Thursday Quiz

1. [2 points] In the SON algorithm, please write down the input/output for the Map/Reduce functions in both phases, if we use a two-phase Map/Reduce. It is okay to describe your answer in English.

   Phase 1 Map Input: Phase 1 Map Output:
   Phase 1 Reduce Input: Phase 1 Reduce Output:
   Phase 2 Map Input: Phase 2 Map Output:
   Phase 2 Reduce Input: Phase 2 Reduce Output:

   Answer-
   Phase
   PHASE 1:
   Map Task 1:
   Input: (k, v) where k is the chunk Id and value is the input data of that chunk
   Output : (F,1) where F is the candidate itemset (support for that item is greater than the (support threshold/Total number of chunks)
   Reduce task 1:
   Input: (F, list of 1's) where F is the frequent itemset
   Output : Candidate itemsets
   PHASE 2:
   Map Task 1:
   Input: (k, v) where k is the chunk Id and value is the input data and the output from phase 1 reducer which is the candidate itemsets
   Output : (F, Count) where F is a Candidate itemset from output of phase 1 and count is the count of that itemset is in that chunk
   Reduce task 1: (Removes duplicates)
   Input: (F, list of Counts) where F is the frequent itemset and list of count represent its count in all the chunks
   Output : (F, Count) where F is the frequent itemset and Count is the sum of all the counts in the list of input, If the count >= support threshold ,It is a frequent itemset.

2. [2 Points] Answer the following questions:
   a. How many 2-shingles does Humpty have?
   b. How many 2-Shingles does "Dumtpy" have?
   c. What is the Jaccard Distance between the two?
   a. {"Hu", "um", "mp", "pt", "ty"} [0.5 point]
   b. {"Du", "um", "mt", "tp", "py"} [0.5 point]
   c. Intersection = 1. Union = 9
   Similarity = 1/9; Distance = 1 - 1/9 [1 point]
   -1 if you have read the input string wrong, but have the right answer for that string.
   -0.5 if you have given Jaccard Similarity.

Appropriate marks have been given if your shingles have included double quotes, but they weren't necessary.

3. [1 point] The Toivonen algorithm will always produce an answer? False
4. [5 points] In one pass (from the top row to bottom row), generate the minhash signature for each set S. There are three minhash functions. The first minhash function is x + 1 mod 5, the second one is 3x + 1 mod 5, and the third one is x+2 mod 5. Compute the Jaccard similarity and Estimated similarity between (S1, S3), (S1, S4), and (S3, S4).

Row	S1	S2	S3	S4
0	1	0	0	1
1	0	0	1	0
2	0	1	0	1
3	1	0	1	1
4	0	0	1	0

Row	S1	S2	S3	S4	x+1 mod 5	3x+1 mod 5	x+2 mod 5
0	1	0	0	1	1	1	2
1	0	0	1	0	2	4	3
2	0	1	0	1	3	2	4
3	1	0	1	1	4	0	0
4	0	0	1	0	0	3	1

Minhash values using h1() h2() and h3():

S1 S2 S3 S4
h1 ∞ ∞ ∞ ∞
h2 ∞ ∞ ∞ ∞
h3 ∞ ∞ ∞ ∞

S1 S2 S3 S4
h1 1 ∞ ∞ 1
h2 1 ∞ ∞ 1
h3 2 ∞ ∞ 2

S1 S2 S3 S4
h1 1 ∞ 2 1
h2 1 ∞ 4 1
h3 2 ∞ 3 2

S1 S2 S3 S4
h1 1 3 2 1
h2 1 2 4 1
h3 2 4 3 2

	S1	S2	S3	S4
h1	1	3	2	1
h2	0	2	0	0
h3	0	4	0	0

	S1	S2	S3	S4
h1	1	3	0	1
h2	0	2	0	0
h3	0	4	0	0

(Represented by new row numbers)

Computation of actual Jaccard similarities and estimated similarities:

	Jaccard Similarity	Estimated Similarity
(S1,S3)	1/4	2/3
(S1,S4)	2/3	1
(S3,S4)	1/5	2/3

Rubrics:

Signature matrix [3 points]

1 point for each row.

Each row has 4 numbers. So 0.25 points for each correct entry

Actual Jaccard similarity [1 point]

Estimated Jaccard similarity [1 point]

Note that the estimated similarities is NOT

(S1,S3)	0
(S1,S4)	1
(S3,S4)	0

In the signature matrix, these are not bits, but row numbers. Hence even if two entries are both 0, they contribute positively to the similarity. This is in contrast to the case when 1 and 0 represent bits. In this case 0 - 0 would not contribute to the similarity.