In class activity week 3

1. A. V_1 Doc 1 = [1, 1, 0, 0, 1]

V_1 Doc 2 = [0, 0, 1, 1, 1]

V_1 Doc 3 = [1, 0, 1, 0, 1]

1.A. TF-IDF

V_1 doc 1 = $[\frac{1}{3}, \frac{1}{3}, 0, 0, \frac{1}{3}]$

TF $[\frac{1}{3}, \frac{1}{3}, 0, 0, \frac{1}{3}]$ x IDF $[\ln(\frac{3}{2}), \ln(\frac{3}{1}), \ln(\frac{3}{3})]$

TFIDF = (0,13, 0,366, 0)
doc 1

free prize m $ now

(0,13, 0,366, 0, 0, 0) TFIDF Doc 1

1A. DOC 2 TF-IDF

TF

[0, 0, 1/3, 1/3, 1/3] x IDF [0, 0, ln(3/2), ln(3/1), ln(1)]

0

0.405, 1.09,

TF IDF [0, 0, 0.135, 0.366, 0] Doc 2

Connor Roave | 1/23/2026 | EE 467

1.A TF IDF doc 3

TF $[\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}]$ IDF $[\ln(\frac{3}{2}), 0, \ln(\frac{3}{1}), 0, \ln(\frac{3}{1})]$

0.405 1.09

1/3

$[0.135, 0, 0.366, 0, 0]$ TF IDF doc 3

1.B

Bow Doc1 = [1, 1, 0, 0, 0]
Bow Doc2 = [0, 0, 1, 1, 0]
Bow Doc3 = [0, 0, 0, 0, 1]

1.B TF IDF

TF $[\frac{1}{2}, \frac{1}{2}, 0, 0, 0]$ × IDF $[\ln(\frac{3}{1}), \ln(\frac{3}{1}), 0, 0, 0]$

0.5 0.5 1.09 1.09

$[0.545, 0.545, 0, 0, 0]$ Doc1 TF IDF

1.B TF IDF

$[0, 0, \frac{1}{2}, \frac{1}{2}, 0]$ × $[0, 0, \ln(\frac{3}{1}), \ln(\frac{3}{1}), 0]$

Doc 2 TF IDF $[0, 0, 0.545, 0.545, 0]$

1.B TF IDF

$[0, 0, 0, 0, \frac{1}{2}]$ × $[0, 0, 0, 0, \ln(\frac{3}{1})]$

$[0, 0, 0, 0, 0.545]$ Doc 3 TF IDF

Connor Roane | 1/23/2026 | EE 467

In class activity week 3

2. 100 spam 1000 ham

2 correct Spam and correct ham

2.A, $TP = \frac{\#\text{ spam correctly classified}}{\text{total \#emails}} = \frac{2}{1100} = TP$

~~FPZ~~

0.0018	0.089
0.0004	0.908

$\frac{98}{1100} = 0.089 = FN$

$TN \ \frac{909}{1100} = 0.908$

$\frac{1}{1100} = 0.0004 = FP$

B, Accuracy $= \frac{TP+TN}{TP+FP+TN+FN} = \frac{0.9098}{1}$

Accuracy = 0.9098

precision $= \frac{TP}{TP+FP} = \frac{0.0018}{0.0018+0.0004} = \frac{0.0018}{0.0022}$

$= .66$

recall $\frac{TP}{TP+FN} = \frac{0.0018}{0.0018+0.089} = \frac{0.0018}{0.0908}$

Recall = 0.0198

$F1 = 2 \times \frac{0.66 \times 0.0198}{0.66 + 0.0198} = \frac{0.026}{0.6798} = 0.0382$

F1 = 0.0382

Connor Roane | 1/23/2026 | EE467

2. C. The evaluation metric that best reflects performance in this situation is the F1 score. This is because it is the average of precision and recall. A high number shows the filter is both good at minimizing false positives and false negatives. This is very important for spam filter. This has a low F2 score showing it is not precise or has high recall. This means overall the positive is not accurate most of the time which is the most important part.

3. A. snippet B is technically correct

3. B. In snippet A since we vectorize then split the model learns on all the data. This makes it so that the test and train data will both show perfect F1 scores when tested on, not giving an accurate test of the model. We need to split first so the model doesn't learn on all the data and can be tested properly on parts of the dataset.