Tuesday Quiz

Suppose that two sets are considered to be similar if their Jaccard similarity is greater than or equal to 0.6. Consider two sets S1 and S2. Suppose that their actual Jaccard similarity is 0.8. Consider their minhash signatures S1' and S2', each having 100 minhash values. Suppose the signatures are divided into 20 bands with 5 rows in each band. That is, b = 20, r = 5. Locality-sensitive hashing (LSH) is then applied to the signatures to obtain candidate pairs of sets.

1. [3 points] What is the probability that S1 and S2 are not identified as a candidate pair (i.e., false negative rate)?
Probability that S1 and S2 are identified as a candidate pair in a single band = s^r = 0.8^5= 0.32768
Probability that S1 and S2 are not identified as a candidate pair in a single band = (1-s^r)
Probability that S1 and S2 are not candidate pair in any bands =>
False Negative rate = (1-s^r)^b
(1-0.8^5)^20 = 3.56X10-4

2. [2 points] Give the formula for computing the predicted threshold. Compute the predicted threshold when b = 20 and r = 5.
t=(1-(1-p)^1/b)^1/r where b=20, r=5 and p=0.5
t=(1-(0.5)^1/20)^1/5 = 0.5087 ~ 0.51
OR
Formula for predicted threshold = (1/b) ^ (1/r)
Predicted threshold = (1/b) ^ (1/r) = (1/20) ^ (1/5) = 0.549

Now let's set b = 10, r = 10, and perform LSH again.
3. [3 points] Compute the new predicted threshold [1 point]. Can you predict if the false negative rate will go up or down, by comparing the new predicted threshold with one in question (2) [2 points]?
t=(1-(1-p)^1/b)^1/r where b=10, r=10 and p=0.5
t=(1-(0.5)^1/10)^1/10 = 0.7631 ~ 0.76
OR
New Predicted threshold = (1/b) ^ (1/r) = (1/10) ^(1/10) = 0.794

Since the predicted threshold increased, the false negative rate will increase. As threshold defines how similar documents are to be considered as a similar pair, so when it increases, the number of documents which will not be identified as pair would also increase, leading to increase in false negative.

4. [2 points] Now compute the false negative rate, when b = 10, and r = 10. Does it go up or down, compared to ones when b = 20 and r = 5? Do you reach the same conclusion as question (3)?

Probability that S1 and S2 are identified as a candidate pair in a single band = s^r = 0.810= 0.10737

Probability that S1 and S2 are not candidate pair in any bands =>

False Negative rate = (1-s^r)^b

(1-0.8^10)^10 = 0.32114

Since 0.32 > > 3.56X10-4, what we predicted is right. The false negative rate increases.