**Netflix-Movies-and-TV-Shows- Clustering**

This project aims to cluster the video content available on Netflix based on the company’s site data. Apart from aiding in the development of an efficient recommendation system, clustering the video content would also provide information about the type of content the company is interested in listing on its site. Thus giving an insight to content creators and filmmakers on the type of video content in demand.

**EDA Summary:**

The notable observations identified during EDA were:

Most movies streaming on the platform were released after 2010. A large portion of the TV Shows streaming on the platform was released after 2015. The year 2017 had the highest number of movie and TV show releases on the platform. Netflix began adding videos to the platform in 2008 and started aggressively adding video content in 2017. It was also found that more stand-alone movies were added as compared to TV shows. The majority of the content is rated for Mature Audiences and for audiences over 14 years old. Drama is the most produced genre in the top non-English speaking countries with exception of Japan and South Korea. Japan is the biggest producer of Anime across the platform which is also the leading genre in Japan. Romance is the most produced genre in South Korea. It is noted that Comedy was the most produced genre in English-speaking countries like the United States of America, the United Kingdom and Canada. Documentaries are predominantly produced in the United Kingdom and the United States of America Duplicates of TV shows were made corresponding to their seasons. Upon doing so, it was observed that the TV shows signed have been higher than the movies in 2016 Although, the movies signed have been higher than TV shows ever since it was prominent that the TV shows signed per year were catching up to the movies signed annually. Hence, we can conclude that it was true that Netflix had been showing more interest in TV shows as compared to movies

**Clustering Summary:**

Clusters were built on the attributes: Director, Cast, Country, Rating, Listed in (genres), Description Steps involved in preprosessing:Tokenize the corpus, Remove non-ascii characters, Convert all words to lowercase, Remove punctuation marks, Replace all numbers with its respective textual representation, Stemming and Lemmatization Text corpus was vectorized using TFIDF vectorizer, and 20000 attributes were generated. More than 80% of the variance is explained just by 4000 components. Hence to handle the curse of dimensionality, only the top 4000 components were taken, which will still be able to capture more than 80% of variance in the data.

* 3.1. K Means Clustering: The optimal number of clusters were built after visualizing the elbow curve and the Silhouette score. Highest Silhouette score is obtained for 6 clusters using k-means clustering. Hence, the number of clusters for k-means clustering was taken as 6. Silhouette score at 6 clusters: 0.0082
* 3.2. Hierarchical Clustering: Clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters were built after visualizing the dendogram. From the dendogram, at an Euclidean distance of 3.8 units, 12 clusters can be built. Hence the number of clusters were taken as 12. Algorithm: Agglomerative clustering Distance: Euclidean Linkage: Ward

**Content Based Recommender system using Cosine Similarity:**

We can build a simple content based recommender system based on the similarity of the shows. If a person has watched a show on Netflix, the recommender system must be able to recommend a list of similar shows that s/he likes. To get the similarity score of the shows, we can use cosine similarity. The similarity between two vectors (A and B) is calculated by taking the dot product of the two vectors and dividing it by the magnitude. We can simply say that the CS score of two vectors increases as the angle between them decreases.

**Conclusion**

In conclusion, tailored recommendations can be made based on information about movies and TV shows. In addition, similar models can be developed to provide valuable recommendations to consumers in other domains. It will solve for improved movie and TV-Show selection times with a considerable growth in satisfaction of the content being consumed leading to more user engagement and greater trust in Netflix recommendations.

**GitHub Link**

[**https://github.com/Abhiraj1501/Netflix-Movies-and-TV-Shows-Clustering**](https://github.com/Abhiraj1501/Netflix-Movies-and-TV-Shows-Clustering)

**Drive Link**

[**https://drive.google.com/drive/u/0/folders/1RGDufaPui50VFHtblA\_j-b7MOQzyQEJ9**](https://drive.google.com/drive/u/0/folders/1RGDufaPui50VFHtblA_j-b7MOQzyQEJ9)