Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

HRITIK SHARMA
EMAIL – hritik.2.sharma@gmail.com

Contribution-

1. DATA CLEANING
2. EXPLORATERY DATA ANALYSIS
3. FEATURE ENGINEERING
4. K MEANS CLUSTERING
5. HIERARCHIAL CLUSTERING

Please paste the GitHub Repo link.

Github Link:-https://github.com/hritiksharma11/NETFLIX-TVSHOWS-AND-MOVIES-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)

The most popular media and video streaming platform is none other than Netflix. It includes over 8000+ movies with tv shows worldwide. Currently Netflix have over 200M Subscribers Globally. The most popular entertainment service used by people around the globe is also NETFLIX. It provides a huge collection of movies and TV shows which are streamed anytime by means of online services.

Netflix observed that in 2018, the number of TV shows has nearly tripled whereas the number of movies has decreased over 2,000 titles since 2010 and it will be very interesting to dig into what all other insights can be derived from the same dataset. The aim of the project is to create a model that can perform Clustering on comparable material by matching text-based attributes.

The tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.The data set contains 7787 rows and 12 columns.

The main goal of our project is to create a model that can perform Clustering on comparable material by matching text-based attributes.

According to the problem statement, the question arises that, understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that we have used K-means Clustering.
We are going to perform the data wrangling on the raw data to get the useful data without NAN values (null values) and observe the summary statistics of the dataset. We prepared a dataset with feature engineering and feature scaling and also,

dropped out the inessential columns. In the analysis, the data was converted into standard scalar form to implement the model.

We executed the exploratory data analysis. Reviewing all the data processing
then the model was trained to form Collections. In which we remark as below k-means clustering model gave insights of silhouette analysis consisting of 2,3,4,5,6 clusters.

For n_clusters = 2 The average silhouette_score is : 0.7049787496083262

For n_clusters = 3 The average silhouette_score is : 0.5882004012129721

For n_clusters = 4 The average silhouette_score is : 0.6505186632729437

For n_clusters = 5 The average silhouette_score is : 0.56376469026194

For n_clusters = 6 The average silhouette_score is : 0.4504666294372765

In the end, we plot boxplot to predict the hypothesis –

❖ After clustering, we can say that our alternative hypothesis is that the number of TV shows launched in the previous few years is not growing.
❖ Our second alternative hypothesis is the number of TVshows added to Netflix is high.

We evaluated that ,TV shows account for 2.8 percent of the total, while movies account for 97.2 percent.Netflix has added a lot more movies and TV episodes in the previous years, but the numbers are still low when compared to movies released in the last ten years.Movies are mostly watched in various countries rather than TV shows