Thursday Quiz

[6 points] Consider a LSH set up with n = 100, b = 10, and r = 10. Suppose the threshold for two sets to be similar is .8.
a. [3 points] Consider two sets with a Jaccard similarity of .9. What is the error rate on these two sets given by the above LSH? Is it a false positive or false negative rate? Show your derivation.

Since the threshold for two sets is 0.8, so two sets with a Jaccard similarity 0.9 are similar.
Probability of two sets identified as a candidate pair in a single band:
t^r = 0.9^10
So prob. that two sets are not candidate pair in any band:
(1-t^r)^b = (1-0.9^10)^10
Error rate = (1-t^r)^b = (1-0.9^10)^10 = 0.013738 [2 points (1 formula and 1 final answer)]
It is a false negative rate. [1 point]

b. [3 points] Consider another two sets with a Jaccard similarity of .3. What is the error rate on these two sets given by the above LSH? Is it a false positive or false negative rate? Show your derivation.
Since the threshold for two sets is 0.8, so two sets with a Jaccard similarity 0.3 are not similar.
Probability of two sets identified as a candidate pair in a single band:
t^r = 0.3^10
So prob. that two sets being candidate pair in at least one band:
1-(1-t^r)^b = 1-(1-0.3^10)^10

error rate = 1-(1-t^r)^b = 1-(1-0.3^10)^10 = [2 points (1 formula and 1 final answer)]
It is a false positive rate. [1 point]

2) [3 points] Prove that the prob. that two signatures agree on all rows in at least one band for LSH is: 1 - (1 – s^r)^b (You also need to explain what s, r, and b are).

b – number of bands (we divide signatures into b bands)
r – r rows per band
t - the probability the minhash signatures for these documents agree in any one particular row of the signature matrix
t^r is the probability of signatures agree on all rows in one band;
(1- t^r) is the probability that they disagree on at least one row in a band; [0.5 points]
(1 – t^r)^b is the probability that they disagree on at least one row in all bands; [1 point]
So,1-(1–t^r)^b is the probability that they agree on all rows in at least one band. [1.5 points]

3. [1 point] What is the effect of following on False positive and False negative:
    a. Increasing B, keeping r constant [0.5 points]
    b. Increasing r, keeping b constant [0.5 points]

a. Decreases false negatives and Increases false positives [0.5, if both mentioned]
b. Increase False Negatives and Decreases False positives [0.5, if both mentioned]