**Netflix Movies & TV Shows Clustering**

**Akash Salmuthe**

**Data Science Enthusiast**

**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.***

**1. 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.

**2. 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.

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.

1. Exploratory Data Analysis
2. Understanding what type of content is available in different countries
3. Is Netflix has increasingly focusing on TV rather than movies in recent years
4. Clustering similar content by matching text-based features

**3. Data Description**

This dataset consists of TV shows and movies available on Netflix as of 2019.

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.

The dataset contains following columns:

* **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

**4. Rise of Netflix**

**Reasons:**

1. Entail for Original Content
2. Personalized Recommendation
3. Series of Successful Shows
4. Unique Marketing Strategy
5. Revolutionizing the Industry

**5. Steps Involved:**

* **Exploratory Data Analysis**

After mounting our drive and fetching and reading the dataset given, we performed the Exploratory Data Analysis for it.

To get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step.

* **Null values Treatment**

*Director* feature have more than 30% of null values. So, dropping feature director.

*Country* feature have 6.51% of null values. Filling null values by mode of feature.

*Cast* feature have 9.22% of null values. Filling null values by ‘missing’.

*Rating* feature have 0.09% of null values. Filling null values by mode of feature.

*Date\_added* feature have 0.12% of null values. Dropping rows corresponding to null values.

* **Duplicate Values Treatment**

Duplicate values dose not contribute anything to accuracy of results.

Our dataset dose not contains any duplicate values

* **Data Type Change**

Features in their appropriate data type provides better understanding and work ability on that data.

Date\_added feature have object data type converting to datetime.

Duration is in combination of integer values and text. Removing text part so as to get integer data type.

* **New Features**

From the feature date\_added; extracted year, month and day to form new columns by name of year, month and day respectively.

* **Data Pre-Processing**

1. **Removing Punctuation**

Punctuations does not carry any meaning in clustering.

So, removing punctuations helps to get rid of unhelpful parts of the data, or noise

1. **Removing Stopwords**

Stopwordsare basically a set of commonly used words in any language, not just in English.

If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

1. **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.

**6. Algorithm**

**i) K-means Clustering:**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

**How the K-means algorithm works:**

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

It halts creating and optimizing clusters when either:

* The centroids have stabilized — there is no change in their values because the clustering has been successful.
* The defined number of iterations has been achieved.



Fig: K-means Clustering

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non overlapping subgroups where each data point belongs to only one group.

**1. Vectorization:**

Here we have textual data.

Clustering algorithms cannot understand textual data.

So, we use vectorization technique to convert textual data to numerical vectors.

**2. Elbow Curve:**

The Elbow Curve is one of the most popular methods to determine this optimal value of k.

The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

**3. 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.

**ii) Recommender System:**

Recommender systems are the systems that are designed to recommend things to the user based on many different factors.

It finds out the match between user and item and imputes the similarities between users and items for recommendation.

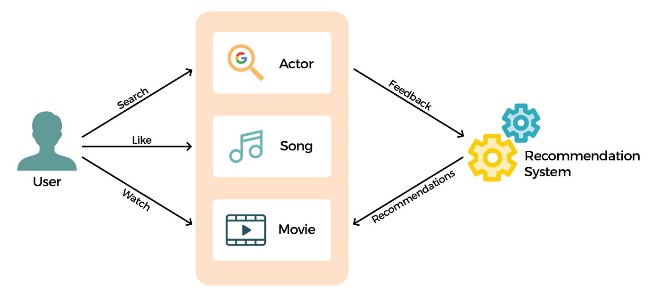
****

Fig: Recommender System

**Cosine similarity:**

Cosine similarity is a metric that measures the cosine of the angle between two vectors projected in a multi-dimensional space.

**7. Conclusion:**

That's it!

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.

**References:**

1. [*https://www.marketing360.in/the-rise-of-netflix*](https://www.marketing360.in/the-rise-of-netflix)
2. [*https://machinelearningmastery.com/clustering-algorithms-with-python*](https://machinelearningmastery.com/clustering-algorithms-with-python)
3. [*https://towardsdatascience.com/introduction-to-machine-learning-algorith*](https://towardsdatascience.com/introduction-to-machine-learning-algorith)
4. [*https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques*](https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques)