- JunaidN7

Movie-Genre-Prediction

Project Description:

This project involved the following sequential steps to analyze and predict movie genres based on their story descriptions:

1. ***Imported Data:*** The project began by importing data from CSV files.
2. ***Data Merging:*** The imported data was merged to create a comprehensive dataset.
3. ***Data Exploration:*** The dataset was explored to gain familiarity with its contents. Preliminary preprocessing was performed, and a word cloud was created for text visualization.
4. ***Rescaling and Bagging:*** The TF-IDF methodology was employed to scale the words in the story descriptions. A bag of words was created and transformed into a matrix. WordClouds were generated to visualize prominent words in the dataset.
5. ***Data Cleaning:*** Rows with null genre information and no plot were dropped. A DataFrame containing three columns - movieId, story, and DramaGenre (in binary format) - was developed.
6. ***Data Preparation:*** The data was prepared for modeling, which included confirming the shape, counting values, checking for missing data, handling duplicates, and lemmatizing text.
7. ***Data Partitioning:*** The data was divided into training and testing sets using a 85:15 ratio.
8. ***Feature Extraction:*** TF-IDF embedding was used for feature extraction in the dataset.
9. ***Model Testing:*** Nine different predictive models were initialized and tested on the training data. Logistic Regression emerged as the best-performing model based on accuracy, precision, and recall.
10. ***Model Deployment:*** The chosen Logistic Regression model was trained on the full dataset.
11. ***Movie Story Evaluation:*** The process of data merging, cleaning, and binary representation of genres was replicated for the movie evaluation dataset.
12. ***Prediction:*** The best model (Logistic Regression) was applied to predict movie genres in the evaluation dataset.
13. ***Evaluation:*** The final step involved reporting the accuracy, precision, recall, and F1 score of the model's predictions on the evaluation dataset.

In summary, this project aimed to create a predictive model for movie genres based on story descriptions, and it concluded with the deployment of the best-performing model and the evaluation of its performance on a separate evaluation dataset.