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.

SOLUTION

If the value in row1, column1 matches row1, column2 AND row2, column1 matches row2, column2, then we have a candidate pair.

The candidate pairs are:

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

In Band1→C1 and C4, C2 and C5

In Band2→C1 and C6

In Band3→C4 and C7, C1 and C3

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?

SOLUTION

N pairs of signatures that are 50% similar. M pairs of signatures that are 20% similar.

If M: N = 1: 1, the number of rows = 3

If M: N = 10: 1, the number of rows = 3

If M: N = 100: 1, the number of rows = 1

If M: N = 1000: 1, the number of rows = 1

QUESTION 3:

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

ABRACADABRA

SOLUTION:

AB, BR, RA, AC, CA, AD, DA, AB, BR, RA

AB, BR, RA, AC, CA, AD, DA

and also for the "document":

BRICABRAC

SOLUTION:

BR, RI, IC, CA, AB, BR, RA, AC

BR, RI, IC, CA, AB, RA, AC

Answer the following questions:

1. How many 2-shingles does ABRACADABRA have?
   Answer: 7

2. How many 2-shingles does BRICABRAC have?
   Answer: 7
3. How many 2-shingles do they have in common?
   Answer: 5

4. What is the Jaccard similarity between the two documents"?

   Answer: 5/9 (5 intersection and 9 union)

   J(A,B) = |A∩B| / |A∪B|
   =5/9

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.

SOLUTION

sim(c1,c2)= 0/5 = 0

sim(c1,c3) = 2/4 = 0.5

sim(c1,c4) = 1/3 = 0.33

sim(c2,c3)= 1/6 = 0.17

sim(c2,c4) = ¼ = 0.25

sim(c3,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.

SOLUTION:

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

Minhash(C1) = R5

Minhash(C2) = R6

Minhash(C3) = R4

Minhash(C4) = R3