Question 1 (TFIDF)

Suppose that we are given the following three documents:

Document 1: Corgis are the best dogs
Document 2: The best movie is Pulp Fiction
Document 3: Corgis aren't in Pulp Fiction

a) Suppose that we use Document 3 as a query document. What are the euclidean distances to all other documents using bag of words? Which document would be classified as most similar?
b) Suppose that we query on a document with a single word: "Corgis". What is the TFIDF computed for each document? Which document would be classified as the most relevant?
c) Compute the TFIDF for the phrase "Corgis Pulp Fiction." For multi-word queries, you may assume that the TFIDF is the sum of the TFIDF for each word in the query. Which document would be classified as the most relevant?

Question 2 (k-Nearest Neighbors)

Suppose that we are given the following dataset :

x	y
1	3
2	4
5	6
7	7
9	10

Given a query point of 2 with $\lambda = 3$, what values would be calculated using the following kernels:

a) Box-car kernel $K(t) = 1$ if $|t| < \lambda$, 0 otherwise
b) Gaussian kernel $K(t) = \exp^{-t^2/\lambda}$

Question 3 (Precision and Recall Practice)

Recall (hehe) that precision and recall are defined:

Precision = TP/(TP + FP)
Recall = TP/(TP + FN)

a. Give a description of what the Precision metric is capturing/Why is it a useful metric?
b. Give a description of what the Recall metric is capturing/Why is it a useful metric?
c. Describe the Precision-Recall tradeoff. What would the precision and recall be of a “perfect/best” model?

II. Given the following confusion matrix below:

Actual	Predicted +	Predicted -
+	73	22
-	39	134

a. What is the precision of the model?
b. What is the recall of the model?

Question 4 (Precision and Recall : A tug of war)

To fully evaluate the effectiveness of a model, you must examine **both** precision and recall. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa. Explore this notion by looking at the following figure, which shows 30 predictions made by an email classification model. Those to the right of the classification threshold are classified as "spam", while those to the left are classified as "not spam.

(a)

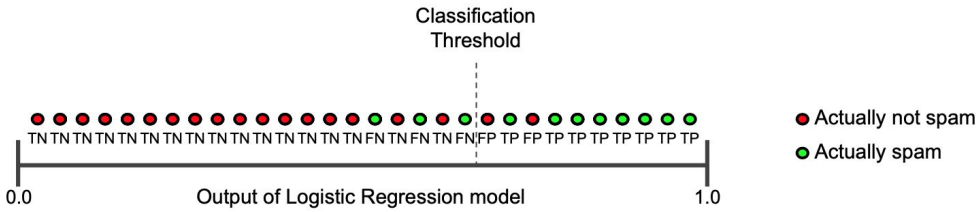

Figure 1. Classifying email messages as spam or not spam

Based on the results shown in Figure 1, what are the number of true positive, true negative, false positive, false negative? Calculate precision and recall.
Note that, precision measures the percentage of **emails flagged as spam** that were correctly classified. While, recall measures the percentage of **actual spam emails** that were correctly classified.
(b)

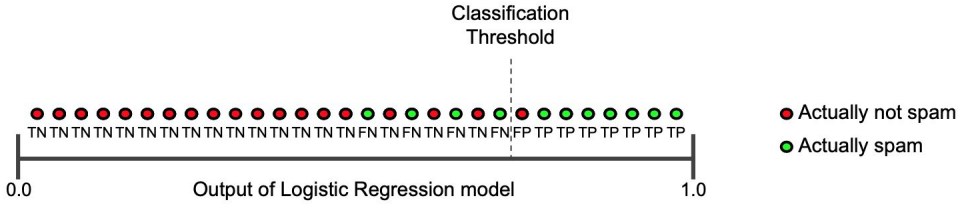

Figure 2. Increasing classification threshold

Based on the results shown in Figure 2, calculate precision and recall.
Note that, the number of false positives decreases, but false negatives increase. As a result, precision increases, while recall decreases.
(c)

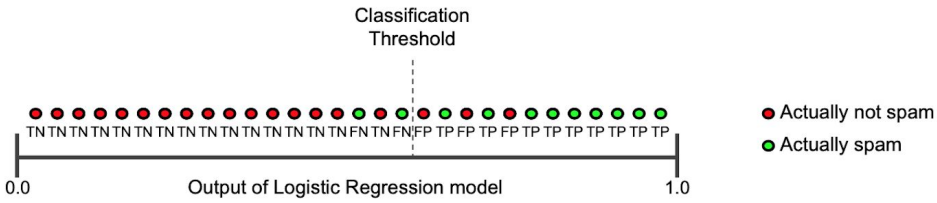

Figure 3. Decreasing classification threshold

Based on the results shown in Figure 3, calculate precision and recall.
Note that, false positives increase, and false negatives decrease. As a result, this time, precision decreases and recall increases.

Various metrics have been developed that rely on both precision and recall. For example, see F1 score: https://wikipedia.org/wiki/F1_score.