Week 2: LSH (Basic)

Help Center

Warning: The hard deadline has passed. You can attempt it, but **you will not get credit for it**. You are welcome to try it as a learning exercise.

□ **In accordance with the Coursera Honor Code, I (Manuel Bordés Rguez.) certify that the answers here are my own work.**

Question 1

The edit distance is the minimum number of character insertions and character deletions required to turn one string into another. Compute the edit distance between each pair of the strings he, she, his, and hers. Then, identify which of the following is a true statement about the number of pairs at a certain edit distance.

○ There is 1 pair at distance 5.

○ There are 3 pairs at distance 4.

○ There is 1 pair at distance 1.

○ There are 3 pairs at distance 1.

Question 2

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. Which of the following is the correct minhash value of the stated column? **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.

○ The minhash value for C1 is R5

○ The minhash value for C1 is R6

○ The minhash value for C2 is R3

○ The minhash value for C3 is R5

Question 3

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, and then identify one of them in the list below.

○ C1 and C7

○ C2 and C7

○ C1 and C2

○ C4 and C7

Question 4

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"?

Then, find the true statement in the list below.

○ There are 4 shingles in common.

○ ABRACADABRA has 9 2-shingles.

○ ABRACADABRA has 7 2-shingles.

○ BRICABRAC has 8 2-shingles.

Question 5

DO NOT ANSWER THIS QUESTION. iT COUNTS ZERO POINTS AND WILL APPEAR IN A LATER HOMEWORK WHERE IT BELONGS.

Here are eight strings that represent sets:

s_1 = abcef

s_2 = acdeg

s_3 = bcdefg

s_4 = adfg

s_5 = bcdfgh

s_6 = bceg

s_7 = cdfg

s_8 = abcd

Suppose our upper limit on Jaccard distance is 0.2, and we use the indexing scheme of Section 3.9.4 based on symbols appearing in the prefix (no position or length information). For each of s_1, s_3, and s_6, determine how many *other* strings that string will be compared with, if it is used as the probe string. Then, identify the true count from the list below.

○ s3 is compared with 2 other strings.

○ s1 is compared with 5 other strings.

○ s1 is compared with 6 other strings.

○ s3 is compared with 6 other strings.

Question 6

Suppose we want to assign points to whichever of the points (0,0) or (100,40) is nearer. Depending on whether we use the L_1 or L_2 norm, a point (x,y) could be clustered with a different one of these two points. For this problem, you should work out the conditions under which a point will be assigned to (0,0) when the L_1 norm is used, but assigned to (100,40) when the L_2 norm is used. Identify one of those points from the list below.

○ (53,15)

○ (53,10)

○ (50,18)

○ (66,5)

□ **In accordance with the Coursera Honor Code, I (Manuel Bordés Rguez.) certify that the answers here are my own work.**

Submit Answers Save Answers

You cannot submit your work until you agree to the Honor Code. Thanks!