**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **S.NO** | **NAME** | **MAIL ID** | **CONTRIBUTION** | |  | 1. | **Sohail Mohammad** | [mohammadsohail825@gmail.com](mailto:mohammadsohail825@gmail.com) | Data Understanding, topic modelling , EDA , k-means cluster analysis, Agglomerative clustering | |  | 2. | **Laveti Syam Sagar** | [shyamsagar651@gmail.com](mailto:shyamsagar651@gmail.com) | Data Understanding, model implementation, EDA, NLP, Exploratory Data Analysis, | |  | 3. | **Bisetty Kodanda Nagesh** | [dpsnnagesh38@gmail.com](mailto:dpsnnagesh38@gmail.com) | Data Preparation, Data Cleaning, Gaussian Clustering, Null value Treatment | |
| **Please paste the GitHub Repo link.** |
| **Github Link:-**  https://github.com/mohammadsohail825/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING/blob/c1e318b4152966f58e77624672c11f1e45317795/NETFLIX\_MOVIES\_AND\_TV\_SHOWS\_CLUSTERING\_Project.ipynb |
| This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.  **APPROACH**  Initially, within the 1st step imported the data set to carry out the examination over the data set to comprehend the points of interest of accessible data and Checked for Null values and treated them. Here, we found more than 30% null values within the director's column. At that point, we take fitting activity for null values agreeing to the circumstances.  We Performed the Exploratory data analysis and tried to get the understanding of the data and how the content is distributed in the data set, it's type and details such as which nations are observing more and which sort of content is in demand. has been analyzed in this step with the help of visualization chart by getting insights from analysis.  **Data preprocessing –** in this we remove the punctuation and stops words also used stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.  We used the k-means clustering algorithm and then checked the model performance using Silhouette’s coefficient and elbow method to find the number of clusters.  Analysing all the factors of the data set and recognizing the solution for given tasks.  Performed hypothesis testing to get the insights on duration of movies and content with regard to diverse variables.  After doing feature engineering and finding the number of clusters, we utilized the k-means algorithm and after that checked the model performance using Silhouette’s coefficient, to identify the best fit Model.  The number of movies on Netflix is growing significantly faster than the number of TV shows. Because of covid-19, there's a significant drop in the number of movies and tv episodes produced after 2019.   * The project's primary objective is to make a model that can perform Clustering on comparable material by matching text-based attributes. * As the problem statement says, understanding what type of content is available in different countries and Is Netflix increasingly focused on TV instead of movies in recent years we need to do clustering on comparative content by matching text-based features. For that we utilized Partiality Proliferation, Agglomerative Clustering, and K-means Clustering. * In Partiality Proliferation, we had 13 clusters and a Outline Coefficient score of 0.244 * In Agglomerative Clustering. There were four clusters, with an average silhouette score of 0.17296314851287742.   The final model we used was k-means clustering, which consisted of 2,3,4,5,6 clusters.  For n\_clusters = 2 The average silhouette\_score is : 0.7049787496083262  For n\_clusters = 3 The average silhouette\_score is : 0.5882004012129721  For n\_clusters = 4 The average silhouette\_score is : 0.6505186632729437  For n\_clusters = 5 The average silhouette\_score is : 0.56376469026194  For n\_clusters = 6 The average silhouette\_score is : 0.4504666294372765 In this project, we done these work:-  * Exploratory Data Analysis * Understanding what type content is available in different countries * Checked, Netflix has increasingly focused on TV rather than movies in recent years. * Clustered similar content by matching text-based features |