Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name - Rohan jagadale.
Email id - rjagadale202@gmail.com
Contribution - Whole project

Please paste the GitHub Repo link.

Github Link:-
https://github.com/Rohan20202/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 company that manages a large collection of TV shows and movies, streaming it anytime via online.

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.

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

As the problem statement says, 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 used Affinity Propagation, Agglomerative Clustering, and K-means Clustering.

In Affinity Propagation, we had 9 clusters and a Silhouette Coefficient score of 0.340.
A dendrogram was used to determine the number of clusters in Agglomerative Clustering. There were two clusters, with an average silhouette score of 0.56590662228136.The final model we used was k-means clustering, which consisted 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

1. 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.

2. Our second alternative hypothesis is the number of TV shows added to Netflix is HIGHER.

Drive link:
https://drive.google.com/drive/folders/1jr-sAQfPmkIX31YKe6ifNC2rzBiOK0cy