Quiz4:Similar Items

1) Give two sets A(1, 2, 3) and B(2, 3, 4), what is the Jaccard similarity between them? (1pt) What is the Jaccard bag similarity of A(1, 1, 2, 2) and B(1, 1, 1, 2, 2, 3)? (1pt)

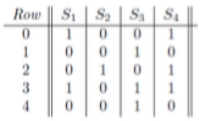
1.Jaccard(A,B) = |A∩B|/|A∪B| = 2/4 = 0.5

2.Jaccard-bag-similarity(A,B) = 4/10 = 0.4

The intersection counts 1 twice and 2 twice, so its size is 4. The size of the union of two bags is always the sum of the sizes of the two bags, or 10 in this case.

2) What are the 2-shingles for "abcdabd"? (1pt)

The (k = 2)-shingles of "abcdabd" are {ab, bc, cd, da, bd}



Row	S_1	S_2	S_3	S_4
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

3) 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. You need to show the intermediate step (3 pts) Compute the Jaccard similarity and Estimated similarity between (S1, S3), (S1, S4), and (S3, S4) (1 pts)

Generate row numbers with three minhash functions: (1pt)

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(): (2pts)

	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: (1pt)

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

4) Proof that the prob. that two signatures agree on all rows in at least one band for LSH is: $1 - (1 - s^r)^b$ (You also need to explain what s, r, and b are). (3 pts)

b – number of bands (we divide signatures into b bands)
r – r rows per band
s - the probability the minhash signatures for these documents agree in any one particular row of the signature matrix
s^r is the probability of signatures agree on all rows in one band;
$(1- s^r)$ is the probability that they disagree on at least one row in a band;
$(1 - s^r)^b$ is the probability that they disagree on at least one row in all bands;
So,$1-(1-s^r)^b$ is the probability that they agree on all rows in at least one band.
(explain what s, r, and b are: 0.5pt for each; proof – 1.5pts)