Netflix Movies and TV Shows Clustering

Candidate Name: Sourav Chowdhury
Email: sourav.20497@gmail.com

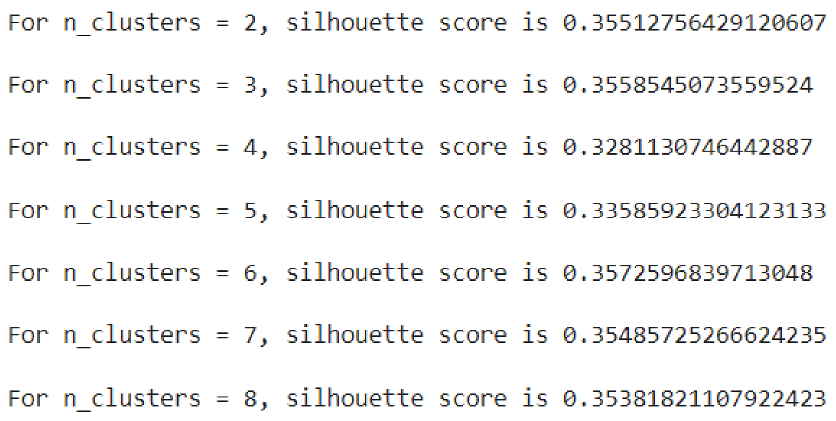
Contributor Roles:

- Importing libraries
- Exploratory Data Analysis
- Data Preparation:
  1.) Text Pre-processing
    a.) Removing Punctuations
    b.) Removing Stopwords
    c.) Stemming
    d.) Creating New Variables for Text Length
  2.) Rescaling the data
- Clustering:
  1.) Implementing K- Means Clustering
  2.) Implementing Hierarchical Clustering
- Uploading on Github
- Making Technical Document
- Presentation Slides.

Netflix Movies and TV Shows Clustering is an Unsupervised Machine learning based project for clustering the dataset into optimal groups , the model is build on the data provided collected by Flixable, a third party Netflix search engine. There are around 7787 observations in the dataset and are mostly textual features.

As the first step, perform data cleaning over the raw data and then doing Exploratory data analysis, so as to come up with conclusions based on visualizations. After that preparing the data, Once the data is prepared the machine learning algorithms were implemented, they are K-means Clustering and Hierarchical clustering. After analysis on every algorithm the different number of clusters were analyzed and an optimal number is being considered.
In the Analysis the more Focus was on Silhouette Score and the dendrogram plotted.

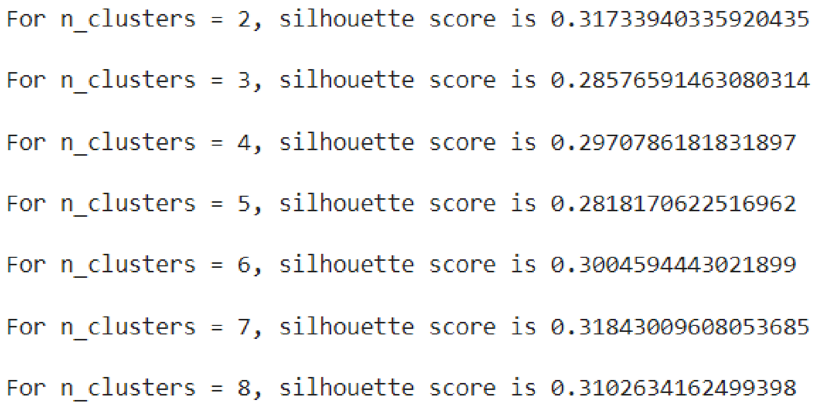
Silhouette score For K-means Clustering:



```
For n_clusters = 2, silhouette score is 0.35512756429120607

For n_clusters = 3, silhouette score is 0.3558545073559524

For n_clusters = 4, silhouette score is 0.3281130746442887

For n_clusters = 5, silhouette score is 0.33585923304123133

For n_clusters = 6, silhouette score is 0.3572596839713048

For n_clusters = 7, silhouette score is 0.35485725266624235

For n_clusters = 8, silhouette score is 0.3538182110792242З
```

Here n_clusters = 6 is giving optimum score, after which there is slight drop in score.

Silhouette score For Hierarchical Clustering:



```
For n_clusters = 2, silhouette score is 0.31733940335920435

For n_clusters = 3, silhouette score is 0.28576591463080314

For n_clusters = 4, silhouette score is 0.2970786181831897

For n_clusters = 5, silhouette score is 0.2818170622516962

For n_clusters = 6, silhouette score is 0.3004594443021899

For n_clusters = 7, silhouette score is 0.31843009608053685

For n_clusters = 8, silhouette score is 0.3102634162499398
```

Here n_clusters = 7 is giving optimum score.

Thus numbers of clusters in which the Netflix dataset can be segregated is 7.

GitHub link: https://github.com/SrvPioneer/Unsupervised-ML-Netflix-and-TV-shows-Clustering-/blob/main/NETFLIX_MOVIES_AND_TV_SHOWS_CLUSTERING.ipynb

Drive Link: https://drive.google.com/drive/folders/19S2ccuiuo8_fYivI3ipIPX4UvjbQ7-iU?usp=sharing