CAPSTONE PROJECT 4 NETFLIX MOVIES AND TV SHOWS CLUSTERING

Capstone Project Summary

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.Firstly,we imported the dataset and carried out exploratory data analysis,in which we found that there are null values in director,cast,country columns.We treated null values. In exploratory data analysis,we found distribution of TV shows and movies,content added monthly,yearly,analyzed various graphs and derived conclusions from them.We then performed feature engineering,text cleaning,and removed punctuations. We used 5 clustering algorithms namely 1.Silhouette score

2.Elbow Method

3.DBSCAN

4.Dendrogram

5.Agglomerative clustering

Principal component analysis was performed inorder to reduce the higher dimensionality which improved the silhouette coefficient to 0.34118.

Clusters are identified for each of the records in the dataset.Recommendation based on cosine similarity is done.

For n_clusters = 2, silhouette score is 0.3492523531807158

For n_clusters = 3, silhouette score is 0.37281514079850314

For n_clusters = 4, silhouette score is 0.3860345117431574

For n_clusters = 5, silhouette score is 0.3640633581237277

For n_clusters = 6, silhouette score is 0.3496010333153998

For n_clusters = 7, silhouette score is 0.3566459651876703

For n_clusters = 8, silhouette score is 0.336616745528675

For n_clusters = 9, silhouette score is 0.335838456193169

For n_clusters = 10, silhouette score is 0.3288370025153447

For n_clusters = 11, silhouette score is 0.3331184524046595

For n_clusters = 12, silhouette score is 0.3354084534086651

For n_clusters = 13, silhouette score is 0.3351674199946564

For n_clusters = 14, silhouette score is 0.33480471464041256

For n_clusters = 15, silhouette score is 0.3370362158317324

Conclusion

1.The dataset contains 7787 rows and 12 columns, cast and director columns have a lot of missing values so we dropped them and we have 10 features for the further analysis.

2.We have two types of content: movies and TV shows.

3.Netflix has 69% of its content as movies, so we can say that movies are clearly more popular on Netflix than TV shows.

4.For a mature audience, there is much more movie content than TV shows. However, for the younger audience (under the age of 17),there are more TV shows than movies.

5.Netflix has started adding content since 2014,the highest number of movies and tv shows added in the year 2019,there is consistent content addition to netflix across the year.

6.The average duration of a movie on netflix is 90 minutes.

7.With respect to available content,the United States is on the top.India is at second followed by the UK and Canada. China is not even close to the top.

8.In terms of genres, Dramas is on the top followed by Comedies and Documentaries.

9.Number of movies added to netflix is higher than that of TV shows. In 2019, netflix added 1497 movies and 656 TV shows. So there we cannot conclude that Netflix has switched focus from movies to TV shows.

10.Principal component analysis was performed inorder to reduce the higher dimensionality which improved the silhouette coefficient to 0.34118.

11.Clusters are identified for each of the records in the dataset.

12.Recommendation based on cosine similarity is done.

Team Member's Name, Email and Contribution:

Contributor Roles:

1. Mihir Kulkarni: mihirkulkarni50@gmail.com

a. Data Preparation and Cleaning

b.Analyzing statistics of the dataset

c.Handling Null Values From the dataset

d. Creating Separate Datasets for Movies and TV Shows

e.Targets based on "rating"

f.Classifying the 'rating' feature into three categories. (Kids, Teenagers, Adults)

g. Fixing the datatype

h.Handling Comma-Separated Values

I Exploratory Data Analysis and Visualization

j.TEXT CLEANING AND FEATURE ENGINEERING

k.Removing punctuations

l.STEMMING
m.Applying Different Clustering Algorithms

n.Recommendation

O. Conclusion

2. Ram Bakale: rambakale@gmail.com

a. Data Preparation and Cleaning
b. Creating Separate Datasets for Movies and TV Shows
c. Exploratory Data Analysis and Visualization
d.TEXT CLEANING AND FEATURE ENGINEERING
e.Creating a new DataFrame vocab_before_stemming
f.VECTORIZATION
g.What type of content is available in different countries?
h.Finding if netflix has increasingly focused on TV shows rather than movies in recent years.
i.Creating a new DataFrame vocab_before_stemming
j. STEMMING
k.VECTORIZATION
l.Using CountVectorizer() to count vocabulary items
m.Creating a new DataFrame vocab_after_stemming_listed_in
n. Applying Different Clustering Algorithms
o. Conclusion

Github link:-
https://github.com/mihirkulkarni50/Netflix-Movies-and-TV-shows-Clustering.git

Google Drive Link :
https://drive.google.com/drive/folders/1oRuXELoCXXSIw5rCV7pSJnCCWhRqXgaO?usp=sharing

I.