Take home quiz

1. [1 point] When we perform minhashing, explain why Probability of h(C1) = h(C2) is the same as Sim(C1,C2)?
   The minhash value of any column is the number of first row, in permuted order, in which column has a 1. We look down the permuted columns C1 and C2 until we see a 1. If it's a type-a row, then h (C1) = h (C2). If a type-b or type-c row, then not. We don't count the type-d rows.

   Sim(C1, C2) for both Jaccard and Minhash are a / (a +b +c).

   Then SIM(C1, C2) = a/(a + b + c). The reason is that a is the size of C1∩C2 and a+b+c is the size of C1 ∪ C2.
   Now, consider the probability that h(C1) = h(C2). If we imagine the rows permuted randomly, and we proceed from the top, the probability that we shall meet a type-a row before we meet a type-b row or type-c row is a/(a+b+c). But if the first row from the top is a type-a row, then surely h(C1) = h(C2). On the other hand, if the first row that we meet is a type-b row or type-c row, then the set with a 1 gets that row as its minhash value. However the set with a 0 in that row surely gets some row further down the permuted list. Thus, we know h(C1) != h(C2) if we first meet a type-b row or type-c row. We conclude the probability that h(C1) = h(C2) is a/(a + b+ c), which is also the Jaccard similarity of C1 and C2.

2. [5 points] Consider the following characteristic matrix of two sets: S1 and S2.

Row #	S1	S2
0	1	0
1	1	1
2	1	1
3	0	1
4	1	1
5	1	0
6	0	1

   What are the minhash values of S1 and S2 based on the permutation using h1(x) = (x + 1) mod 7 (2 pt) and h2(x) = (x + 4) mod 7 (2 pt).

   Construct a signature for S1 and S2 based on the minhash values obtained from h1(x) and h2(x) above. Estimate the Jaccard similarity of S1 and S2 using the signature. What is the actual Jaccard similarity of S1 and S2 ? Is the estimate close to the actual Jaccard similarity? If not, suggest a way to improve the estimate.

Answer- [2 points]

Final	S1	S2
h1(x)	1	0
h2(x)	1	0

Estimated Jaccard Similarity (S1,S2) = 0 (as no elements in the intersection) [1 point]

Actual Jaccard Similarity (S1,S2) = Intersection/Union = 3/7 [1 point]

No, the estimated Jaccard Similarity is not close to the Actual Jaccard similarity. This can be improved if we use a greater number of hash functions to generate more signatures with different permutations. [1 point]

3. [1 point] Explain why the Apriori algorithm cannot be used to calculate similar documents.
   Apriori outputs frequent itemsets that have support greater than threshold. This does not account for the fact that 2 itemsets although they are not frequent, yet they can be very similar.
4. [1 point] What are the conditions for an itemset to be in the negative border?
   Item is not frequent, but all of its subsets are frequent
5. [2 points] How will you deal with False positives and false negatives in Random Sampling?
   Answer-
   False positives:
   1. **Eliminate False positives:**
      ! Make a second pass through the full dataset
      ! Count all itemsets that were identified as frequent in the sample
      ! Verify that the candidate pairs are truly frequent in entire data set
      " But this doesn't eliminate false negatives
      ! Itemsets that are frequent in the whole but not in the sample
      ! Remain undiscovered

   " Reduce false negatives
      ! Before, we used threshold ps where p is the sampling fraction
      ! Reduce this threshold: e.g., 0.9ps
      ! More itemsets of each size have to be counted
      ! If memory allows: requires more space
      ! Smaller threshold helps catch more truly frequent itemsets