Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Shubham Bhadouria (shubhambhadouria47@gmail.com): 1.1. Descriptive Analysis 1.1.1. Data frame description 1.1.2. Data frame shape 1.2. Dealing with missing values 1.2.1. The attribute 'director', 'cast', 'country', 'date added', 'rating' consists of missing values. The missing values of attributes 'country', 'director' and 'cast' is replaced by 'Unknown'. The missing values of attribute 'rating' is replaced by its mode values. 1.3. Data Wrangling 1.3.1 Extracting the information from date added string for further data analysis. 1.3.2 Extracting the information from director, cast, country, duration columns for problem reading 1.4 Data Visualization 1.4.1 Proportion of movies and TV shows on Netflix 1.4.2 Top 10 directors on Netflix 1.4.3 Top 10 actors on Netflix 1.4.4 Top 10 countries where the highest number of movies and Tv shows has been produced 1.4.5 Distribution of content rating on Netflix 1.4.6 Top 10 genres in which the movies and Tv shows has been released 1.4.7 Frequency of movies added on Netflix month-wise in various years 1.4.8 Length distribution of movies on Netflix 1.4.9 Variation of TV shows added on Netflix with respect to the seasons 1.4.10 Number of movies and TV shows added on Netflix over the years 1.4.11 Number of movies and TV shows listed on Netflix with specific rating 1.4.12 Target audience proportion in top 10 countries 1.5 Text Processing 1.5.1 Removing punctuations, stopwords, non ASCII characters 1.5.2 Lemmatizing, Tokenizing and Vectorisation 1.6 Unsupervised Machine learning Clustering algorithm 1.6.1 Fitting the clustered data in K means and Hierarchical clustering algorithm 1.7 Recommendation system 1.7.1 Content based recommendation system
Please paste the GitHub Repo link.
Github Link:- https://github.com/shubham-bhadouria/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)

Netflix is a streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more on thousands of internet-connected devices. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential. The Aim of this project is to form the clusters based on K mean clustering and Agglomerative clustering and build a simple recommender system.

In this project, I worked on a text clustering problem where we had to classify/group the Netflix movies and shows into certain clusters such that the shows and movies within a cluster are similar to each other and the shows and movies in different clusters are dissimilar to each other.

The dataset contained about 7787 records, and 12 attributes. I began by dealing with the dataset's missing values and doing exploratory data analysis (EDA).

It was found that Netflix hosts more movies than TV shows on its platform, and the total number of shows and movies added on Netflix is growing exponentially. Also, majority of the shows were produced in the United States, and most of the shows on Netflix were created for adults and teens.

Once obtained the required insights from the EDA, we start with Pre-processing the text data by removing the punctuation and stop words. This filtered data is passed through TF - IDF Vectorizer since we are conducting a text-based clustering and the model needs the data to be vectorized in order to predict the desired results.

It was decided to cluster the data based on the attributes: director, cast, country and description. The values in these attributes were tokenized, pre-processed, and then vectorized using TFIDF vectorizer.

Through TFIDF Vectorization, we created a total of 20000 attributes. Principal Component Analysis (PCA) has been used to handle the curse of dimensionality. 5000 components were able to capture more than 95% of variance, and hence, the number of components were restricted to 5000. We first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to be 9. This was obtained through the elbow method and Silhouette score analysis.

Then clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters came out to be 11. This was obtained after visualizing the dendrogram.

A simple content-based recommender system was built using the similarity matrix obtained after using cosine similarity. This recommender system will make 10 recommendations to the user based on the type of show they watched.