix. In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. It will be interesting to explore what all other insights can be obtained from the dataset. The given challenge is to cluster similar content by matching text-based features.

Data Summary

- There are 7,787 observations with various types of field in our dataset.
- List of columns:-

1. show_id : unique id for every movie/tv show
2. type : Identifier(movie or show)
3. title : Title of the movie/show
4. director : Director of the movie/show
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
9. rating : TV rating of the movie/show
10. duration : Total duration(minutes or number of seasons)
11. listed_in : Genre
12. description: The summary description

Null and duplicate values

- Null values present in director, cast, country, date_added and rating columns.

- No duplicate rows are found in the dataset.

```
# checking if any null values are present in our dataset
count_of_null_values = netflix_df.isnull().sum()
count_of_null_values
```

```
show_id            0
type               0
title              0
director        2389
cast             718
country          507
date_added        10
release_year       0
rating             7
duration           0
listed_in          0
description        0
dtype: int64
```

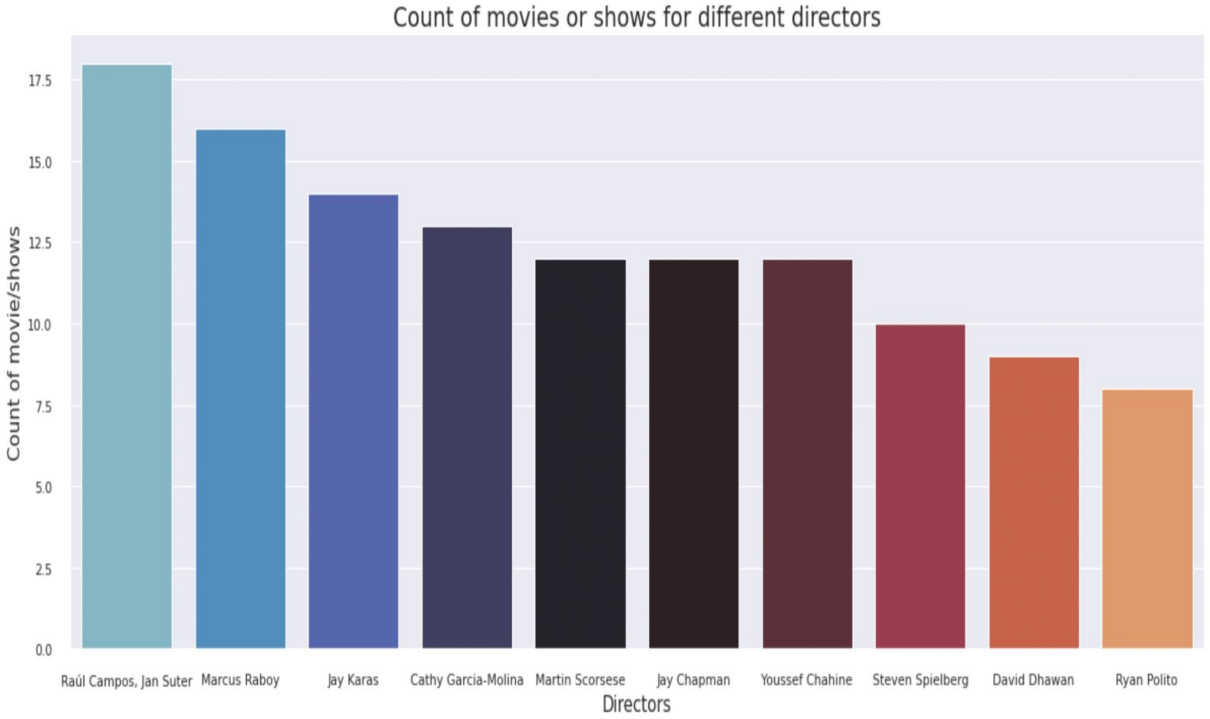
```
# checking duplicates in our dataset
value = len(netflix_df[netflix_df.duplicated()])
print("Total no. of duplicates = ", value)
```

```
Total no. of duplicates =  0
```

Exploratory Data Analysis

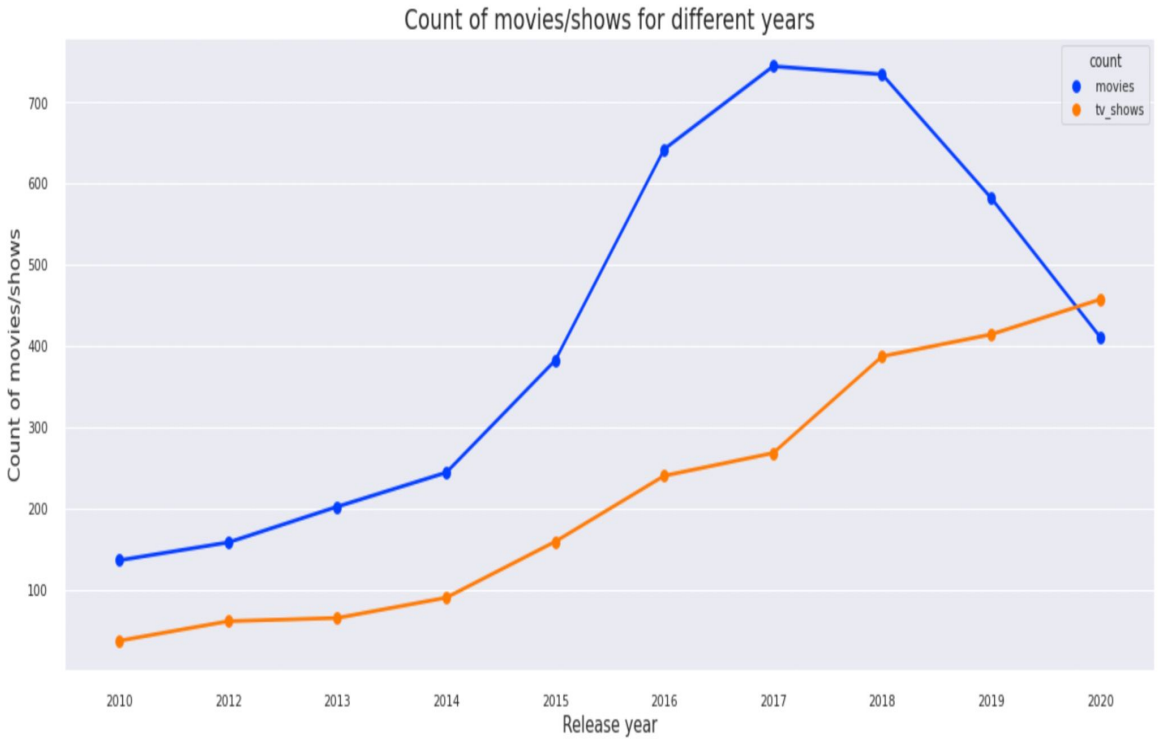
Top 10 directors with the most number of movies

- Raúl Campos, Jan Suter have directed the most number of movies followed by Marcus Raboy.
- Ryan Polito directed the least number of movies.
- Jay Chapman, Martin Scorsese and Youssef Chahine have directed the same amount of movies.



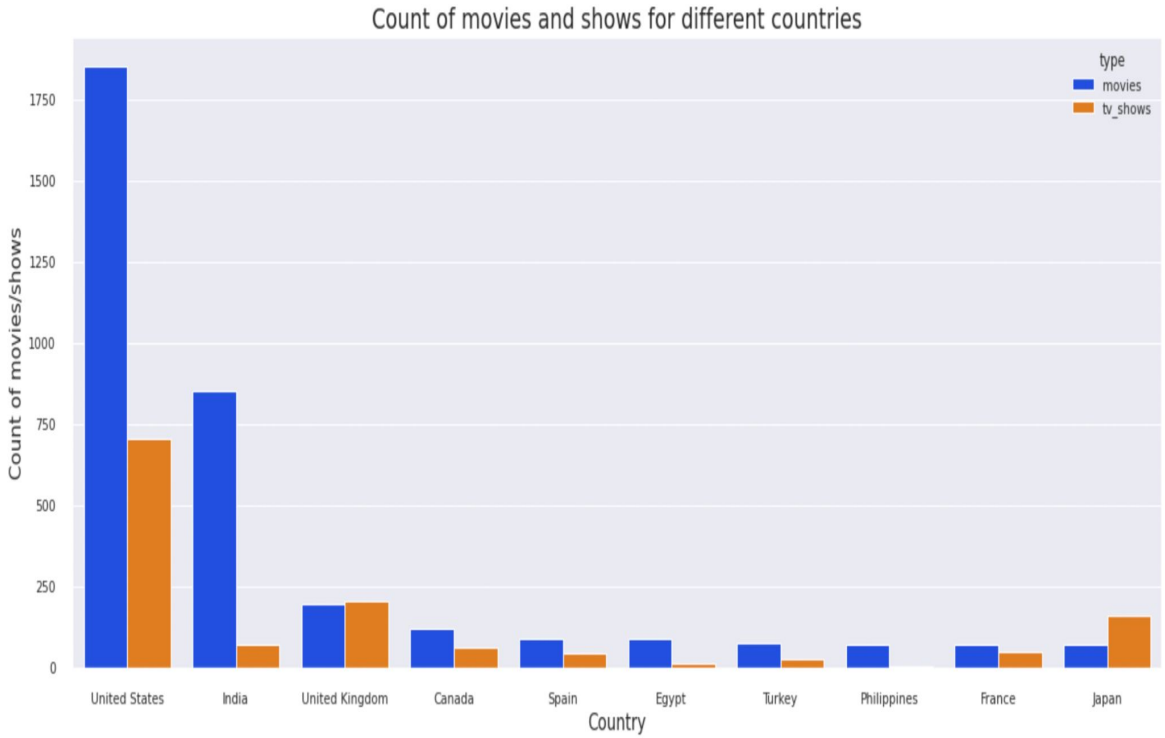
Is Netflix focusing on TV shows rather than movies in recent years?

- The highest number of movies and tv shows were released in 2017 and 2020 respectively.
- With each year number of tv shows keeps on increasing.
- In 2020 we see a dip in movies and an increase in tv shows due to Covid.

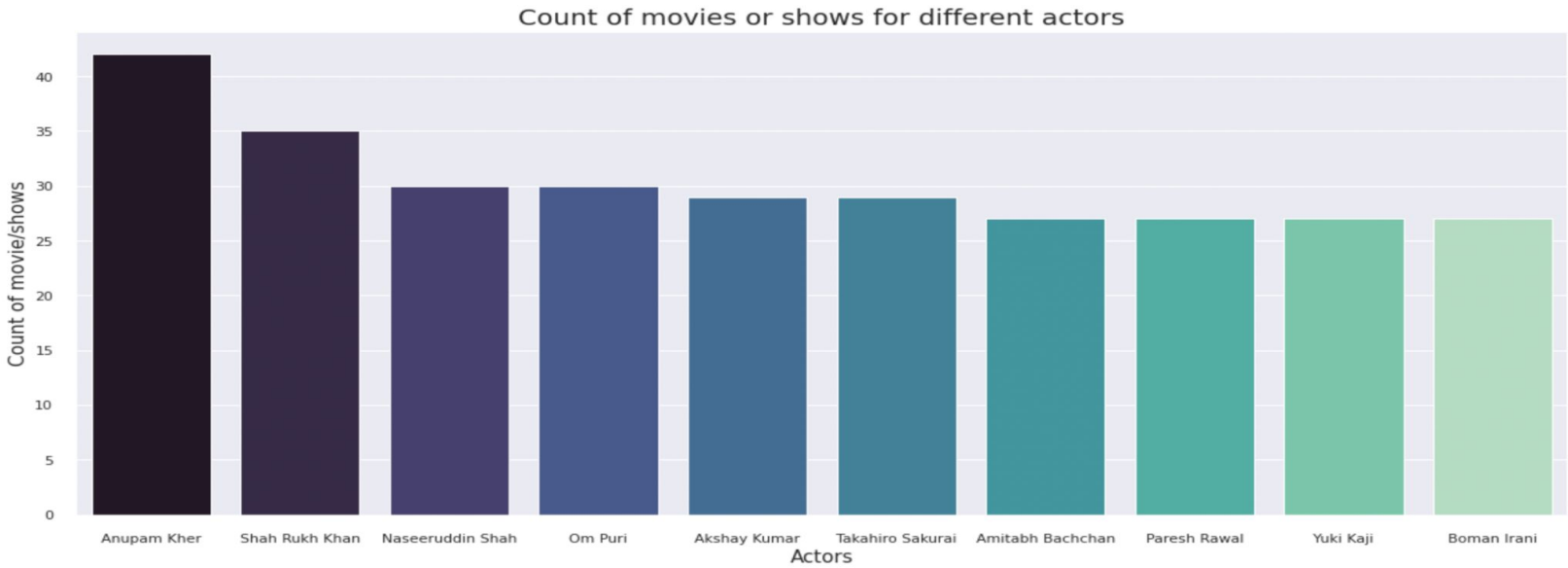


Understanding what type is available in different countries

- United States has produced the highest number of movies and shows.
- India is second in terms of movies produced.
- For the United Kingdom count of movies and tv shows are same.
- Japan has more TV shows than movies on Netflix.



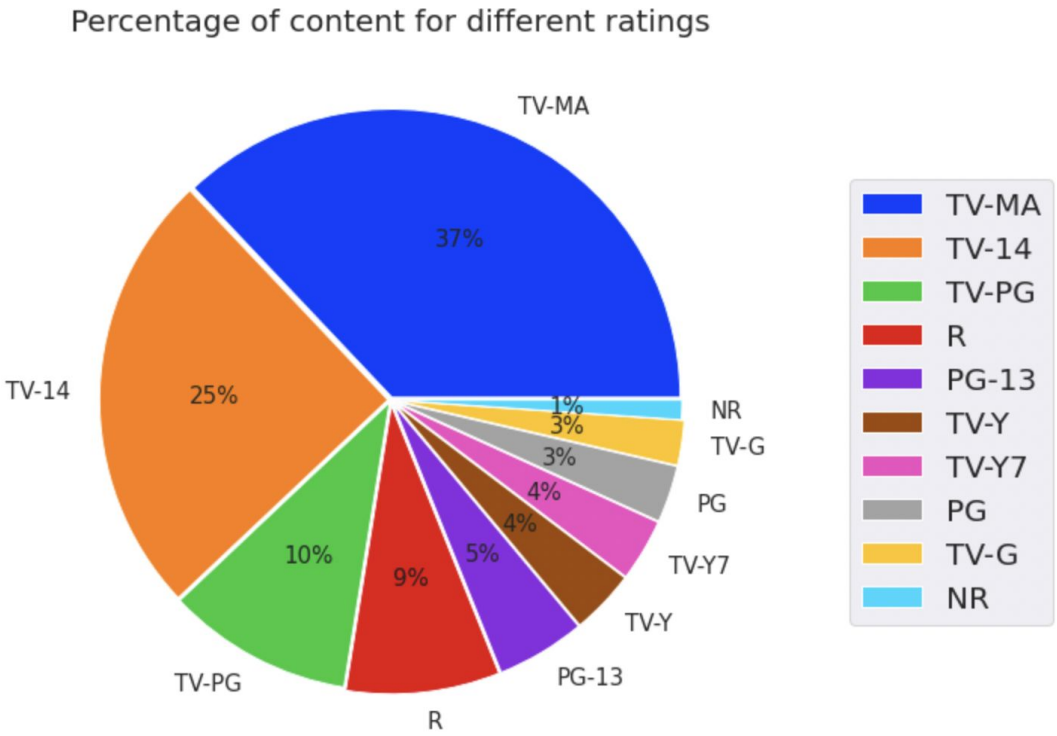
Top 10 actors that acted in most number of movies



- ➔ Anupam Kher has acted in the most number of films and TV shows.
- ➔ Count of movies is the same for Amitabh Bachchan, Yuki Kaji, Paresh Rawal and Boman Irani.

Which rating has the most number of movies/tv shows on Netflix?

- We can clearly see from the pie plot that most of the content on Netflix is for mature audiences.
- Not rated content is the lowest on Netflix



Feature Engineering

Text pre-processing

The text preprocessing is an important step where the words that are insignificant for the machine learning model are removed such as punctuation, special characters, stopwords etc. The objective is to remove words that carry less weightage in context to the text.

This is done in three steps:

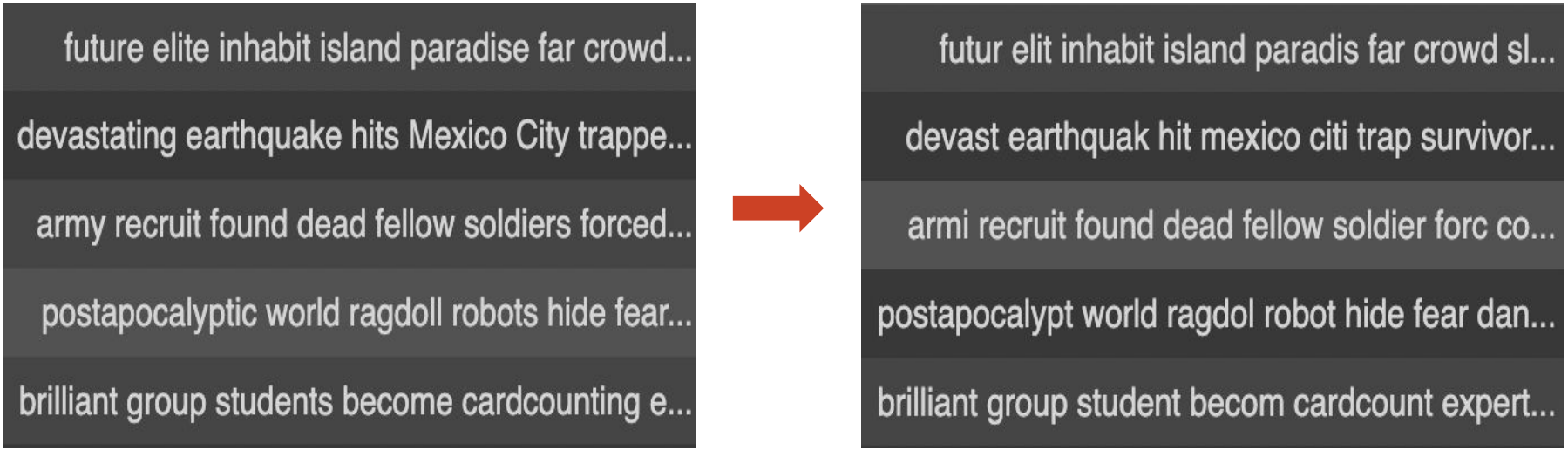
1. Remove punctuation
2. Remove stopwords
3. Apply stemming

Step 1- Remove punctuation

In a future where the elite inhabit an island ...	In a future where the elite inhabit an island ...
After a devastating earthquake hits Mexico Cit...	After a devastating earthquake hits Mexico Cit...
When an army recruit is found dead, his fellow...	When an army recruit is found dead his fellow ...
In a postapocalyptic world, rag-doll robots hi...	In a postapocalyptic world ragdoll robots hide...
A brilliant group of students become card-coun...	A brilliant group of students become cardcount...

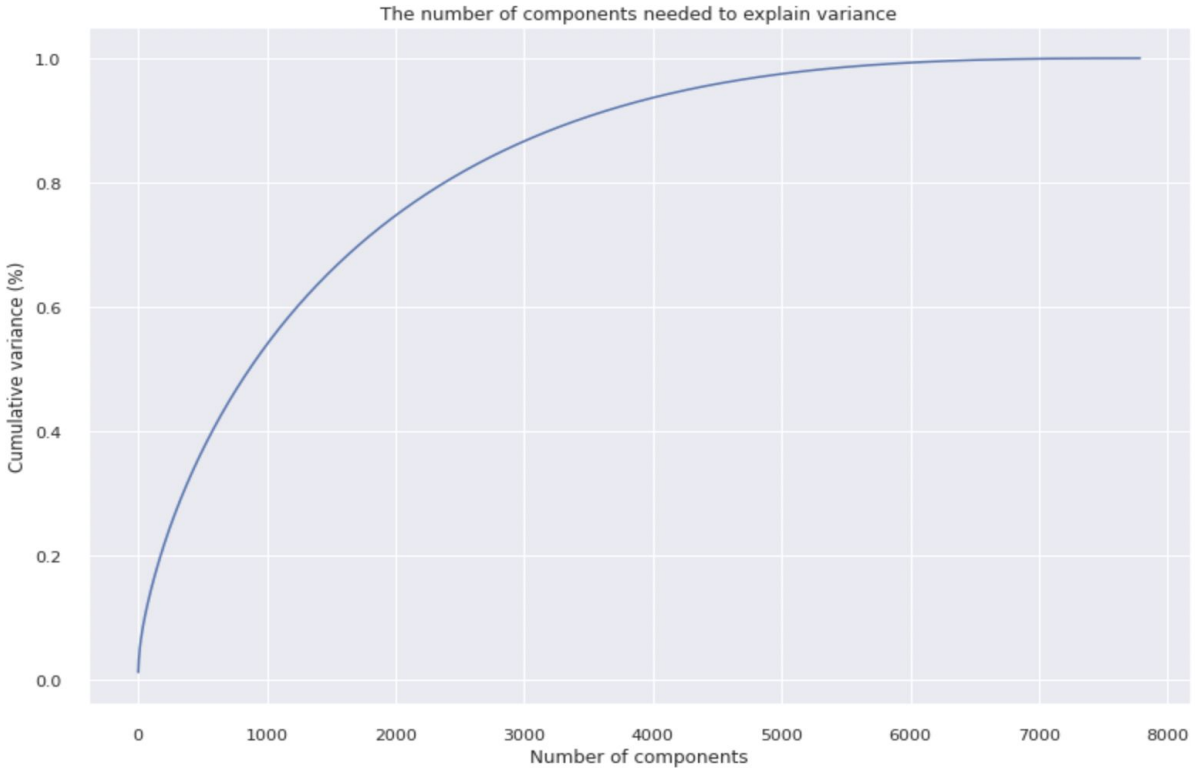
Step 2 - Remove stopwords

Step 3 - Stemming



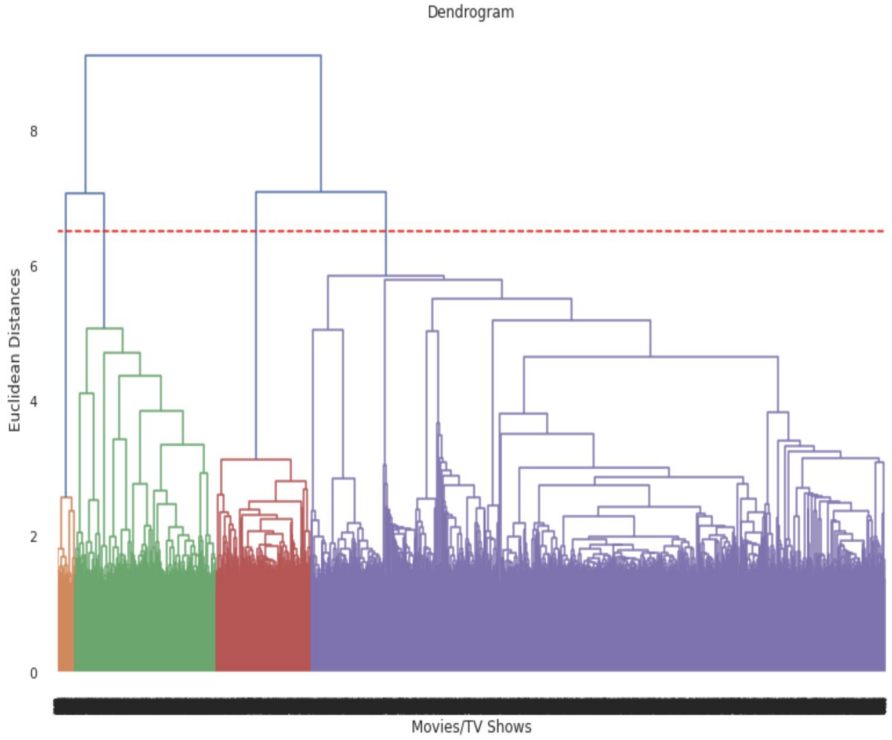
PCA for dimensionality reduction

In this case, to get the 95% of variance explained, 5000 principal components are required.

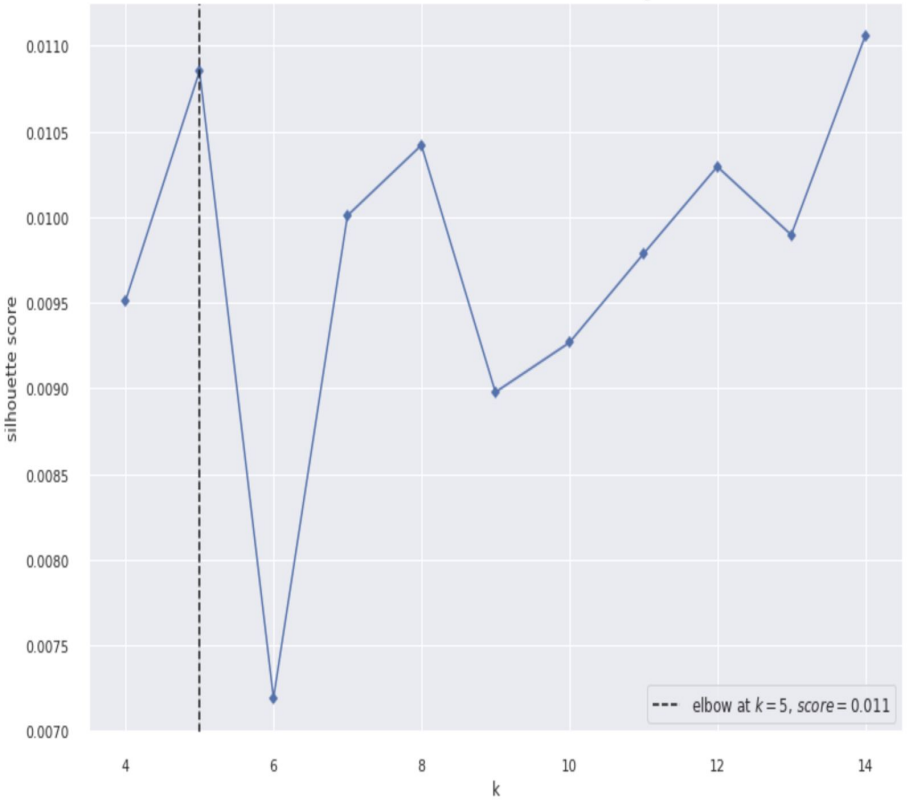


Clustering Algorithms

Hierarchical Clustering

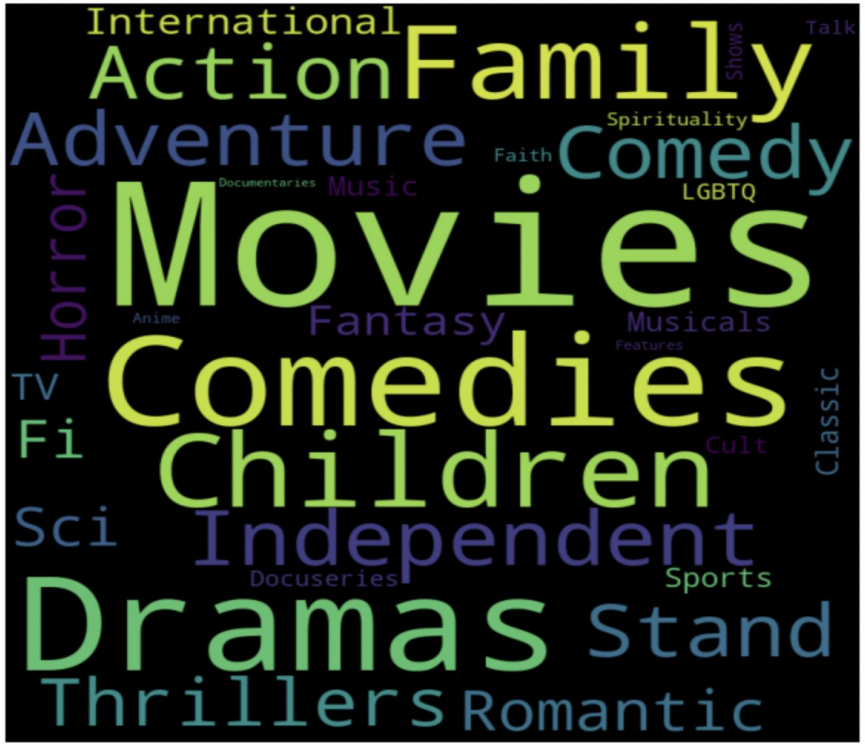


Silhouette Score Elbow for KMeans Clustering

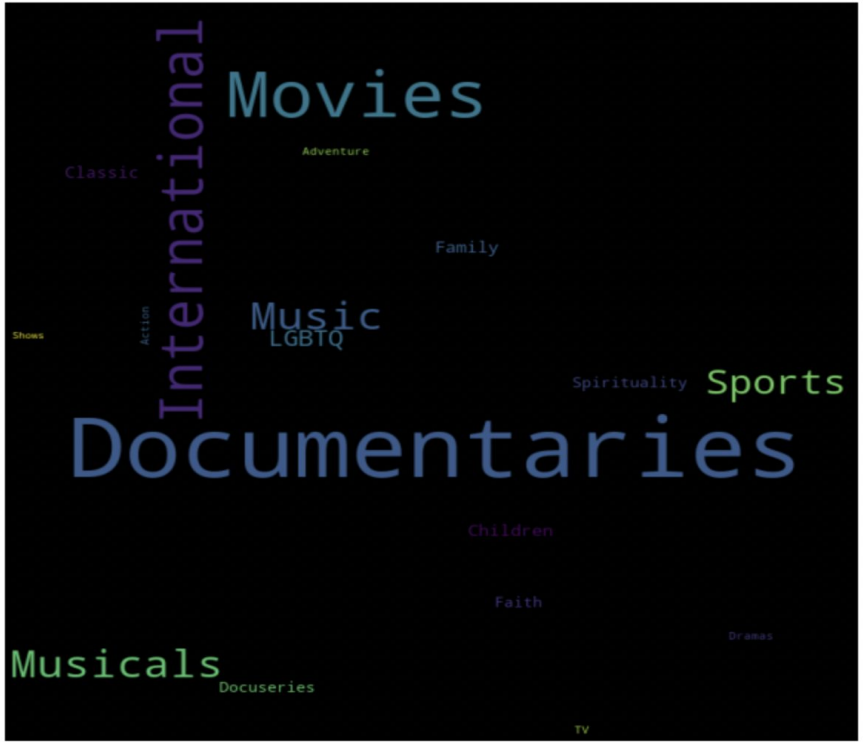


Word Clouds

Cluster 0: "Family movies - Comedy and Drama"



Cluster 1: "Documentaries and sports movies"



Word Clouds

Cluster 2: "International TV Shows - Crime, Drama and Romantic"

Cluster 3: "International Movies - Adventure, Comedy and Drama"

Word Cloud

Cluster 4: “Kids TV Shows - Animation”

Conclusion

- Most of the content available on Netflix is for mature audiences which shows the demand for kids' content is low.
- In recent years, Netflix has increasingly focused on TV shows rather than movies.
- Most movies/TV shows available on Netflix were directed by Raúl Campos and Jan Suter.
- Based on the content available on Netflix, the United States has produced the highest number of movies and TV shows.
- Japan has more TV shows than movies on Netflix.
- Anupam Kher has acted in most movies and shows available on Netflix.
- Hierarchical clustering formed 4 clusters whereas K-means formed 5 clusters.
- The optimal value of k comes out to be 5 using the elbow method and Silhouette score.
- Content is divided into 5 clusters: Family movies, Documentaries, International TV Shows, International movies and Kids' TV shows.