NETFLIX MOVIES AND TV SHOWS CLUSTERING

Name - Vikas Kumar

Abstract- This project on 'Netflix movies and TV shows clustering' is using Machine learning unsupervised clustering method. In this project, 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.

Keywords- Machine learning algorithm, EDA, Clustering, and NLP.

Problem Statements:

- Introduction
- Problem Statement
- Data Description
- Data Wrangling
- Exploratory Data Analysis
- Data pre-processing
- Word cloud
- Removing Stop words
- Removing Punctuation
- Stemming
- Clustering
- PCA
- Elbow and Silhouette Method
- K-Means Clustering
- Inference

Introduction :

Netflix employs data science to always provide us with the appropriate content. They categorise all of the information that people in a specific area are now seeing using a clustering and classification algorithm. Also, they employ a recommender system to predic a person's preferences in the future given a specific quantity of sparse data.
In this project, 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.

Steps involved in this Project : 📢

Step 1:

In the first step,I performed data wrangling to get proper data from given data. I filled the null value with NA in three features. Which are cast, country, and directors.

Step 2:

In the second step, I performed our exploratory data analysis. I analysed the trends for type vs counts, top 15 countries (content maker), director names, months added, year added, and release years etc. I also made a graph for duration which shows that the most of the movie’s duration comes between 80 to 120 minutes.

Step 3:

In the third step, I came across Natural language processing. In which I performed many techniques to get insight from the given dataset. Which words are most frequently used in the dataset. I also performed in Word cloud, Removing Stopwords, Stemming, and Removing Punctuation.

Step 4:

In the last step, I performed dimensionality reduction using PCA (Principal components analysis) then I performed elbow method and silhouette method to get an optimized value of k. After getting optimised value of k I make clusters using k-means clustering.

Conclusions 🎯

- In this given dataset of Netflix there are a total of 7787 rows and 12 columns.
- There are some null values present in some features like director, cast, country, data added, release year and rating.
- After analysing the netflix dataset, it shows 5372 movies and 2398 tv_shows.
- There are more movies than tv_shows present on netflix.
- In the USA the number of content released or added is highest followed by india.
- Most of the content were added in 2019 on netflix. I can also observe that most of the movies were added in 2019 and tv_shows in 2020.
- After 2018 the popularity of tv_shows increases with respect to movies.
- TV-MA has the highest rating for both movies and tv_shows which shows that mature and teen content is most popular on netflix. Least rating for tv_shows is TV-y7FV and movie is NC-17 which is an adult category.
- The highest duration of the movies lies between 80 to 120 minutes and most of the popular tv_shows have only 1 season.
- After that I use NLP on the text columns of our dataset in which I perform punctuation removal, removing of stopwords and then stemming.
- After doing all that I do Tf-Idf on listed_in(genre) and description column.
- So I can conclude that frequent words listed_in are tv, thriller, teen and least frequent words are adventure, action. So the popularity of tv_shows is higher than any genre of the movies. in movies most popular genre is thriller followed by teen.
- I apply Tf-Idf on clustered data which is corpus of words. Then I do PCA for dimensionality reduction.
- Then I apply clustring. For finding optimum value of k I use Elbow method and silhouette method. After that I apply clustring so that I obtain the best clustring arrangement.
- From all the analysis we have done, I can improve the future content quality of

the netflix i.e which type of content is popular among the citizens and of which genre and duration etc.
