Netflix Movies & TV Shows Clustering

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 Fixable, 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 Fixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflixhas 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.

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

A brief summary of the dataset is given below:

Feature Name	Feature 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**

Design and Methodology

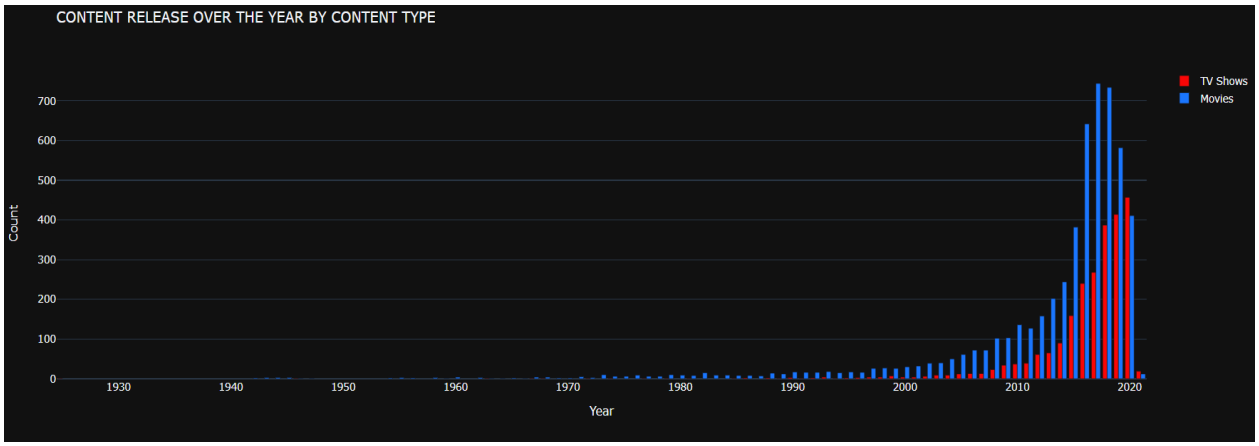
In this section, we will discuss the framework, extraction and preprocessing features, feature selection, and clustering algorithms.

Exploratory Data Analysis

The first step involved in the analysis is to load the dataset into the panda's data frame. Before exploring the data using different libraries available in python we should if the dataset is ready torun 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

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



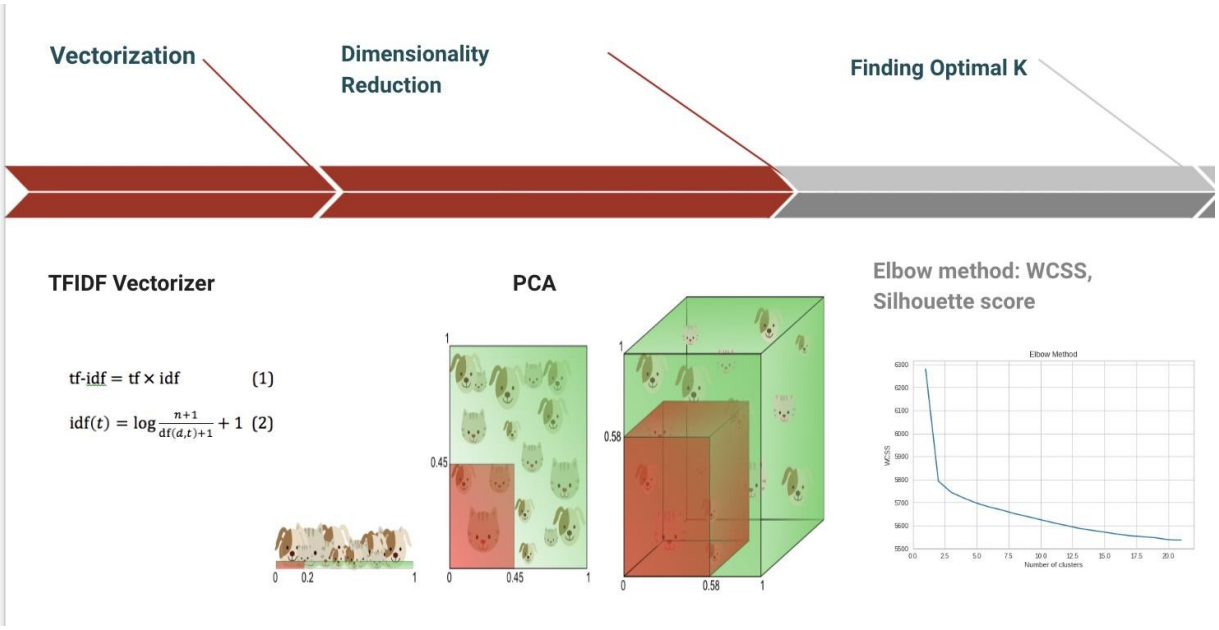
Yes, Netflix is increasingly focusing on TV Shows now, which is clear from the graph, in 2020, there were more Shows than Movies. Also, Movie's preference shows a declining graph, while shows are increasing.

- **Textual Columns**

We clubbed 6 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 stop words, tokenization, and stemming.

1. CLEANING	2. STOPWORDS	3. TOKENIZATION	4. STEMMING
• Cleaned Null values • All Columns: Only characters selected by regex • All words to lowercase • Merged text columns	• Removed Stop words • Normal english words & problem specific	• Splitted sentences to tokens • Used word_tokenise from nltk	• Transformed words to roots • Used Snowball Stemmer

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**

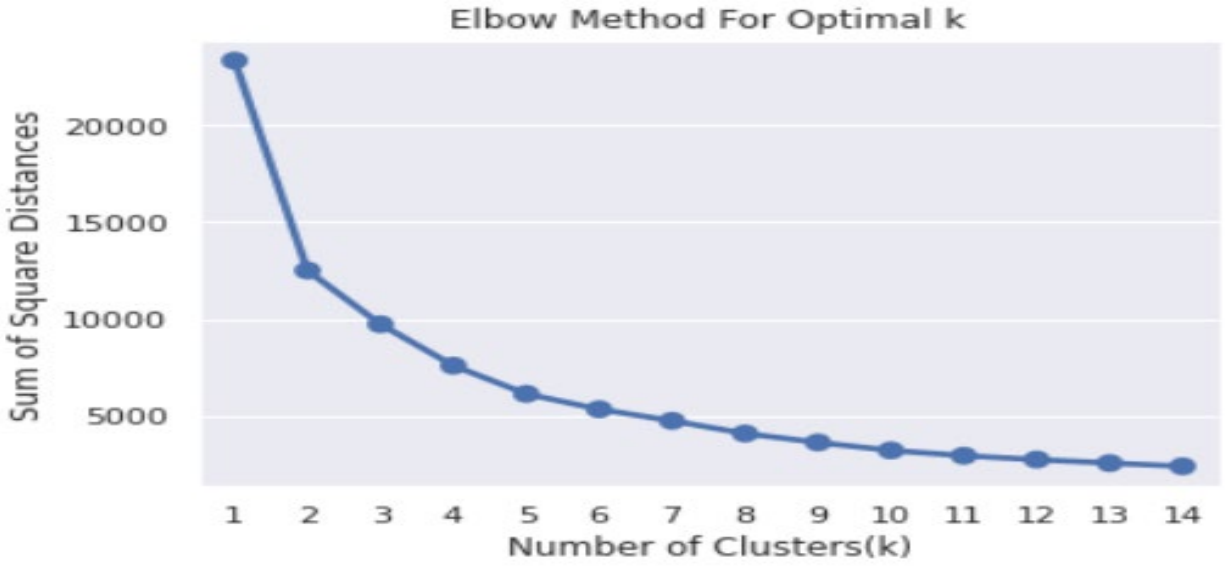
We have used 5 algorithms for the clustering of our model.

1. SILHOUETTE SCORE

	n clusters	silhouette score
0	2	0.428
1	3	0.383
2	4	0.374
3	5	0.371
4	6	0.369
6	8	0.369
8	10	0.364
7	9	0.362
5	7	0.360
9	11	0.358
10	12	0.355
11	13	0.352
12	14	0.336
13	15	0.326

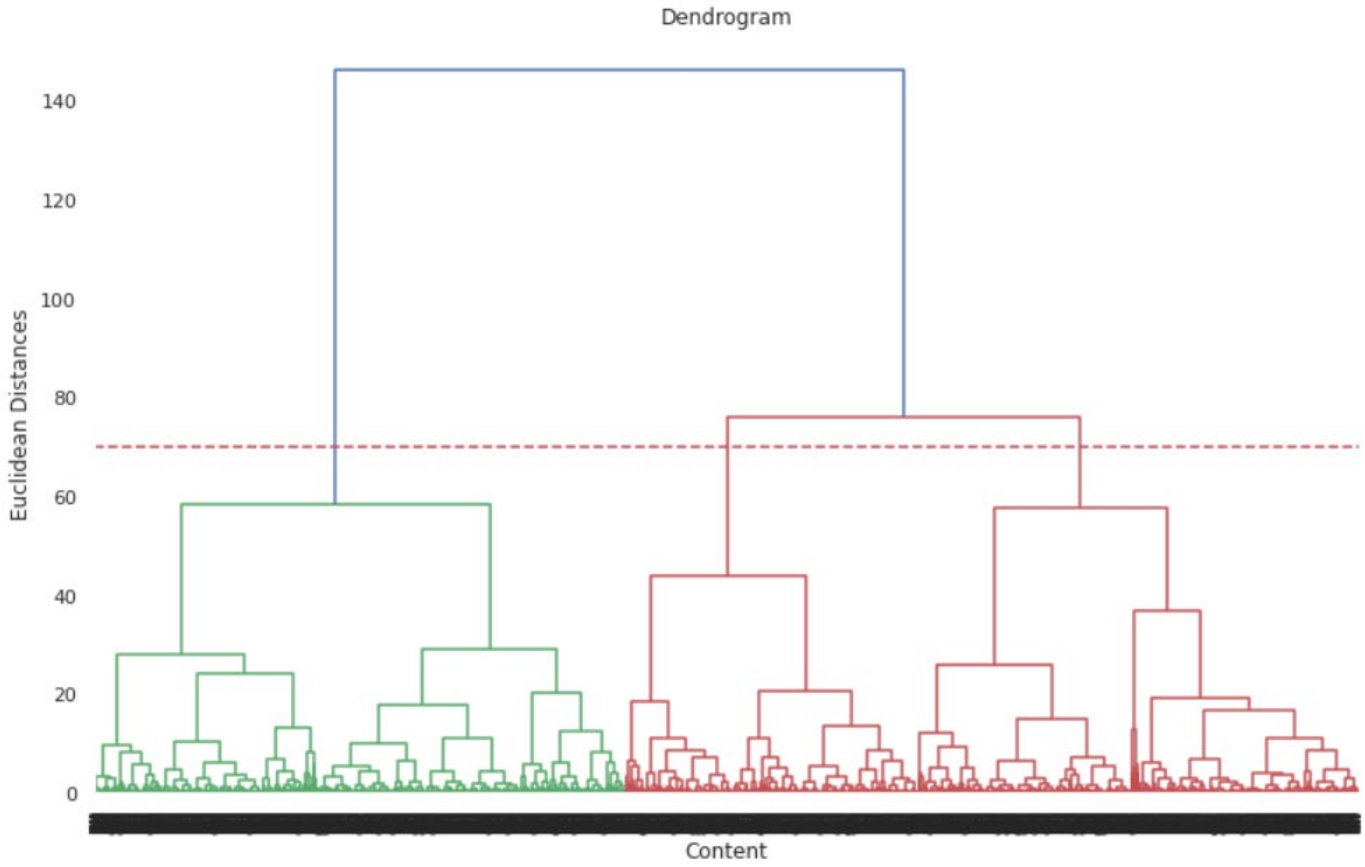
The value of the **silhouette coefficient** is between [-1, 1]. **A score of 1 denotes the best meaning that the data point I is very compact within the cluster** to which it belongs and far away from the other clusters. *The worst value is -1*. Values near 0 denote overlapping clusters

2. ELBOW METHOD



From this we can conclude that the clusters will be using is k=3 as the elbow graph shows the elbow bend at the point 3.

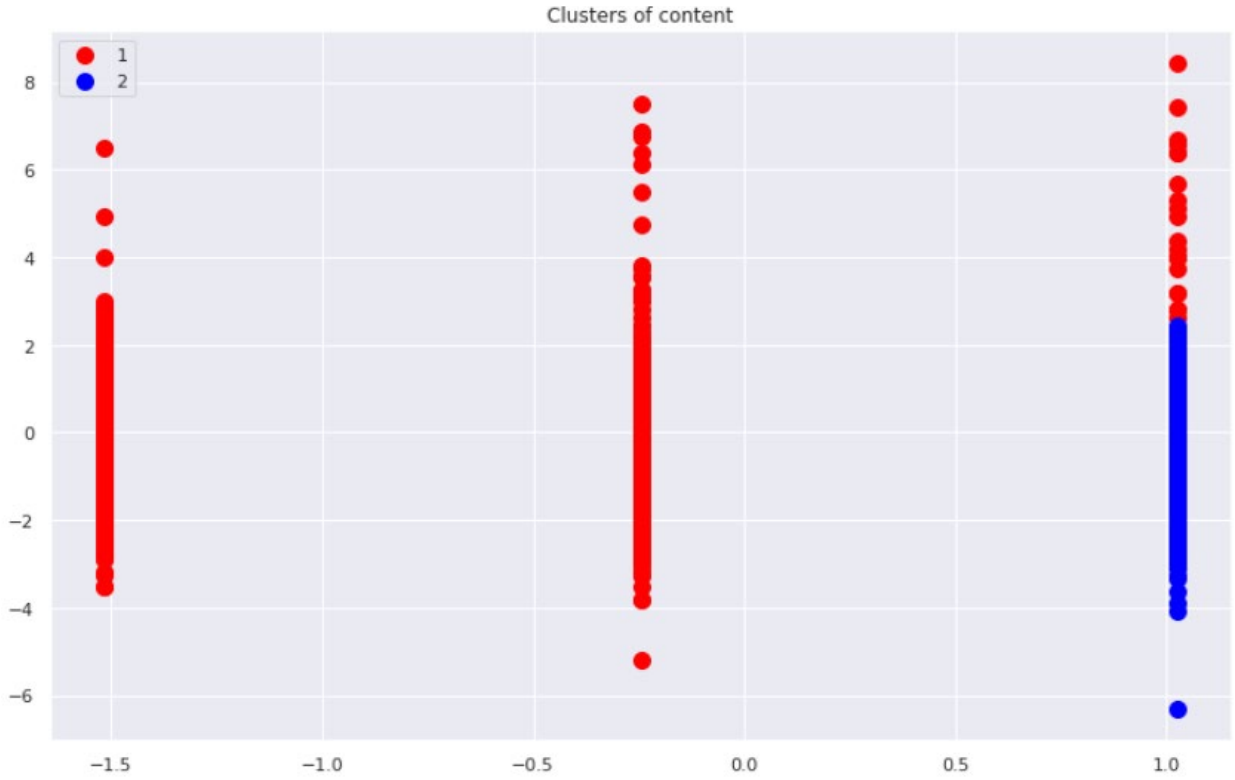
3. DENDROGRAM



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

No. of Cluster = 3

4. AGGLOMERATIVE CLUSTERING



The agglomerative clustering is the **most common type of hierarchical clustering used to group objects in clusters based on their similarity**. It's also known as AGNES (Agglomerative Nesting). The algorithm starts by treating each object as a singleton cluster.

CONCLUSION

1. Director and cast contain a large number of null values so we will drop these 2 columns.

2. In this dataset there are two types of contents where 30.86% includes TV shows and the remaining 69.14% carries Movies.

3. We have reached a conclusion from our analysis from the content added over years that Netflix is focusing movies and TV shows (Form 2016 data we get to know that Movies is increased by 80% and TV shows is increased by 73% compare)

4. From the dataset insights we can conclude that the greatest number of TV Shows released in 2017 and for Movies it is 2020

5. On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows.

6. Most of the movies are belonging to 3 categories

7. TOP 3 content categories are international movies, dramas, comedies.

8. In text analysis (NLP) I used stop words, removed punctuations, stemming & TF-IDF vectorizer and other functions of NLP.

9. Applied different clustering models like Kmeans, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements.

10.By applying different clustering algorithms to our dataset. we get the optimal number of clusters is equal to 3