Assignment: Spam Classification

Objective

In this assignment, you will build Logistic Regression and Naive Bayes from scratch to classify spam vs. not spam messages. You'll experiment with feature representations, regularization, and compare their performance using standard metrics.

Experiment-3: Logistic Regression (from scratch)

1. **Baseline:** Train logistic regression using your implementation with default hyperparameters (learning rate, epochs, no regularization). Report accuracy, precision, recall, F1-score, and confusion matrix.
2. Apply feature scaling so that all features have zero mean and unit variance.(Use `StandardScaler` from `sklearn.preprocessing`.)
3. **Feature Representation:** Compare the effect of using `CountVectorizer` vs. `TfidfVectorizer`
4. **Effect of Regularization:** Add L2 regularization with different strengths (λ) (experiment with 3 different value of λ . Compare performance
5. **Learning Dynamics:** Plot the loss curve across epochs.

Experiment-4: Naive Bayes (from scratch)

1. Implement the multinomial Naive Bayes algorithm with Laplace smoothing:
$$P(\text{spam} \mid d) \propto P(\text{spam}) \prod_{w \in d} P(w \mid \text{spam})$$
2. Train and evaluate Naive Bayes on the same dataset using both `CountVectorizer` and `TfidfVectorizer`.

3. Report accuracy, precision, recall, F1-score, and confusion matrix.

Part III: Comparative Analysis

1. Create a results table to summarize your findings. Example format:

Model	Vectorizer	Reg. λ	Accuracy	Precision	Recall	F1
Logistic Regression	Count	0.1	0.87	0.80	.85	.41
Logistic Regression	TF-IDF	0.1	0.91	0.68	0.55	0.61
Naive Bayes	Count	–	0.89	0.62	0.70	0.66
Naive Bayes	TF-IDF	–	0.92	0.75	0.65	0.70