Capstone Project Submission

Name: Priyanka Shinde

Email: shindepriya7709@gmail.com

Contributor: Individual Project

Github Link:https://github.com/priyankashinde-DS/Capstone_project-Netflix_Movies-TV_Shows/blob/main/NETFLIX_MOVIES_AND_TV_SHOWS_CLUSTERING.ipynb

The purpose of this project is to Clustering similar content by matching text-based features.we are going to explore the dataset and we would like to find out how long the Netflix platform takes a movie or a TV show to release on its platform, how many movies and TV shows are released in specific time frame, how many movies and TV shows are release in the recent ten years on the platform, and what were the top 10 genres that the audience of the Netflix platform liked the most. From here, we would like to apply a machine learning approach to understand the data fully.

After loading csv_file the data cleaning is the most important step.in which we treat NaN values beacause,this wouldn't be beneficial to our project since there is loss of information.Using pandas, seaborn, and matplot, I did some visualization in the second stage.After visualization in text analysis (NLP) I used stop words, removed punctuations , stemming & TF-IDF vectorizer and other functions of NLP. For clustring algrorithm I used only three feature and on this feature I use different clustring method for finding the optional cluster value(best value of n_cluster=3).

We have drawn many interesting inferences from the dataset Netflix ,here's a summary of the few of them:

1) From the dataset insights we can conclude that the most number of TV Shows released in 2020 and for Movies it is 2017.
2) On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows.
3) We observe that the content added over years that Netflix is focusing movies and TV shows (Fom 2016 data we get to know that Movies is increased by 80% and TV-shows is increased by 73% compare).
4) The end & beginnings of each year seem to be Netflix's preference for adding content.Least number of contents are added in the month of February.
5) Applied different clustering models like Kmeans, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements.
6) By applying different clustering algorithms to our dataset .we get the optimal number of cluster is equal to 3.
7) *In this dataset there are two types of contents where 30.86% includes TV shows and the remaining 69.14% carries Movies.*

8) ***TOP 3 content categories are International movies , dramas , comedies.***
9) On netflix there is much more content for a more mature audience.

The most popular actor on Netflix movie, based on the number of titles, is Anupam Kher.

The most popular actor on Netflix TV Shows based on the number of titles is Takahiro Sakurai.

DriveLink:https://drive.google.com/drive/u/0/folders/1new4PqqzhTT3JZxqzbElWCA6rxjs6Xnz