Feedback — Week 2A: LSH (Basic)

Help Center

You submitted this quiz on **Thu 12 Feb 2015 9:34 PM PST**. You got a score of **5.00** out of **5.00**.

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.

Your Answer		Score	Explanation
○ There are 2 pairs at distance 3.			
○ There are 2 pairs at distance 2.			
○ There is 1 pair at distance 3.			
◉ There are 3 pairs at distance 3.	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

We need to calculate the edit distance between each of the six pairs of words. Consider d(he,she), an easy case. You can convert "he" into "she" by one edit: insert "s" at the beginning. Alternateively, convert "she" into "he" by the single edit of deleting the first character. Thus, d(he,she) = 1. For a harder case, consider d(she,his). There are two ways to convert "she" into "his" but both take four edits. We could delete "he" from "she", leaving only the "s", and then insert "hi" in front of the "s". Or we could delete "s" and "e" from "she" and then follow the remaining "h" by "is". Either way, 4 edits are needed. Thus, d(she,his) = 4.
In a similar manner, we can discover d(he,his) = 3, d(he,hers) = 2, d(she,hers) = 3, and d(his,hers) = 3. A useful rule is that the edit distance is the sum of the lengths of the words minus twice the length of the longest common subsequence. For instance, the longest common subsequence of "his" and "hers" is "hs", so their edit distance is |his| + |hers| - 2|hs| = 3 + 4 -2*2 = 3.

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.

Your Answer		Score	Explanation
○ The minhash value for C1 is R6			
○ The minhash value for C4 is R2			
◉ The minhash value for C4 is R3	✔	1.00	
○ The minhash value for C2 is R3			
Total		1.00 / 1.00	

Question Explanation

Look at the rows in the stated order R4, R6, R1, R3, R5, R2, and for each row, make that row be the minhash value of a column if the column has not yet been assigned a minhash value.
We sart with R4, which only has 1 in column C3, so the minhash value for C3 is R4.
Next, we consider R6, which has 1 in C2 only. Since C2 does not yet have a minhash value, R6 becomes its value.

Next is R1, with 1's in C2 and C3. However, both these columns already have minhash values, so we do nothing.

Next, consider R3. It has 1's in C2 and C4. C2 already has a minhash value, but C4 does not. Thus, the minhash value of C4 is R3.

When we consider R5 next, we see it has 1's in C1 and C3. The latter already has a minhash value, but R5 becomes the minhash value for C1. Since all columns now have minhash values, we are done.

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.

Your Answer		Score	Explanation
C3 and C4			
C1 and C7			
C2 and C4			
C2 and C5	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

In the first band (first two rows) C1 and C4 both have (1,2), so they form a candidate pair. Also, C2 and C5 both have (2.3), so that is another candidate pair.
In the second band (rows 3 and 4) we find only C1 and C6 agree, and in the third band we find C1-C3 agree and C4-C7 agree. Thus, the five candidate pairs are C1-C4, C2-C5, C1-C6, C1-C3, and C4-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.

Your Answer		Score	Explanation
◉ The Jaccard similarity is 5/9.	✔	1.00	
○ The Jaccard similarity is 5/7.			
○ BRICABRAC has 8 2-shingles.			
○ ABRACADABRA has 9 2-shingles.			
Total		1.00 / 1.00	

Question Explanation

The 2-shingles for ABRACADABRA: AB, BR, RA, AC, CA, AD, DA.
The 2-shingles for BRICABRAC: BR, RI, IC, CA, AB, RA, AC.

There are 5 shingles in common:AB, BR, RA, AC, CA.

As there are 9 different shingles in all, the Jaccard similarity is 5/9.

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.

Your Answer	Score	Explanation
○ s3 is compared with 5 other strings.		
○ s3 is compared with 6 other strings.		
○ s3 is compared with 2 other strings.		
○ s6 is compared with 6 other strings.		
Total	0.00 / 0.00	

Question Explanation

First, we index a string of length L on the symbols appearing in its prefix of length floor(0.2L+1). Thus, strings of length 5 and 6 are indexed on their first two symbols, while strings of length 4 are indexed on their first symbol only. Thus, the index for *a* consists of {s_1, s_2, s_4, s_8}; the index for *b* consists of {s_1, s_3, s_5, s_6}, the index for *c* consists of {s_2, s_3, s_5, s_7}, and no other symbol is indexed at all.
For s_1, we examine the indexes for *a* and *b*, which contains all strings but s_7. Thus, s_1 is compared with 6 other strings.

For s_3, we examine the indexes for *b* and *c*, which together contain s_1, s_2, s_3, s_5, s_6, and s_7. Thus, s_3 is compared with five other strings.

For s_6, we examine only the index for *b*. Thus, s_6 is compared only with the three other strings s_1, s_3, and s_5.

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.

Your Answer		Score	Explanation
◉ (59,10)	✔	1.00	

- (53,10)
- (66,5)
- (63,8)

Total 1.00 / 1.00

Question Explanation

The L_1 distance from (x,y) to (0,0) is x+y. The L_1 distance from (x,y) to (100,40) is 140-x-y. Thus, (x,y) is assigned to (0,0) using the L_1 norm if x < 70-y.
When comparing L_2 distances, it is often better to use the squares of the distances. The square of the L_2 distance from (x,y) to (0,0) is x^2+y^2, and the square of the L_2 distance from (x,y) to (100,40) is $(100-x^2)+(40-y)^2 = 11600-200x-80y+x^2+y^2$. Thus, for (x,y) to be clustered with (100,40) according to the L_2 norm, we must have 200x+80y > 11600, or x > 58-2y/5. Thus, each of the correct answers is an (x,y) pair with 58-2y/5 < x < 70-y. For example, if y=10, we must have 54 < x < 60.