Assignment 2 answers:

Question 1: Which model do you think performs best on the test set?

The Logistic Regression model with TF-IDF representations performs best on the test set. This model achieved an accuracy of 0.55 (55%), precision of 0.545, recall of 0.55, and F1-score of 0.543. While Logistic Regression with Word2Vec performed similarly (0.545 accuracy), the TF-IDF version had a slight advantage in precision. Logistic Regression significantly outperformed Naive Bayes with both representations (Naive Bayes + TF-IDF: 0.38 accuracy, Naive Bayes + w2v: 0.415 accuracy). Logistic Regression performs better because it can model complex decision boundaries through weighted feature combinations, while Naive Bayes assumes feature independence, which does not hold well for text data where word co-occurrences matter. TF-IDF likely works better than Word2Vec here because it captures domain-specific vocabulary through inverse document frequency weighting, which helps distinguish genre-specific terminology in movie summaries.

Question 2: Which metric(s) did you base your decision on?

I based my decision on all four metrics (accuracy, precision, recall, and F1-score) with particular emphasis on accuracy and F1-score. The Logistic Regression + TF-IDF model achieved the highest accuracy (0.55), meaning it correctly classified the most movies overall. Its F1-score (0.543) was also the highest, which is important because F1-score balances precision and recall, showing the model performs consistently well on both false positives and false negatives. The precision of 0.545 and recall of 0.55 being nearly equal demonstrates balanced performance without favoring one type of error over another. While the differences between Logistic Regression + TF-IDF (0.55 accuracy) and Logistic Regression + w2v (0.545 accuracy) were small, the TF-IDF version had consistently stronger metrics across the board.

Question 3: Why did you decide that these metrics were most important for this task?

For movie genre classification, I prioritized a balanced evaluation approach because different metrics reveal different aspects of model quality. Accuracy is important because we want to correctly classify as many movies as possible across all four genres. However, accuracy alone can be misleading, which is why I also considered precision, recall, and F1-score. Precision matters because incorrect genre classifications could frustrate users. If someone searches for a thriller but gets recommended a romantic comedy, they lose trust in the system. Recall is equally important because missing movies of a certain genre (false negatives) means users will not discover content they would enjoy. The F1-score combines

these concerns into a single metric, making it valuable for comparing models. Since our dataset has four genres that need to be distinguished from each other, and the consequences of misclassification are similar across genres, a balanced metric like F1-score combined with overall accuracy provides the most comprehensive view of model performance.

Question 4: Briefly describe two instances where your best model predicted incorrectly.

Instance 1 (ID: 9271903, "Trois"): The model predicted this as a Romantic Comedy (3) when it was actually a Thriller (0). The movie summary describes a "young corporate tycoon who is newly married" with his "lovely and supportive wife," which contains romantic relationship keywords that likely misled the model.

Instance 2 (ID: 7134109, "White Lady"): The model predicted this as Drama (1) when it was actually a Thriller (0). The summary mentions "a simple and smart girl from the province" in a "dark and eerie campus," where the focus on character description and setting may have emphasized dramatic elements over the thriller aspects.

Question 5: What potentially might have led the model to predict them incorrectly?

The misclassifications likely occurred due to overlapping vocabulary between genres. In Instance 1, words like "married," "wife," and "lovely" are strongly associated with romantic comedies in the TF-IDF representation, even though the overall plot is a thriller. In Instance 2, the character-focused language ("simple girl," "from the province") is commonly found in dramas, causing the model to miss the thriller elements suggested by "dark and eerie." TF-IDF focuses on word frequency rather than narrative structure or plot development, so it may miss the suspenseful nature of stories when relationship or character descriptions dominate the text. The model also lacks context about how these words work together to create genre-specific atmospheres. Another factor is that the training data may not have enough examples that combine romantic elements with thriller plots, making it difficult for the model to learn these nuanced patterns.