Capstone Project Submission

Instructions:

i) Please fill in all the required information. ii) Avoid grammatical errors.

Team Member's Name:

Name: Khushboo Chaurasiya Email: Sharmakhushboo771@gmail.com

Role: -

1) Data Cleaning: -

- Dealing with null values, duplicate data

2) Exploratory Data Analysis: -

- Extracting the facts from the dataset

3) Feature Engineering: -

- Text preprocessing(Lower case, removed punctuations, remove stop words, tokenization, stemming, vectorization using TF-IFD)

Name: Abhishek Jain Email: klearpixeloff@gmail.com

4) Model Implementation: -

- Clustering: Used K-Means and hierarchical clustering.
- Find clusters using dendogram and WCSS.

5) Model performance Evaluation:-

- Evaluated both models using DBS(Davies bouldin score)

6) Conclusion:-

from elbow and sillhoute score ,optimal of 26 clusters formed , K Means is best for identification than Hierarchical as the evaluation metrics also indicates the same.in k-means cluster 0 has the highest number of data points and evenly distributed for other cluster

Netflix has 5372 movies and 2398 TV shows, there are more number movies on Netflix than TV shows.

TV-MA has the highest number of ratings for tv shows i,e adult ratings

highest number of movies released in 2017 and 2018 highest number of movies released in 2020 The number of movies on Netflix is growing significantly faster than the number of TV shows. We saw a huge increase in the number of movies and television episodes after 2015. there is a significant drop in the number of movies and television episodes produced after 2020. It appears that Netflix has focused more attention on increasing Movie content than TV Shows. Movies have increased much more dramatically than TV shows

the most content is added to Netflix from October to January

Documentaries are the top most genre in Netflix which is followed by standup comedy and Drams and international movies

kids tv is the top most TV show genre in Netflix

most of the movies have duration of between 50 to 150 highest number of tv_shows consisting of single season Those movies that have a rating of NC-17 have the longest average duration. When it comes to movies having a TV-Y rating, they have the shortest runtime on average

usa has the highest number of content on the Netflix ,followed by India

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.

Please paste the GitHub Repo link.

GitHub Link:-
https://github.com/Klearpixeloff/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING

https://github.com/Geniuskhushboo/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING

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