Capstone Project Submission

Team Member's Name, Email and Contribution:

Contributor role :

1) **Data Wrangling -**
   i. Analyze the Data set

2) **Data cleaning-**
   i. Delete unnecessary data.
   ii. Null value treatment/Duplicate values treatment

- **Name: Sameer Satpute**
  Email: sameersatpute7@gmail.com
  Contribution:

1) **Data Visualization**
   - Countplot for rating
   - Countplot for rating w.r.t Type
   - Top 10 genres
   - Top 20 director
   - Country wise trend analysis
   - Wordcloud on genres
   - Country wise distribution
   - Distibution of TV & Movies

2) **Text Classification**
   - Count vectorizer
   - TF-IDF
   - Lower case
   - Remove stopwords
   - Remove words and digit containing digits
   - Rephrase text

3) **Algorithms**
   - principal component analysis
   - hierarchical clustering
   - K-Means clustering
   - DBSCAN

Please paste the GitHub Repo link.

GitHub Link: - https://github.com/sameersat96/Netflix-movies-and-TV-shows-clustering.git
G-Drive link - https://drive.google.com/drive/folders/1qjFTwTi-6TWl52xAz2AV7dB-jwRbqvp1?usp=sharing

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Netflix The firm was founded in 1999 and is now recognized as one of the largest international firms in this field. Netflix has maintained a competitive strategy by having a business model based upon fast delivery, no late-fees policy and a very useful return in the mail system

My First Step was to import the dataset using Pandas then data wrangling and know the features in the dataset. There are around 30% null values in director column.

The next step is EDA in that I use different feature to know insights from dataset. I plot Countplot with respect to different features in this step with help of visualization more insights we get from analysis. Data processing in this step I remove stopwards, punctuation, stemming.

I apply **K-means clustering, DBSCAN, hierarchical clustering, PCA** algorithms and also check model performance also silhouette coefficient and elbow method use for finding number of clusters

Conclusion:

- In EDA part we observed that
    i. Netflix has 69% of its content as movies, so movies are more popular on Netflix than TV shows.
    ii. United States has the most number of movies and shows followed by India and United Kingdom.
    iii. TV-MA rated content is maximum in number in the dataset. This rating indicates that the content is for mature and adult audience above the age of 17.
    iv. There is an exponential raise in the number of TV shows and movies distributed by Netflix in the recent years.
- Text cleaning and vectorization was done on the combined features of the dataset which includes origin country, leading cast member, rating type, content type and description for clustering analysis.
- Optimal number of clusters were found out to be 25 with silhouette coefficient value of 0.0279
- Principal component analysis was performed in order to reduce the higher dimensionality which improved the silhouette coefficient to 0.35. Even though there's improvement in the silhouette score, these cannot be compared as these are two different method of pre processing is involved.
- Recommendation based on cosine similarity is also done on the same transformed data.