Netflix Movies & TV Shows Clustering

Md Shahrukh Khan, Anis Bagwan
Dayanand Kshirsagar Ankit Rai
Data Science Trainees,
AlmaBetter, Bangalore

Abstract

With the advent of streaming platforms, there's no doubt that Netflix has become one of the important platforms for streaming. The dataset that we have used for EDA and clustering has been collected by Flixable, a third-party Netflix search engine. There are 12 features and around 7700 observations in the dataset and are mostly textual features.

Through univariate and multivariate analysis, we found trends that will help in understanding what content is being consumed country-wise, depending on some categorical features like rating, type, genres, cast, directors, etc. Clustering was performed along with NLP on textual columns and then a mini-recommendation system was built out of it.

Keywords—Machine Learning, Explanatory Data Analysis, Netflix, TV Shows, Movies, Genre,Clustering, K Means.

Introduction

Unsupervised Learning is a machine learning technique in which the models are not supervised by the training set instead we find hidden patterns and insights from the given data. It is a machine learning technique in which models are trained on the unlabeled data set without any supervision.

A cluster is a collection of elements that are similar to each other but dissimilar to the elements belonging to other clusters. Clustering can be done using various kinds of distances such as Euclidean distance, Manhattan distance, gomer distance, etc. We can do different kinds of clustering based on the data pattern in space such as spherical clustering, K-means clustering, etc.

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. It will be interesting to explore what all other insights can be obtained from the same dataset.

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

Our goal here is to make an unsupervised clustering model, which will help in garnering insights on Netflix and how its content is being consumed.

<u>A brief summary of the dataset is given below:</u>

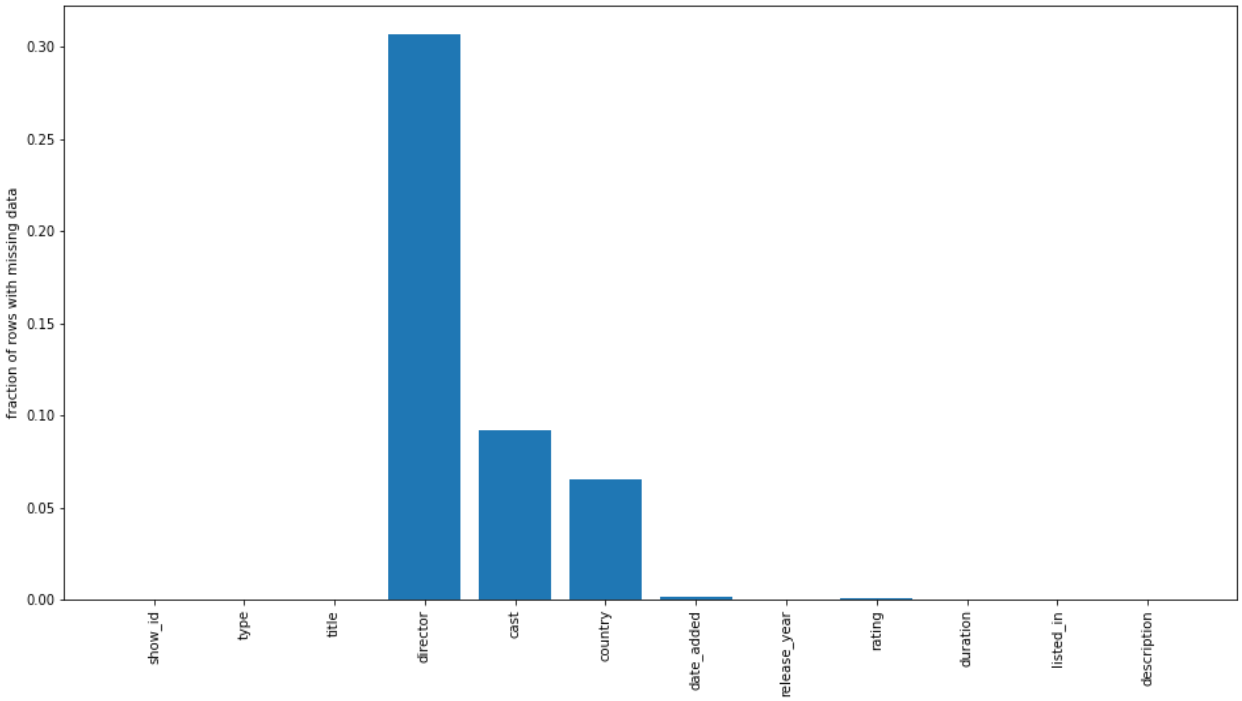
Show id: Unique ID for every Movie / TV Show
type – Identifier - A Movie or TV Show
title – Title of the Movie / TV Show
director-director of the content
cast –Actors involved in the movie / show
country – Country where the movie / show was produced
date_added – Date it was added on Netflix
release_year – 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

Exploratory Data Analysis

The first step involved in the analysis is to load the dataset into the pandas data frame. Before exploring the data using different libraries available in python we should if the dataset is ready to run the operations on it.

- **Data Cleaning:** Data Cleaning is one of the important steps before we start building models, in fact, there will be a significant increase in Model Performance when we have a clean, rich dataset. So here, we decided to replace null values with an empty string.
  - There are 2389 null values in Director column
  - There are 718 null values in cast column
  - There are 507 null values in country column
  - There are 10 null values in date added column
  - There are 7 null values in rating column

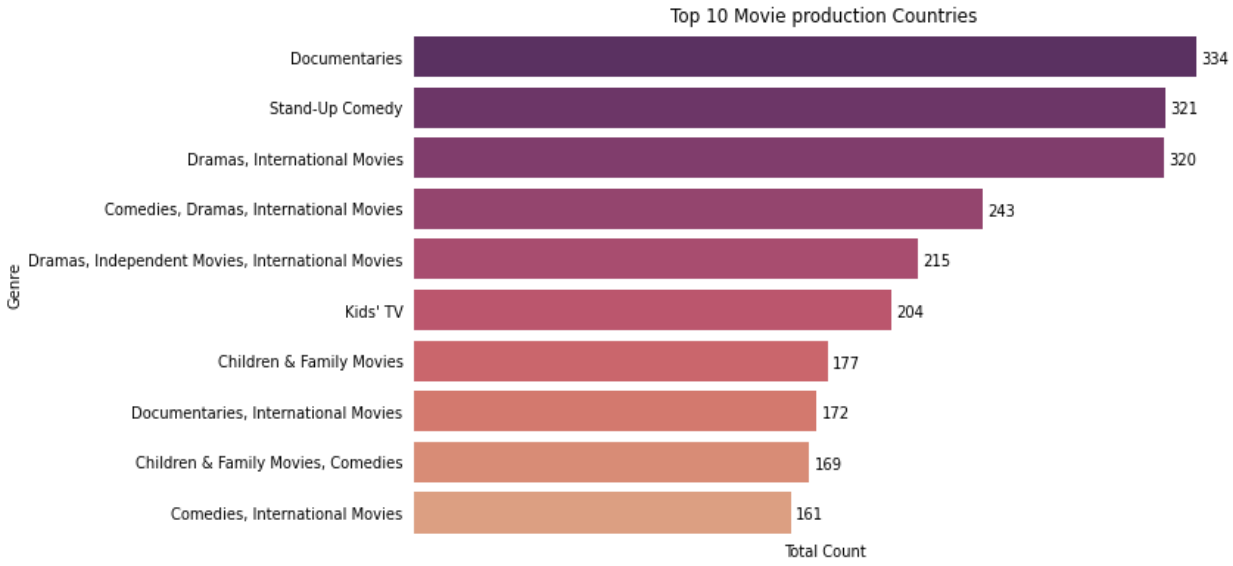
- **Handling Duplicates:** There were no duplicates.



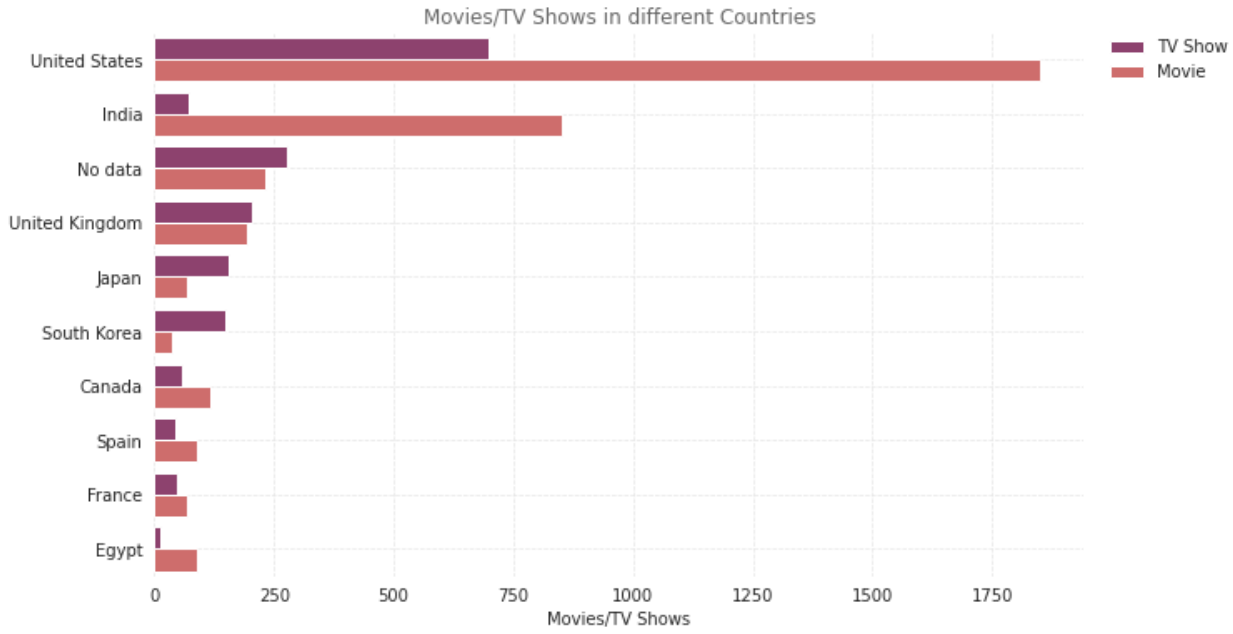
Design and Methodology

In this section, we will discuss the framework, extraction and preprocessing features, feature selection, and clustering algorithms.

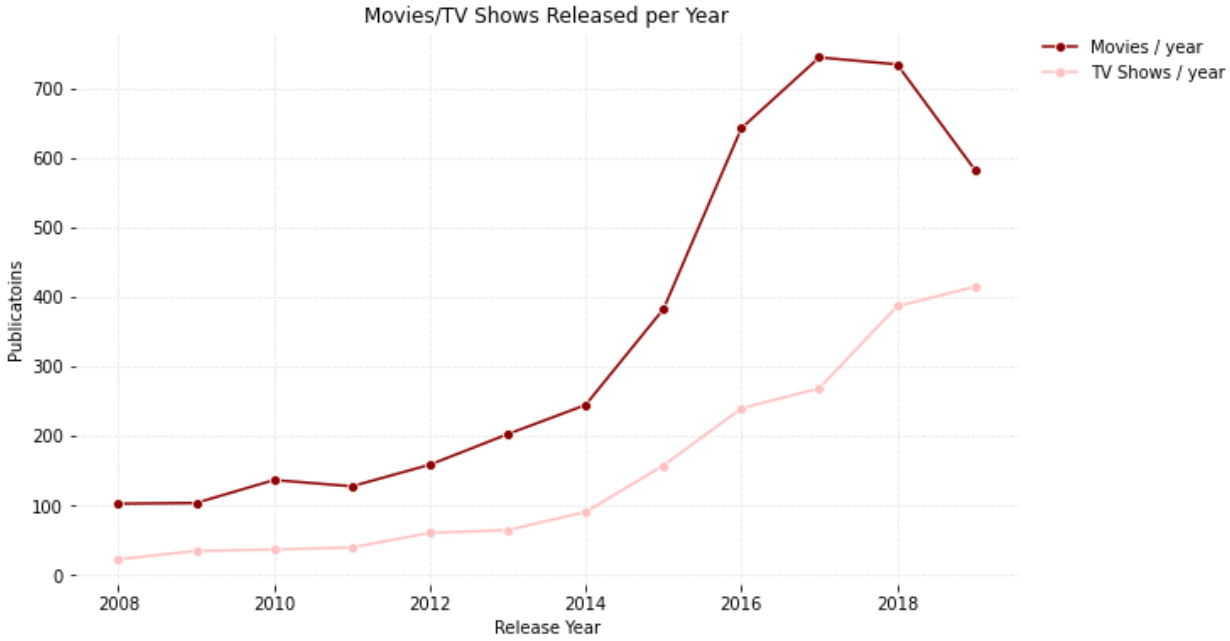
Total Count of Top 10 Movie production Countries



Movies/TV Shows in different Countries



- **Is Netflix increasingly focused on TV rather than Movies in recent years?**

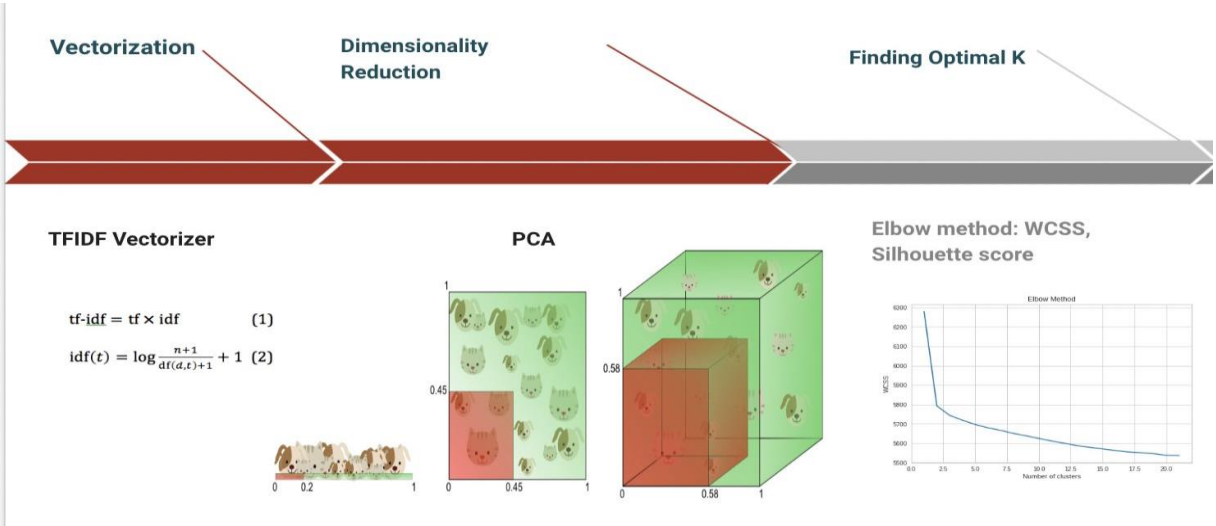


Yes, Netflix is increasingly focusing on TV Shows now, which is clear from the graph, in 2020, there were more Shows than Movies. Also, Movies preference shows a declining graph, whileshows are increasing.

Textual Columns

We clubbed various textual columns together and merged them into one final feature, which we used for clustering. We first started off by replacing null values in the columns with an empty string, followed by the removal of stopwords, tokenization, and stemming.

Finally, after we were done with textual preprocessing, we performed vectorization of the final text column using TFIDF followed by dimensionality reduction using PCA.

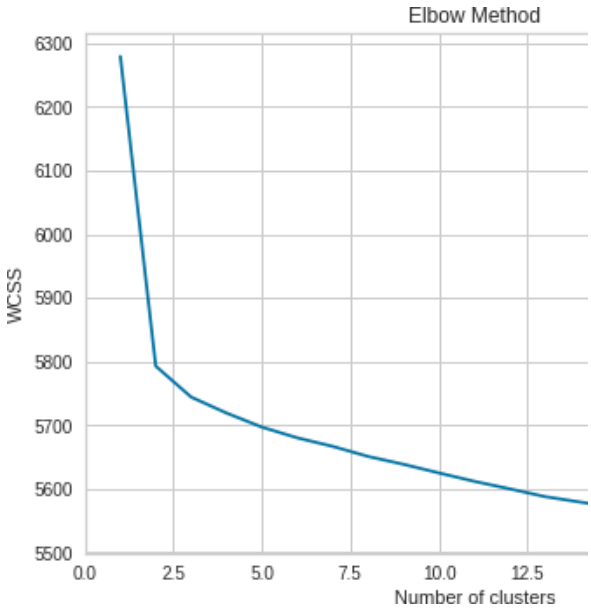


Clustering

1. K-means Clustering

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.

We created the sample data using build blobs and used range_n_clusters to specify the number of clusters we wanted to utilize in k means.



Here we will see the output of 2,3,4 and 5 number of clusters.

Silhouette score:

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 observations belonging to all the clusters.

For n_clusters = 2, silhouette score is 0.3551415129065328
For n_clusters = 3, silhouette score is 0.35586172779109915
For n_clusters = 4, silhouette score is 0.32858920525532515
For n_clusters = 5, silhouette score is 0.3348785202036102

So in this model the 3 clusters are giving best result.
So that we will consider 3 clusters as optimum clusters

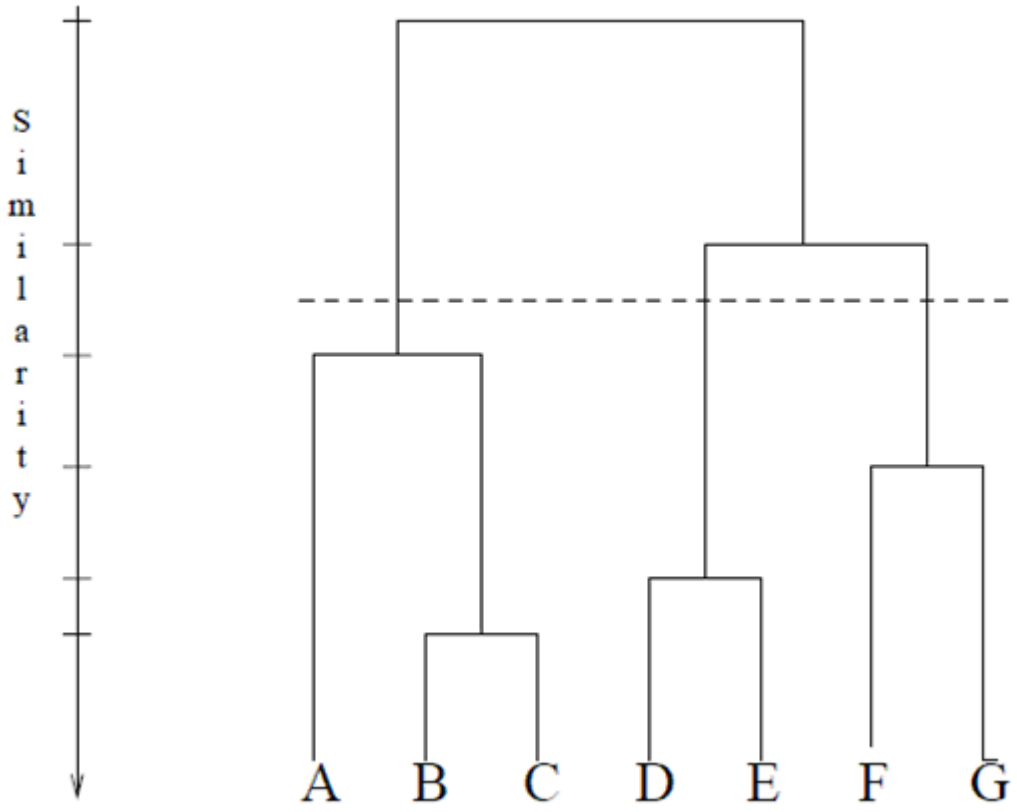
We have also done an interactive clustering visualisation in the notebook.

2. Hierarchical clustering

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. As shown in following figure:

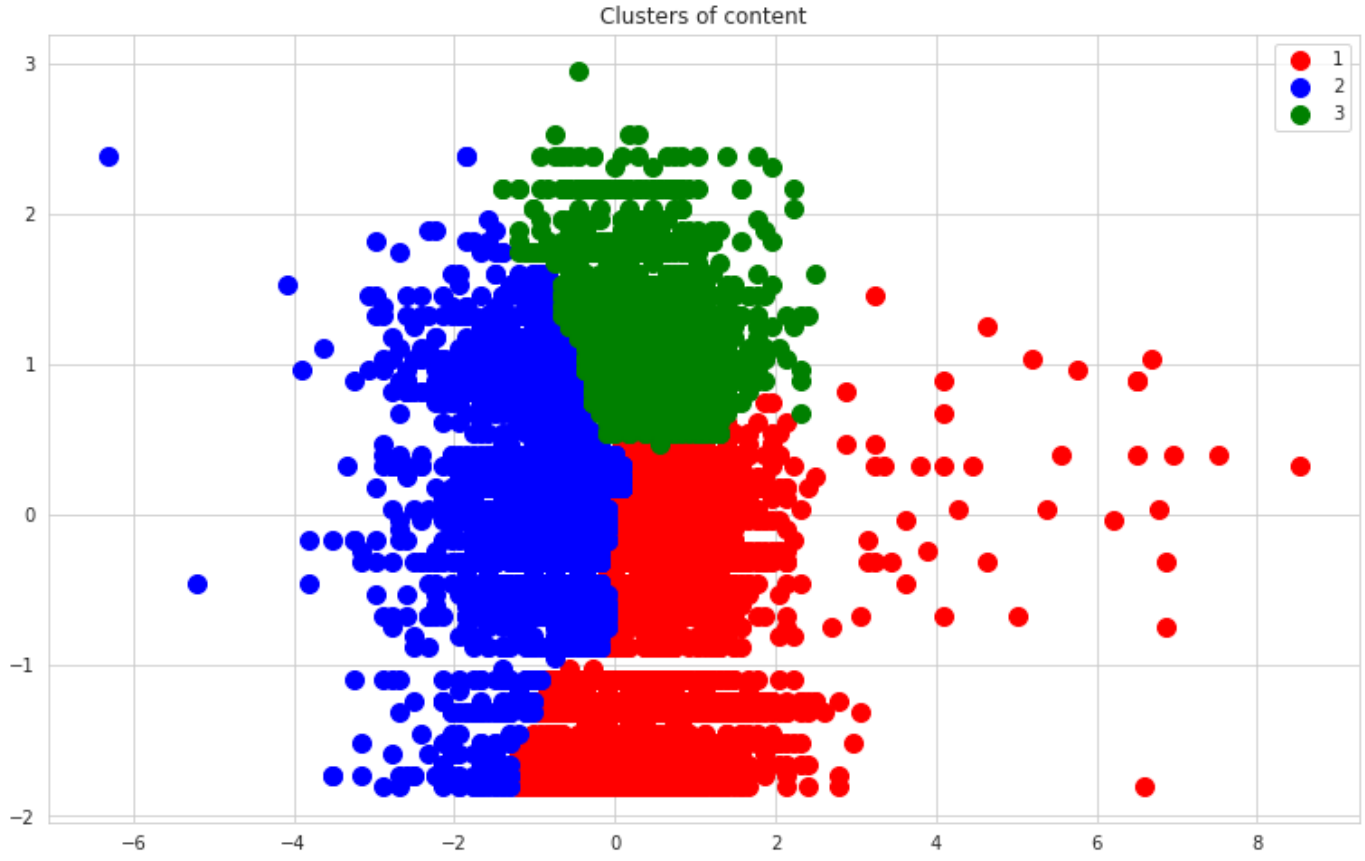


The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold

So we consider, no. of Cluster = 3

3. 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.



Future Scopes

- More Post Cluster Analysis
- Integrate the Netflix dataset with other datasets and present more insights and clusters.
- We could have done some more research on the recommendation system. (Based on TFIDF, rather than cosine similarity)

Conclusion

After the data building and preprocessing we came up with 12 features and 7.7 k records. We clubbed textual features together and found 9 optimal clusters based on silhouette score and elbow graph and performed K-means clustering and named those clusters after inferring the data we got in each one of them.

Following conclusion were drawn

- We started by removing NaN values and converting the Netflix added date to year, month, and day using date time format.
- Most films were released in the years 2018, 2019, and 2020.
- The months of October, November, December and January had the largest number of films and television series released.
- TV shows account for 2.8 percent of the total, while movies account for 97.2 percent.
- The United States, India, the United Kingdom, Canada, and Egypt are the top five producer countries.
- Netflix has added a lot more movies and TV episodes in the previous years, but the numbers are still low when compared to movies released in the last ten years.
- We did feature engineering, which involved removing certain variables and preparing a data frame to feed the clustering algorithms.
- For the clustering algorithm, we utilized type, director, nation, released year, genre, and year.
- Affinity Propagation, Agglomerative Clustering, and K-means Clustering were utilised to build the model.
- In Affinity Propagation, we had 9 clusters and a Silhouette Coefficient score of 0.340.
- A dendrogram was used to determine the number of clusters in Agglomerative Clustering. There were two clusters, with an average silhouette score of 0.56590662228136.
- The final model we used was k-means clustering, which consisted of 2,3,4,5,6 clusters. 3 numbers of clusters gives us good fitting.
- After clustering, we can say that the number of TV shows launched in the previous years is NOT growing.
- The number of TV shows added to Netflix is higher in the last three years.

Reference

[1] Applied Science Article MDPI

[2] GeeksforGeeks

[3] Wikipedia

[4] DataCamp