Movie Genre Prediction using TF-IDF and Naive Bayes

# 1. Introduction

This document explains a machine learning project that predicts movie genres based on plot summaries. The approach involves text preprocessing, TF-IDF vectorization, handling class imbalance with oversampling, and classification using the Multinomial Naive Bayes algorithm.

# 2. Code Breakdown

## 2.1 Data Loading

Three datasets are loaded:  
- `train\_data.txt`: Contains movie ID, title, genre, and plot for training.  
- `test\_data.txt`: Contains movie ID, title, and plot for testing.  
- `test\_data\_solution.txt`: Contains movie ID and the correct genre for evaluation.

## 2.2 Text Cleaning

A cleaning function is applied to the plot summaries to remove punctuation and convert text to lowercase, ensuring uniformity before feature extraction.

## 2.3 TF-IDF Vectorization

The `TfidfVectorizer` transforms text data into a matrix of TF-IDF features, limited to the top 5000 most relevant tokens.

## 2.4 Oversampling

RandomOverSampler is used to handle class imbalance by increasing the number of samples for underrepresented genres, creating a balanced dataset for training.

## 2.5 Model Training

The Multinomial Naive Bayes classifier is trained on the resampled training data.

## 2.6 Test Set Preparation

Test data is matched with the solution file based on IDs. The merged data is used for prediction and evaluation.

## 2.7 Prediction and Evaluation

The trained model predicts genres for the test data. Performance is evaluated using precision, recall, and F1-score.

# 3. Results

## 3.1 Model Evaluation

Accuracy: 47%  
  
Macro average F1-score: 0.33  
Weighted average F1-score: 0.49  
  
Genres like 'western', 'documentary', and 'music' showed relatively strong performance. Weak performance was observed in underrepresented genres such as 'biography', 'history', and 'mystery'.

## 3.2 Genre Distribution

The original training data had significant class imbalance. Most common genres:  
- Drama: 13613  
- Documentary: 13096  
- Comedy: 7447  
  
Least represented genres included:  
- War: 132  
- News: 181  
- Game-show: 194

# 4. Conclusion

The TF-IDF + Naive Bayes model provides a baseline for multi-class movie genre classification. Further improvements can be made using advanced sampling techniques, better text representation (e.g., word embeddings), and model tuning or more complex classifiers.