Capstone Project Submission

Team Member's Name, Email, and Contribution:
Name : Ajinkya Shingote email id : shingoteajinkya65@gmail.com Contribution: • Data Wrangling • Handling Missing and duplicate values • Exploratory Data Analysis • Performing Vectorization(TFIDF Vec)
GitHub Repo link.
GitHub Link:- ***https://github.com/Ajinkya6597/NETFLIX_MOVIES_AND_TV_SHOWS_CLUSTERING***
Please write a summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)

Team Reality Neflix Movies & TV shows Clustering

Business Problem: The main objective is to use unsupervised machine learning algorithm to cluster the given dataset into optimal number of clusters.

❖ **Understanding the problem statement**

The Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

❖ Know your Data

Dealing with a huge data set is a time-consuming part. To understand the dataset we are working on, initially, we find the important columns and the data using “head()” and “info()”. To minimize the workload and efforts we must have to distribute data and analyze the content first.

❖ Handling missing and duplicate values

Removing the unnecessary part while handling missing values and data cleaning. During the analysis, we found that there are four columns with null values. To minimize the errors, we dropped the numeric columns having null values with 0 or NaN using the “dropna()” function. The duplicate values were also removed so as to improve the performance of the model.

❖ Exploratory Data analysis

The Exploratory Data Analysis (EDA) as the name suggests, is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. It also helps to understand the relationship between the variables (if any) and it will be useful for feature engineering.

❖ Natural Language Processing (NLP) Model

For the NLP portion of this project, I first converted all plot descriptions to word vectors so they can be processed by the NLP model. Then, the similarity between all word vectors will be calculated using cosine similarity (measures the angle between two vectors, resulting in a score . Removing stopwords, punctuations and lemmatisation. After that we did vectorization using TFIDF Vectorizer and count vectorizer.
between -1 and 1, corresponding to complete opposites or perfectly similar vectors). Finally, I extracted the 5 movies or TV shows with the most similar plot description to a given movie or TV show.

❖ Agglomerative Clustering

The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.

❖ Conclusion

- Exploratory Data Analysis was done for all the attributes to study the deep insights from the given dataset.

- Analyzed various trends in Countries and the corresponding analysis was visualised to get a clear picture of the analysis.
- TV Shows or Movies? Yes, over the period of time the popularity has been moving towards N etflix series instead of Movies. We tried to analyse this with graphical representation as well on yearly basis.
- We used TFIDF Vectorizer and Sigmoid Kernel in order to recommend movies based on the s imilarities in the Textual Attributes.
- Identified 4 distinct clusters and used Interactive Visualizations to dive deeper into the clust ered data.