**Capstone Project Submission**

| Team Member’s Name, Email, and Contribution: |
| --- |
| Name: ShahFaissal I Dharwad  email id: [shahfaissal21@gmail.com](mailto:shahfaissal21@gmail.com)  Contribution:   * Feature Engineering * NLP * Model Building   Name: Ajinkya Satish Jumde  email id: jumde.ajinkya@gmail.com  Contribution:   * Data Wrangling * Handling Missing and duplicate values * Exploratory Data Analysis * Performing Vectorization (TFIDF Vec)   Name: Shreyash Movale  email id: [shreyash9m@gmail.com](mailto:shreyash9m@gmail.com)  Contribution:   * Data Wrangling * Handling Missing values * Creating a word cloud * Performing Vectorization (Count Vectorizer) * Recommender system using sigmoid kernel |
| GitHub Repo link. |
| GitHub Link:- https://github.com/shahfaissal/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 Netflix Movies & TV shows Clustering**  **Netflix is ​​a company that manages a large collection of TV shows and movies and streams them online anytime. This business is profitable because users pay to access the platform on a monthly basis. However, customers can cancel their subscription at any time. Therefore, the company must keep users on the platform and not lose their interest. This is where recommendation systems start to play an important role and providing valuable suggestions to users is essential. Working with a huge data set is the time-consuming part. To understand the dataset we are working on, we first find the important columns and data using "head ()" and "info ()". In order to minimize the workload and effort, we must first distribute the data and analyze the content. Removing the unnecessary part when handling missing values ​​and cleaning data. During the analysis, we found that there are four columns with zero values. To minimize errors, we dropped numeric columns with null values ​​with 0 or NaN using the "dropna()" function. Duplicate values ​​were also removed to improve model performance. Exploratory data analysis (EDA), as the name suggests, is used to analyze and examine datasets and summarize their main characteristics, often using data visualization techniques. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, uncover anomalies, test hypotheses, or check assumptions. It also helps to understand the relationship between variables (if any) and will be useful for feature engineering. For the NLP part of this project, we first converted all graph descriptions into word vectors for the NLP model to process. Then the similarity between all word vectors will be calculated using cosine similarity (measures the angle between two vectors, resulting in a score. Stop words removal, punctuation and lemmatization. Then we vectorized using TFIDF vector and count vectorizer.**  **between -1 and 1, corresponding to complete opposites or perfectly similar vectors). Finally, we extracted 5 movies or TV shows with the most similar plot description to that movie or TV show. Agglomerative clustering is the most common type of hierarchical clustering used to group objects into clusters based on their similarity. Next, pairs of clusters are successively merged until all clusters are merged into one large cluster containing all objects**   * **Conclusion:** * **Exploratory Data Analysis was done for all the attributes to study the deep insights from the given dataset.** * **Univariate & multivariate analysis.** * **Visualized Data, inferred insights** * **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 Netflix 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 similarities in the Textual Attributes.** * **Identified 4 distinct clusters and used Interactive Visualizations to dive deeper into the clustered data.** |