### Unsupervised ML Clustering

Capstone Project

of

Netflix Movies and TV Shows Clustering

by: Prajwal M V

  Data Science Trainee, Alma Better

**Abstract:**OTT platforms have become a normal and usual mode of entertainment in today’s world. There are many popular OTT platforms like Amazon Prime, BigFlix, Arre, Discovery+, etc., competing with each other to increase their user base. Netflix is one of the leading OTT platforms, not only in India but also internationally. The success of the OTT platforms depends on two things- the variety of content and appropriate recommendations to the users. Clustering is a useful technique to achieve the best possible recommendations and increase the viewership of the platform.

**1.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 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. Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

**Attribute Information:**

The dataset provided contains 7787 rows and 12 columns.

The following are the columns in the dataset:

* Show id: Unique identifier of the record in the dataset
* Type: Whether it is a TV show or movie
* Title: Title of the show or movie
* Director: Director of the TV show or movie
* Cast: The cast of the movie or TV show
* Country: The list of the country in which a show/ movie is released or watched
* Date added: The date on which the content was on boarded on the Netflix platform
* Release year: Year of the release of the show/ movie
* Rating: The rating informs about the suitability of the content for a specific age group
* Duration: Duration is specified in terms of minutes for movies and in terms of the number of seasons in the case of TV shows
* Listed in: This columns species the category/ genre of the content
* Description: A short summary about the storyline of the content.

**3. Steps Involved**

**I. Exploratory Data Analysis:**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations. It gives us better idea of which feature behaves in which manner compared to target variable.

**II. Data Cleaning:**

Our Data contains some null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project to get better results.

**III. Feature Scaling:**

Feature scaling is essential for machine learning algorithms that calculate distances between data. If not scale, the feature with a higher value range starts dominating when calculating distances.

**IV. Fitting Models:**

We used the k-means clustering algorithm and then checked the model performance using Silhouette’s coefficient and elbow method to find the number of clusters.

**4.Model Building Pre-Requisites**

1)Removing Punctuations:

* Punctuations does not carry any meaning clustering.
* So, removing punctuations helps to get rid of unhelpful parts of the data, or noise.

2)Removing Stop words:

* Stop words are basically a set of commonly used words in any language, not just English.
* If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

3)Stemming:

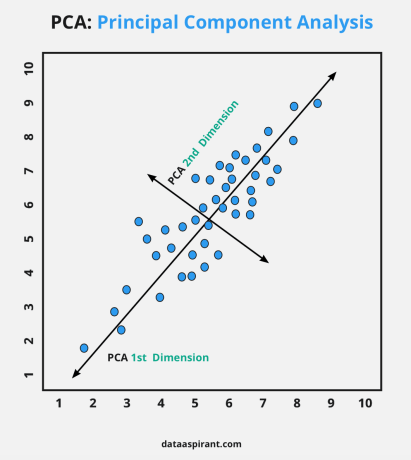
* Stemming is the process of removing a part of a word, or reducing a word to its stem or root.
* Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language

4)Vectorizing the text data using TF\_IDF Vectorizer:

* Here we have textual data.
* Clustering algorithms cannot understand textual data.
* So, we use vectorization technique to convert textual data to numerical vectors

**5. Model Building**

1) Principal Component Analysis:



A dimensionality-reduction technique in which transformation of high dimensional correlated data is performed into a lower-dimensional set of uncorrelated components also referred to as principal components.

Its Main Objectives are:

* To visualize the high dimensionality data.
* To introduce improvements in classification.
* To obtain a compact description.
* To capture as much variance in the data as possible.
* To decrease the number of dimensions in the dataset.
* To search for patterns in the dataset of high dimensionality.
* To discard noise.

# 2)K-Means Clustering Algorithm

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

## What is K-Means Algorithm?

K-Means Clustering is an [Unsupervised Learning algorithm](https://www.javatpoint.com/unsupervised-machine-learning), which groups the unlabelled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means [clustering](https://www.javatpoint.com/clustering-in-machine-learning) algorithm mainly performs two tasks:

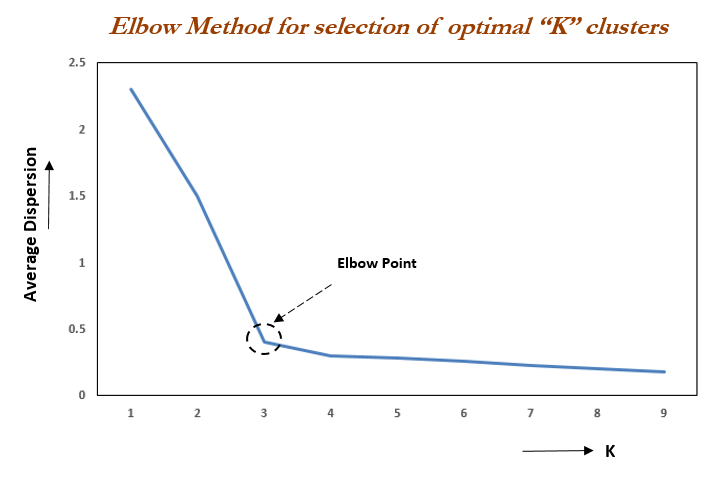
* Determines the best value for K centre points or centroids by an iterative process.
* Assigns each data point to its closest k-centre. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has data points with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:

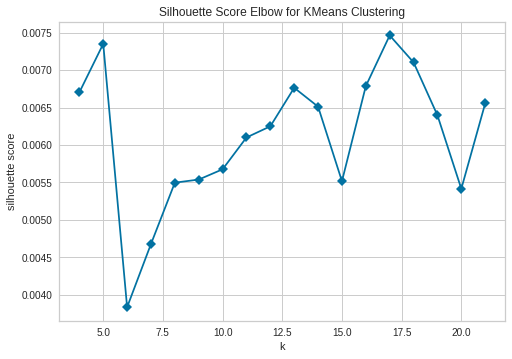


### **3)Elbow method to find optimum k value:**



Elbow Method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. Pick the value of k, where the average distance falls suddenly.

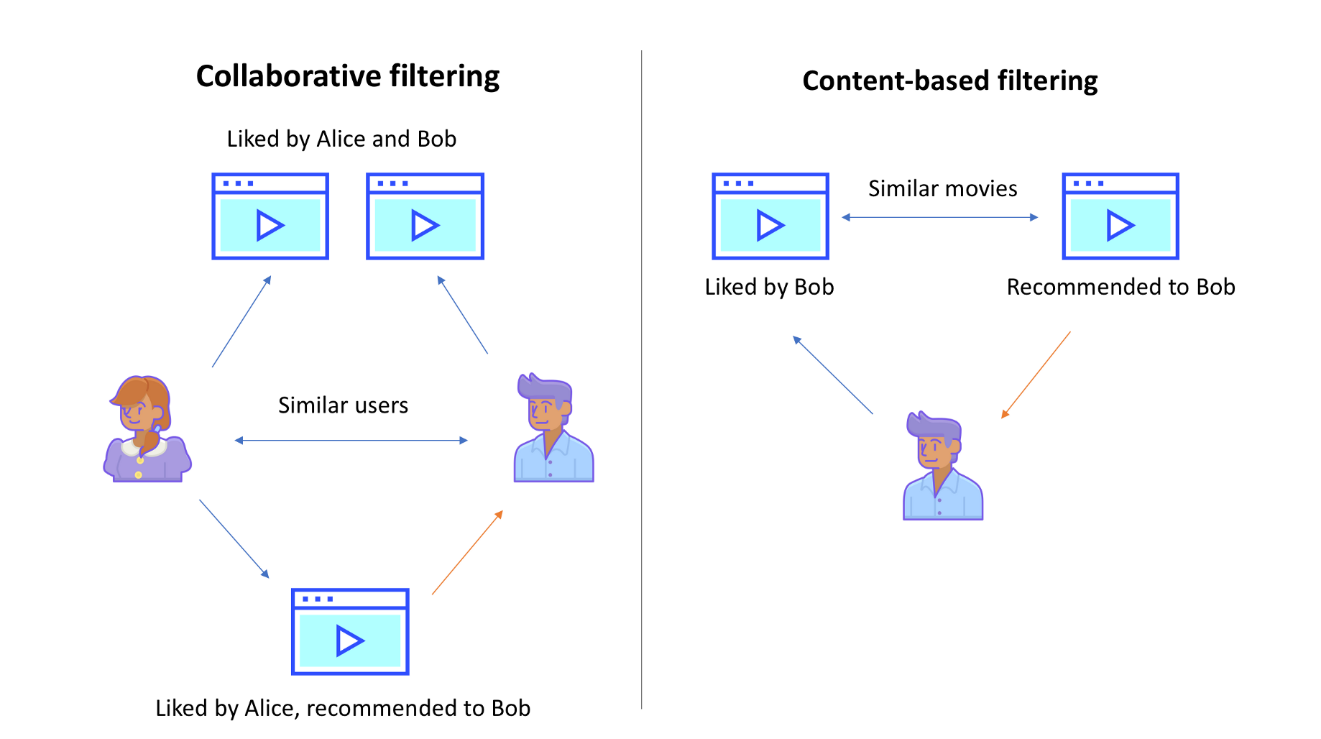
## 4)Find the optimal value of ‘k’ using Silhouette Analysis:



The silhouette Method is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. by providing a succinct graphical representation of how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. Silhouette coefficients are computed for each of point, and averaged out for all the samples to get the silhouette score.

5)RECOMMENDER SYSTEMS

A recommendation system is a subclass of Information Filtering Systems that seeks to predict the rating or the preference a user might give to an item. In simple words, it is an algorithm that suggests relevant items to users. E.g.: In the case of Netflix which movie to watch, In the case of e-commerce which product to buy, or in the case of kindle which book to read, etc.



## **Collaborative Filtering**

### This method makes automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on a set of items, A is more likely to have B's opinion for a given item than that of a randomly chosen person.

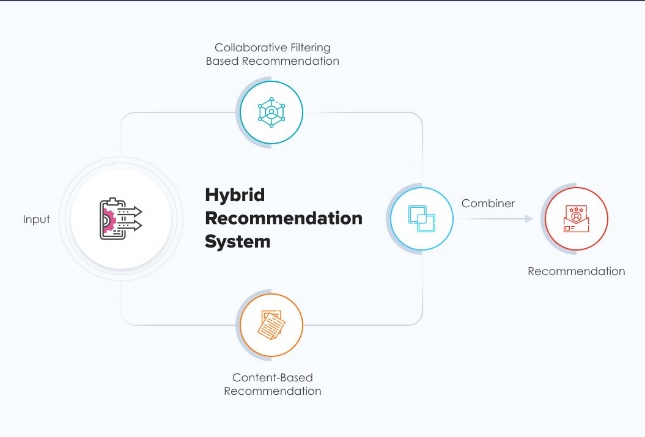
## **Content-Based Filtering**

### This method uses only information about the description and attributes of the items users has previously consumed to model user's preferences. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past (or is examining in the present). In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended.

## **Hybrid Approach**

### Recent research has demonstrated that a hybrid approach, combining collaborative filtering and content-based filtering could be more effective than pure approaches in some cases. These methods can also be used to overcome some of the common problems in recommender systems such as cold start and the sparsity problem.

Example: NETFILX



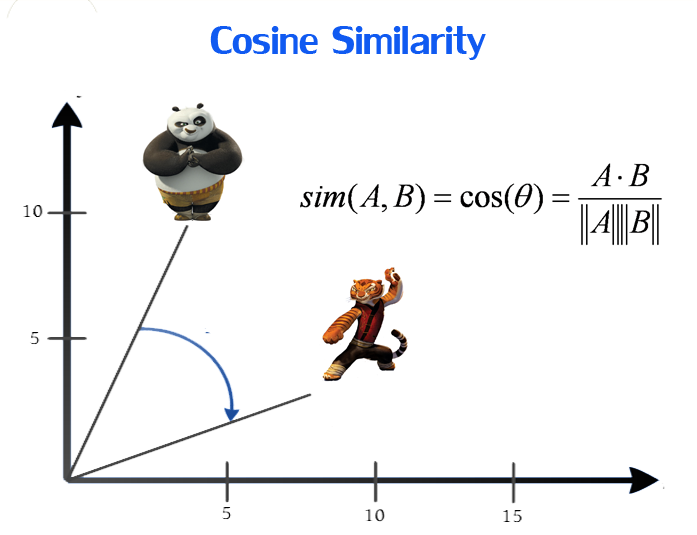
# 6)Cosine Similarity

# What is Cosine Similarity?

# Cosine similarity is a metric used to measure the similarity of two vectors. Specifically, it measures the similarity in the direction or orientation of the vectors ignoring differences in their magnitude or scale. Both vectors need to be part of the same inner product space, meaning they must produce a scalar through inner product multiplication. The similarity of two vectors is measured by the cosine of the angle between them.

# How to calculate Cosine Similarity?

# We define cosine similarity mathematically as the dot product of the vectors divided by their magnitude. For example, if we have two vectors, A and B, the similarity between them is calculated as:



Cosine Similarity is widely used in Data Science and Machine Learning applications. Examples include measuring the similarity of:

* Documents in natural language processing
* Movies, books, videos, or users in recommendation systems
* Images in computer vision

**8. Conclusion:**

The analysis is done with Netflix data. We used the k-means clustering algorithm and then checked the model performance using Silhouette’s coefficient and elbow method to find the optimal number of clusters and finally plotted the Word cloud to visualize various clusters.

**The experimental results show that:**

* It was interesting to find that majority of the content available on Netflix is Movies.
* But in the recent years it has been focusing more on TV-Shows.
* Most of these contents are released either in the year ending or the beginning.
* United States and India are among the top 5 countries that produce all of the available content on the platform.
* Also 6 of the actors among the top ten actors with maximum content are from India.
* TV-MA tops the charts, indicating that mature content is more popular on Netflix.
* k=10 was found to be an optimal value for clusters using which we grouped our data into 10 distinct clusters.
* Using the given data, a simple recommender system was created using cosine\_similarity and recommendations for Movies and TV Shows were obtained.

# **Future Scope:**

* Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.
* More time could be given into building a better recommender system, which later can be deployed on web for usage.