Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team **Member's Name, Email and Contribution:**

1. **Deveshya Gupta(deveshyagupta9454@gmail.com)**

    a. Data processing
    b. Topic modeling
        - LSA(Latent Dirichlet Allocation)
        - LDA(Latent Semantic Allocation)

2. **Minal Kharbade(kharbademeenal@gmail.com)**

    a. Data cleaning
    b. EDA(Exploratory data analysis)
    c. Model implementation
        - K-means clustering

Please paste the Git-Hub Repo link.

Github Link:-https://github.com/DEVESHYA3/Netflix-movies-and-tv-shows.git

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 is a streaming service that offers a widely variety of award winning TV shows, movies, anime,documentaries, etc.In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010.

First we extracted data from Netflix dataset and then categorized it to identify, analyze behavior of data and pattern. We have done EDA on the dataset. Topic modelling and K-Means cluster help us to do the clustering of same type of content.

Some introductory features of dataset are as below:

- 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

In this project we are required to do

- Exploratory Data Analysis
- Understanding what type content is available in different countries.
- Is Netflix has increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text based features.

Netflix Movies and TV shows clustering involves various steps such as below:

- Loading the data
- Data Description
- Exploratory analysis and visualizations
- Data Processing
- Topic modelling
  1)Latent semantic analysis (LSA)
  2)Latent dirichlet allocation (LDA)
- K-Means clustering

- Conclusion

Throughout the project we learn many things such as problem statement , technical side of Topic modelling .We deal with show_id, type, title, date, release date, etc. as Netflix dataset and learn how we do topic modelling and clustering of the same type of content on the basis of their content type, genre, rating, etc.

Drive link:
https://drive.google.com/drive/folders/1RGBKHnNeb7R790-HxJzKgrAgNuUNgSze?usp=sharing