Capstone Project Submission

Team Member's Name, Email and Contribution:

Name – Baishnavee Mahato
Email. mbaishnavee@gmail.com

Data Understanding

- Data Analysis
- Data preprocessing

Data Visualization

- Bar plot
- Count plot
- Line plot
- Word Cloud
- Correlation Heatmap

Data Preprocessing

Clusters Implementation

- K Means Clustering
- Hierarchical Clustering

Conclusion

Technical Documentation

GitHub Repo link.

GitHub Link:
https://github.com/mbaishnavee05/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)

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 services 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

In this project, we are required to do
Exploratory Data Analysis
Understanding what type content is available in different countries
Is Netflix has increasingly focusing on TV rather than movies in recent years.
Clustering similar content by matching text-based features

Conclusion

1.The dataset contained about 7787 rows, and 12 columns.
2.I began by dealing with the datasets missing values and doing exploratory data analysis (EDA).
3.There are more number movies on Netflix than TV shows. It is evident that there are more movies on Netflix than TV shows. Netflix has 4673 movies, which is more than double the quantity of TV shows.
4.The number of releases has significantly increased after 2015 and have dropped in 2021 due to Covid 19.
5.TV-MA ratings are the highest in movies and TV shows, because TV-MA standing for Mature Audience Only. Because this programme is intended for adults, it may not be appropriate for children under the age of 17.
6.I used Principal Component Analysis (PCA) to handle the

curse of dimensionality. 3000 components were able to capture more than 80% of variance, and hence, the number of components were restricted to 3000.

7.I first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to be 8. This was obtained through the elbow method and Silhouette score analysis.

8.Then clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters came out to be 12. This was obtained after visualizing the dendrogram.