Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email, and Contribution:
1) Pruthvi Raj Pruthviraj1698@gmail.com • Data wrangling • Exploratory Data Analysis • Handling Outliers • Data Preprocessing • Data Visualizations • Training and testing model • Performed NLTK operations • Recommender System

Please paste the GitHub Repo link.

Github Link:- https://github.com/Pruthviraj009/NetFlix_movies_TvShows

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

The dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is 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 services number of movies has decreased by more than 2000 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.

Here we will perform different EDA techniques and clustering on the data to get ideal clusters for this problem. And also, we will build recommender system.

Movies is the majority here which covers around 69% and TV Shows covers the rest which is 31% of the dataset.

In the Movies section 48% of the movies are Adult rated, 31% rated as Teen, 15.9% as 7&above, and the rest as Kids.
In the TV Shows section 41% of the shows are Adult rated, 27% rated as Teen, 20% as 7&above and the rest 10% as kids Rated.

The year 2017 saw the highest content ever on Netflix for movies where 744 movies were released and the yesr 2020 saw the highest for 457 TV Shows.

Our analysis on movies duration showed that movies ranging between 90mins to 150mins were the highest with 3481 movies followed by movie duration less than 90mins with 1653 movies and finally movies with more than 150mins duration has 243 movies.

Performed Recommender systems using cosine similarity because The cosine similarity is beneficial because even if the two similar data objects are far apart by the Euclidean distance because of the size, they could still have a smaller angle between them. Smaller the angle, higher the similarity.

The recommendations after all the tuning and model selection we are getting pretty good range of recommendations and satisfied with the recommendations.