Capstone Project-4 Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Contributor's Role:

Ranjit Biswal

- Introduction
- Data cleaning
- Null Value Treatment
- Exploratory Data Analysis
- Univariate
- Data cleaning & pre-processing for clustering
- Encoding the categorical data
- K means clustering
- Hierarchical Clustering
- Silhouette analysis
- Conclusion

Abhishek Kumar

- Introduction
- Data cleaning
- Null Value Treatment
- Exploratory Data Analysis
- Univariate
- Data cleaning & pre-processing for clustering
- Encoding the categorical data
- K means clustering
- Hierarchical Clustering
- Silhouette analysis
- Dash App recommendation System
- Conclusion

Suvendu dey

- Introduction
- Data cleaning
- Null Value Treatment
- Exploratory Data Analysis
- Univariate
- Data cleaning & pre-processing for clustering
- Encoding the categorical data
- K means clustering
- Hierarchical Clustering
- Silhouette analysis
- Dash App recommendation System
- Conclusion

Please paste the GitHub Repo link.

Suvendu Dey:- https://github.com/devsuvendu/Netflix-Movies-and-TV-Shows-Clustering
Abhishek Kumar:- https://github.com/abhishekkumar/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING
Ranjit Biswal:-https://github.com/Ranjitcnb/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

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from flixable 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 the details of available data and Checked for Null values and treated them. Here, we found more than 30% null values in the director's column. Then, we take appropriate action for null values according to the circumstances.

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.

Analyzing all the variables of the data set and identifying the solution for given tasks.
Performed hypothesis testing to get the insights on duration of movies and content with respect to different variables.
After doing feature engineering and finding the number of clusters, we used the k-means algorithm and then checked the model performance using Silhouette's coefficient, to identify the best fit Model.
The number of movies on Netflix is growing significantly faster than the number of TV shows. Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.

- 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. Agglomerative Clustering, and K-means Clustering.

CONCLUSION

1. In this project, we worked on a text clustering problem where in we had to classify/group the Netflix shows into certain clusters such that the shows within a cluster are similar to each other and the shows in different clusters are dissimilar to each other.

2. The dataset contained about 7787 records, and 12 attributes. We began by dealing with the dataset's missing values and doing exploratory data analysis (EDA).

3. It was found that Netflix hosts more movies than TV shows on its platform, and the total number of shows added on Netflix is growing exponentially. Also, majority of the shows were produced in the United States, and the majority of the shows on Netflix were created for adults and young adults age group.

4. Once obtained the required insights from the EDA, we start with Pre-processing the text data by removing the punctuation, and, stop words. This filtered data is passed through TF - IDF Vectorizer since we are conducting a text-based clustering and the model needs the data to be vectorized in order to predict the desired results.

5. It was decided to cluster the data based on the attributes: director, cast, country, genre, and description. The values in these attributes were tokenized, pre-processed, and then vectorized using TFIDF vectorizer.

6. Through TFIDF Vectorization, we created a total of 20000 attributes. We used Principal Component Analysis (PCA) to handle the curse of dimensionality. 4000 components were able to capture more than 80% of variance, and hence, the number of components were restricted to 4000. We first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to be 6. This was obtained through the elbow method and Silhouette score analysis.

7. 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.

8. A content based recommender system was built using the similarity matrix obtained after using cosine similarity. This recommender system will make 10 recommendations to the user based on the type of show they watched.