Tuesday Quiz

1. [2 points] For random sampling algorithm, briefly explain:

   a) How to pick samples from the basket file?(0.5 points)

   If probability is p. i.e. the probability of choosing a sample is p. Then each sample is chosen with a probability of p. Accordingly, the threshold for the samples are set. From each sample, itemsets frequent in that sample are chosen.

   b) How to reduce false positives?(0.75 points)

   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

   c) How to reduce false negatives?(0.75 points)

   2. 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

2. [1 point] Consider Set1 = {0,0,0,1} and Set2 = {1,0,2,0}. What is the jaccard distance between the two sets? 3/4

   2nd item is common to both. So jaccard similarity = ¼. Thus jaccard distance = ¾.
   A couple of you have stated some assumptions regarding the Sets, if your answer is in accordance with the assumption and the assumption is valid, you have been given marks.
   0.5 points if you have written Jaccard similarity

   This is a slightly open ended question, if you still feel what you have done is correct, you can drop by in the office hours and we can discuss.

   Note: Similarity between the 2 Sets is NOT 0. Because 2 is an element of Set2, it clearly means that these are not bits. Given the fact that they are not bits, 0 occurring as the 2nd element of both Sets, means the intersection is 1, and union is 4, hence similarity is 0.25 and distance is 0.75

3. [1 point] The SON algorithm produces false positives? False

4. [2 points] Suppose we have items {A, B, C, D, E} and we have found the following itemsets to be frequent in the sample: {A}, {B}, {C}, {D}, {B,C}, {C,D}. What all sets will belong to the negative border?
   {E}, {A,B}, {A,C}, {A,D}, {B,D}
   -0.5 for every itemset missed
   -0.25 for every extra itemset in the answer that does not belong to the negative border.
5. [4 points] In one pass (from the top row to bottom row), generate the minhash signature for each set S. There are two minhash functions. The first minhash function is x + 3 mod 5, the second one is 3x + 1 mod 5. You need to show the intermediate step. Compute the Jaccard similarity and Estimated similarity between (S1, S3), (S1, S3), and (S3, S2).

Row	S1	S2	S3
0	0	1	1
1	1	1	0
2	0	0	1
3	1	1	0
4	0	1	1

Row	S1	S2	S3	x+3 mod 5	3x+2 mod 5
0	0	1	1	3	1
1	1	1	0	4	4
2	0	0	1	0	2
3	1	1	1	1	0
4	0	0	1	2	3

Final signature set: [2 points]

H	S1	S2	S3
h1	1	1	0
h2	0	0	1

Actual Similarity [1 point]
Jaccard Sim(S1, S3) = 0/5 = 0
Jaccard Sim(S3, S2) = 2/5
Jaccard Sim(S1, S2) = 2/4

Estimated [1 point]
Estimated Sim(S1, S3) = 0/2 = 0
Estimated Sim(S3, S2) = 0/2 = 0
Estimated Sim(S1, S2) = 2/2 = 1