Netflix Movies and TV Shows Clustering

Objectives:

- Conduct Exploratory Data Analysis.
- Try understanding what type content is available in different countries.
- Check if Netflix is increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features.

Methods used:

- Descriptive Statistics.
- Data Visualization.
- Machine Learning.

Libraries utilized:

- NumPy and Pandas - For dataset cleaning and analysis.
- Matplotlib, Plotly and Seaborn - For Data Visualization.
- SkLearn and nltk - For machine learning and clustering.

Dataset used:

This dataset consists of tv shows and movies.
It is collected from a third-party Netflix search engine.

```
Attribute Information:

show_id : Unique ID for every Movie / Tv Show
type : Identifier - A Movie or TV Show
title : Title of the Movie / Tv Show
director : Director of the Movie
cast : Actors involved in the movie / show
country : Country where the movie / show was produced
date_added : Date it was added on Netflix
release_year : Actual Releaseyear of the movie / show
rating : TV Rating of the movie / show
duration : Total Duration - in minutes or number of seasons
listed_in : Genere
```

```
description: The Summary description
```

Project Overview:

Netflix, is an American subscription streaming service and production company. It was founded in 1997 by Reed Hastings and Marc Randolph in Scott's Valley, California.

It offers a library of films and television series through distribution deals as well as its own productions, known as Netflix Originals.

Our objective is to conduct an Exploratory Data Analysis to understand what content is available in different countries and if Netflix has been increasingly focusing on TV rather than movies in recent years. And use these insights to cluster similar content by matching text-based features.

After loading the data, we start by observing the first and last five values to understand the dataset. Next, we treat the null values by dropping them if the respective variables contain <1% of null values. This is followed by feature engineering to extract new variables from the datetime variable date_added.

This cleaned data is then used to conduct EDA in order to understand it better and identify the underlying trends.

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.

Finally, K–Means clustering is utilized to form 10 distinct clusters with similar data points.

Using the data provided, we also implemented a simple recommender system that successfully generates Ten similar Movies or Tv-Shows for the given title.

Contributing Member:

Name	Email
Bharath P	Bharath0924.bp@gmail.com