Name : Sneha k

**8. Movie Success Prediction and Sentiment Study**

**Objective: Predict movie success using IMDB/Kaggle data, and analyze sentiment of viewer reviews.**

**Tools: Python (NLTK, VADER, Sklearn), Excel**

**Mini Guide:**

**Scrape or import IMDB movie + rating data**

**Use VADER for sentiment on user reviews**

**Create regression model to predict box office success**

**Analyze genre-wise sentiment trends**

**Deliverables:**

**Python notebooks**

**Sentiment visuals**

**Predictive model summary**

**Movie Success Prediction Using Regression Techniques**

## Introduction

The movie industry constantly seeks tools to predict the success of films. With abundant data on movies, from descriptions to financial and critical metrics, it's possible to build predictive models that assess potential success. This project leverages regression modelling to forecast movie success using structured and textual features from a dataset of 1000 films.

## Abstract

This project presents a data-driven approach to predict movie success using linear regression. A dataset containing information such as movie title, description, cast, runtime, rating, revenue, and metascore was processed. Natural Language Processing (NLP) techniques were used to compute sentiment scores from movie descriptions. After cleaning and transforming the data, a regression model was trained and evaluated. The model achieved an $R^2$ score of 0.66, indicating a fair predictive capability. The outcome of the project includes an interpretable model and a performance summary exported to an Excel file.

## Tools Used

- **Programming Language**: Python
- **Libraries**: Pandas, NumPy, Scikit-learn, NLTK (VADER Sentiment Analysis)
- **IDE**: JupyterLab
- **File Format for Output**: Excel (.xlsx)

## Steps Involved in Building the Project

1. **Data Loading**: Imported a CSV file containing movie-related data.
2. **Data Cleaning**: Removed null values and infinite entries, and corrected column names (e.g., 'Aniimation' to 'Animation').
3. **Preprocessing**: Stripped whitespace, ensured proper data types, and converted year/rank to integers.
4. **Sentiment Analysis**: Applied NLTK's VADER sentiment analyzer to generate a sentiment score for each movie description.
5. **Feature Selection**: Chose features like year, runtime, rating, votes, revenue, metascore, and sentiment for the regression model.
6. **Model Training**: Used Scikit-learn's LinearRegression on a training set.
7. **Evaluation**: Assessed model with $R^2$ score and Mean Absolute Error (MAE) on a test set.
8. **Output Generation**: Saved predictions, actual values, and model coefficients to an Excel file.

## Conclusion

The project demonstrates the feasibility of using linear regression to predict movie success. With a 66% variance explained by the model, it provides valuable insights for stakeholders in the film industry. Combining structured numeric data with NLP-based sentiment analysis improves predictive power. This approach offers a foundation that can be enhanced with more advanced models or additional features in future iterations.