Capstone Project

Netflix Movies & TV Shows Clustering

Akash Salmuthe

Netflix and Chill !!!

1. Problem Statement
2. Introduction
3. Data Summary
4. Data Cleaning
5. Data Pre-processing
6. Exploratory Data Analysis
7. Data Pre-processing
8. K-Means Clustering
9. Recommender System
10. Conclusion

Problem Statement

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

Introduction

- Netflix is a streaming service with wide variety of content. The idea of this project is to analyze and perform clustering to determine various patterns related to the content available in Netflix.

- The data is gathered from a third party engine.

- Based on the attributes related to the TV shows or movies, we will be implementing different clustering algorithms which comes under unsupervised Machine learning category.

Data Summary

- 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.

Data Description

- **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

Data Cleaning

1. **Duplicate Values Treatment:**
   - • Duplicate values dose not contribute any thing to accuracy of results.
   - • Our dataset dose not contains any duplicate values.

Data Cleaning

2. **Null Value 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.