Locality-Sensitive Hashing

Question 1:

Here is a matrix representing the signatures of seven columns, C1 through C7.

C1	C2	C3	C4	C5	C6	C7
1	2	1	1	2	5	4
2	3	4	2	3	2	2
3	1	2	3	1	3	2
4	1	3	1	2	4	4
5	2	5	1	1	5	1
6	1	6	4	1	1	4

Suppose we use locality-sensitive hashing with three bands of two rows each. Assume there are enough buckets available that the hash function for each band can be the identity function (i.e., columns hash to the same bucket if and only if they are identical in the band). Find all the candidate pairs.

Give that locality sensitive hashing is done with three of two rows each.

Here, it is also given that the hashing function for each band can be the identity function. (i.e., columns hash to the same bucket if and only if they are identical in the band)

Therefore, according to the above rules that candidate pair for each band are:

Band 1: (C1, C4), (C2, C5)

Band 2: (C1, C6)

Band 3: (C1, C3), (C4, C7)

Therefore the candidate pairs are:

(C1, C4), (C2, C5), (C1, C6), (C1, C3), (C4, C7)

Question 2:

Suppose we have computed signatures for a number of columns, and each signature consists of 24 integers, arranged as a column of 24 rows. There are N pairs of signatures that are 50% similar (i.e., they agree in half of the rows). There are M pairs that are 20% similar, and all other pairs (an unknown number) are 0% similar.

We can try to find 50%-similar pairs by using Locality-Sensitive Hashing (LSH), and we can do so by choosing bands of 1, 2, 3, 4, 6, 8, 12, or 24 rows. Calculate approximately, in terms of N and M, the number of false positive and the number of false negatives, for each choice for the number of rows. Then, suppose that we assign equal cost to false positives and false negatives (an atypical assumption). Which number of rows would you choose if M:N were in each of the following ratios: 1:1, 10:1, 100:1, and 1000:1?

Give that

N = 50

M = 20

S = 20

FP = 1-$(1 - s^r)^b$

FN = $(1 - s^r)^b$

r	b	FP	FN
1	24	0.7752	0.224
2	12	0.3872	0.612
3	8	0.0622	0.937
4	6	0.095	0.9905
6	4	0.0002	0.997
8	3	0.0799	0.99992
12	2	0.1719	0.99979
14	1	0	1

Question 3:

Find the set of 2-shingles for the "document":

ABRACADABRA

and also for the "document":

BRICABRAC

Answer the following questions:

1. How many 2-shingles does ABRACADABRA have?
2. How many 2-shingles does BRICABRAC have?
3. How many 2-shingles do they have in common?
4. What is the Jaccard similarity between the two documents"?

Set of 2-shingles for “ABRACADABRA” : {AB, BR, RA, AC, CA, AD, DA}

Set of 2-shingles for “BRICABRAC” : {BR, RI, IC, CA, AB, RA, AC}

Common set of 2-shingles : {BR, CA, AB, RA, AC}

Union of shingles : {AB, BR, RA, AC, CA, AD, DA, RI, IC}

Jaccard similarity between the two documents:

Number of shingles common to both the documents = 5

Total number of shingles present in both the documents = 9

Jaccard Similarity = (Number of common shingles / Total number of shingles)

= 5 / 9

= 0.556

Question 4:

Consider the following matrix:

	C1	C2	C3	C4
R1	0	1	1	0
R2	1	0	1	1
R3	0	1	0	1
R4	0	0	1	0
R5	1	0	1	0
R6	0	1	0	0

Compute the Jaccard similarity between each pair of columns.

Jaccard Similarity: $\frac{Number\ of\ items\ common\ to\ both\ the\ columns}{Union\ of\ all\ the\ items\ (Excluding\ 0s)}$

1. Jaccard Similarity between C1 and C2 = 0 / 5 = 0
2. Jaccard Similarity between C1 and C3 = 2 / 4 = 0.5
3. Jaccard Similarity between C1 and C4 = 1 / 3 = 0.33
4. Jaccard Similarity between C2 and C3 = 1 / 6 = 0.167
5. Jaccard Similarity between C2 and C4 = 1 / 4 = 0.25
6. Jaccard Similarity between C3 and C4 = 1 / 5 = 0.2

Question 5: Consider the following matrix:

	C1	C2	C3	C4
R1	0	1	1	0
R2	1	0	1	1
R3	0	1	0	1
R4	0	0	1	0
R5	1	0	1	0
R6	0	1	0	0

Perform a minhashing of the data, with the order of rows: R4, R6, R1, R3, R5, R2.

Note: we give the minhash value in terms of the original name of the row, rather than the order of the row in the permutation. These two schemes are equivalent, since we only care whether hash values for two columns are equal, not what their actual values are.

Given Matrix

		C1	C2	C3	C4
R4	R1	0	1	1	0
R6	R2	1	0	1	1
R1	R3	0	1	0	1
R3	R4	0	0	1	0
R5	R5	1	0	1	0
R2	R6	0	1	0	0

Signature Matrix after Min-hashing:
Updated matrix will be

C1	C2	C3	C4	
0	1	1	0	3
1	0	1	1	6
0	1	0	1	4
0	0	1	0	1
1	0	1	0	5
0	1	0	0	2

Signature Matrix:

C1	C2	C3	C4
5	2	1	4

The Min hash values for C1, C2, C3 and C4 according to the R4, R6, R1, R3, R5 and R2 are **R5, R6, R4, R3**.