## Take Home Quiz

1. [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 => 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]

2. [1 Point] Suppose we want to find similar items and we do so by min hashing the set 10 times and then applying LSH with 5 bands and 2 rows each. If two sets have Jaccard Similarity 0.5, what is the probability that they will be identified by LSH as candidate pairs? (^ represents power/exponential)
   A.) (1-0.5^2) ^5
   B.) (1-(1-0.5^2) ^5) ^10
   C.) 1-0.5^2
   D.) None of the above
   Ans.) D
3. [2 points] 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 25 bands with 4 rows in each band. That is, b = 25, r = 4. Locality-sensitive hashing (LSH) is then applied to the signatures to obtain candidate pairs of sets.

   What is the probability that S1 and S2 are not identified as a candidate pair (i.e., false negative rate)?
   s^r = 0.8^4 = 0.4096
   (1-s^r)^b = (1-0.4096)^25
   =0.5904^25
   =0.00019%

4. [3 points] When we perform fingerprint matching with LSH, there are 2 kinds of problems that we discussed in class. Describe these 2 problems and how the technique of LSH is applied.

   (1) Many to many problem: take an entire database of fingerprints and identify if there are any pairs that represent the same individual
   1. Define a locality-sensitive family of hash functions:
   Each function f in the family F is defined by 3 grid squares
   Function f says "yes" for two fingerprints if both have minutiae in all three grid squares, otherwise, f says "no"
   "Yes" means the two fingerprints are candidate pairs !
   2. Sort of "bucketization"
   Each set of three points creates one bucket
   Function f sends fingerprints to its bucket that have minute in all three grid points of f
   3. Compare all fingerprints in each of the buckets.

   (2) Many to one problem: A fingerprint has been found at a crime scene, and we want to compare it with all fingerprints in a large database to see if there is a match
   1. Could use many functions f from family F
   2. Precompute their buckets of fingerprints to which they answer "yes" on the large database! 3. For a new fingerprint:
   Determine which buckets it belongs to
   Compare it with all fingerprints found in any of those buckets.