Name: _____ USC ID: _____

# Quiz 6: LSH – 2 (10 points), **15 minutes**

TA handling this quiz: Yiwen Tang yiwentan@usc.edu

1. [6 points] Consider a LSH set up with n = 100, b = 20, and r = 5. 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:

      $s^r = 0.9^5 = 0.59049$ [1 point]

      So prob. that two sets are not candidate pair in any band:

      $(1-s^r)^b = (1-0.9^5)^{20} = 1.76e-8$

      error rate = $(1-s^r)^b = (1-0.9^5)^{20} = 1.76e-8$ [1 point]

      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:

      $s^r = 0.3^5 = 0.00243$ [1 point]

      So prob. that two sets being candidate pair in at least one band:

      $1-(1-s^r)^b = 1-(1-0.3^5)^{20} = 0.0474$

      error rate = $1-(1-s^r)^b = 1-(1-0.3^5)^{20} = 0.0474$ [1 point]

      It is a false positive rate. [1 point]

2. [4 points] Describe the procedure of finding similar documents using LSH. Make sure you also state the formula for computing predicted threshold.
   1. Construct k-shingles, turn them into integers [0.5 point]
   2. Build minhash signatures of length n [0.5 point]
   3. Choose b and r (s.t., br = n) to adjust (predicted) threshold t' [0.5 point]
      set t' to the value where p = 0.5
      $p = 1-(1-s^r)^b \Rightarrow t' = (1-(1-p)^{1/b})^{1/r}$ [1 point]
   4. Construct candidate pairs [0.5 point]
   5. Examine signatures of candidate pairs to see if the fraction of their common values ≥ _t' [0.5 point]
   6. May check if documents are indeed similar, when their signatures are similar [0.5 point]