EE476 Lab 1 API Reference

Quick reference for APIs used to complete the missing code cells in Lab 1.

Pandas

drop_duplicates — removes duplicated rows from a pandas DataFrame.

Scikit-Learn (sklearn)

Data preprocessing

- **CountVectorizer** — converts a collection of text documents into a matrix of token counts.
- **train_test_split** — splits arrays or matrices into random train and test subsets.

Machine learning model

- **LogisticRegression** — logistic regression (aka logit, MaxEnt) classifier.
- Use **model.fit(X, y)** to train the model, where **X** are features and **y** are labels.
- Use **model.predict(X)** to predict class labels from features.
- **model.fit_predict(X, y)** is shorthand for training and predicting in one step.

Evaluation metrics

- **classification_report** — builds a text report showing main classification metrics.
- **confusion_matrix** — computes a confusion matrix to evaluate classification accuracy.
- **accuracy_score** — accuracy classification score.