Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) Vinay V. Lanjewar
E-mail: lanjewarvinay@gmail.com

- EDA.
- Approach towards plan.
- Data preprocessing & feature engineering
- Methods to find k value.
- PPT and Technical documentation.
- Project summery template.

2) Debabrata Sahoo
E-mail: debabratas688@gmail.com

- Data analysis.
- Data Wrangling.
- Feature Engineering.
- Frame work of project.
- Model building.
- K-means clustering.

Problem definition:

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Fixable 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 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.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

Approach :

- Initially, in the 1st step imported the data set to carry out the analysis over the data set to comprehend and handle the missing values and duplicate values .
- Performed the Exploratory data analysis and tried to get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step with the help of visualization graph by getting insights from analysis.

- Data preprocessing – in this we remove the punctuation and stops words also used stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.
- We used the k-means clustering algorithm and then checked the model performance using Silhouette's coefficient and elbow method to find the number of clusters.

Conclusion :

- Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it and we have 10 features for the further implementation.
- We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies)
- By analyzing the content added over years we get to know that in recent years Netflix is focusing movies than TV shows.
- The greatest number of the movies and TV shows release in 2017 and 2020 respectively and united nation have the maximum content on Netflix
- On Netflix, Drama's genre contains the maximum content among all of the genres and the most of the content added in December month and less content in February
- By applying the silhouette score method for n range clusters on dataset we got best score which is 0.348 for 3 clusters it means content explained well on their own clusters, by using elbow method after k = 3 curve gets linear it means k = 3 will be the best cluster
- By applying different clustering algorithms to our dataset, we get the optimal number of clusters is equal to 3.

Please paste the GitHub Repo link.

GitHub Link:- https://github.com/Debabarata308/Netflix-Movies-and-TV-Shows-Clustering.git

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