**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Prajwal M V**  **Email:** [**prajwalpraju786@gmail.com**](prajwalpraju786@gmail.com) |
| **Please paste the GitHub Repo link.** |
| GitHub Link : <https://github.com/prajwal-mv/Netflix_Movies_And_TV_Shows_Clusterig> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| **Introduction:** OTT platforms have become a normal and usual mode of entertainment in today’s world. There are many popular OTT platforms like Amazon Prime, BigFlix, Arre, Discovery+, etc., competing with each other to increase their user base. Netflix is one of the leading OTT platforms, not only in India but also internationally. The success of the OTT platforms depends on two things- the variety of content and appropriate recommendations to the users. Clustering is a useful technique to achieve the best possible recommendations and increase the viewership of the platform.  **Problem Statement**: This dataset consists of TV shows and movies available on Netflix as of 2019.The dataset is collected from Fixable 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**: Here first we imported data set and performed the Exploratory data analysis and tried to get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step with the help of visualization graph 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. Finally used word cloud to visualize various clusters and also obtained recommendations for Movies and TV Shows using Cosine similarity. |
| **Conclusion:**   * It was interesting to find that majority of the content available on Netflix is Movies. * But in the recent years it has been focusing more on TV-Shows. * Most of these contents are released either in the year ending or the beginning. * United States and India are among the top 5 countries that produce all of the available content on the platform. * Also 6 of the actors among the top ten actors with maximum content are from India. * TV-MA tops the charts, indicating that mature content is more popular on Netflix. * k=10 was found to be an optimal value for clusters using which we grouped our data into 10 distinct clusters. * Using the given data, a simple recommender system was created using cosine\_similarity and recommendations for Movies and TV Shows were obtained.  **Future Scope:**  * Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings. * More time could be given into building a better recommender system, which later can be deployed on web for usage. |