Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) Chetan Rajput: **Email: Chetan.rajput91@yahoo.com**

Contribution:

a). Data Wrangling
b). Data Cleaning
c). EDA- TYPE OF CONTENT
d). EDA- Top Movies & TV Shows Rating Based
e). EDA– Top 10 Genre's On Netflix.
f). Distribution of Movies & TV Shows Duration
g). HYPOTHESIS TESTING
h). Data Preprocessing
i). Implementing K-means
j). Conclusion

2). Anas Mustafa. **Email: Mustafaanas84464@gmail.com**

Contribution:

a). Data Wrangling
b). Data Cleaning
c). EDA- Number Of TV Shows And Movies In Top 10 Countries
d). EDA- Number of TV Shows & Movies Release in last 15 Years
e). EDA – Director's With Most Movies & TV Shows.
f). EDA- Year & Month wise Analysis
g). HYPOTHESIS TESTING
h). Data Preprocessing
i). Implementing K-means
j). Conclusion

3) Sarthak Rastogi. **Email: sartakrastogi1@gmail.com**

Contribution:

a). Data Wrangling
b). Data Cleaning
c). EDA- Top 10 Countries
d). EDA- Top 10 Genres For TV Shows & Movies.
e). EDA – Top 10 Actor's On Netflix
f). HYPOTHESIS TESTING
g). Data Preprocessing
h). Finding number of Clusters from Silhouette's score
i). Implementing K-means
j). Conclusion

Please paste the GitHub Repo link.

Github link:- https://github.com/Chetanrajput1331/Netflix-Capstone-Projects.git

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Introduction

As we all know that Netflix is a prominent OTT platform with a wide variety of content to view from a variety of nations and genre. But as of 2019, the dataset contains TV shows and movies available on Netflix. Fixable, a third-party Netflix search engine, provided the data for this study. The purpose is to forecast clusters based on similar content by comparing text-based features, in this example, the description column, which is a brief graphic overview of the contents.

Problem Statement

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.

Approach

Our first step was to import the dataset through pandas 'read_csv' then data wrangling and feature engineering in our dataset. This dataset consists of TV shows and movies available on Netflix as of 2019.Dataset contains 7787 rows & 12 columns. It will be interesting to explore what all other insights can be obtained from the same dataset There are null values in four columns. Then we did Null value treatment and remove all the null values. After this we perform various EDA on given data set.
The idea was to use text-based variables to anticipate clusters of related content. The dataset is subjected to exploratory data analysis in order to extract insights from it, but the initial null results are ignored. In addition, using EDA's findings, some hypothesis testing was done. After that, our target variable, the description column, must be feature engineered, with NLP operations such as symbol removal, stop words, punctuation, tokenization, and vectorization using TFIDF done on it. All that was left was to discover the clusters, fit our models based on the number of clusters, and evaluate the model using evaluation metrics.

Conclusion

There are about 70% movies and 30% TV shows on Netflix.The United States has the highest number of content on Netflix by a huge margin followed by India.Raul Campos and Jan Sulter collectively have directed the most content on Netflix. Anupam Kher has acted in the highest number of films on Netflix. Drama is the most popular genre followed by comedy. More of the content is released in holiday season October, November, December and January.The number of releases have significantly increased after 2015 and have dropped in 2021 because of Covid 19.