TI2736-C Assignment 1

Joost Pluim, jwpluim, 4162269
Pascal Remeijsen, premeijsen, 4286243

February 26, 2016

Shingles

1 Question 1

1.1 Question 1.1

When a shingle appears twice in a set, it is only saved once. Duplicates therefore arent́ saved. In the example set "ab" appears twice in the string, but it is only one time in the ShingleSet

2 Question 3

2.1 Question 3.1

The word "touchdown" appears in both strings and is larger than 5 (which is the shingle size). The rest of the words are different. This means that A is sentence 1, and B is sentence 2, that $|A \cap B|$ is 1 because of the space. The rest of the sentence isnt́ the same. This is bigger than 5 and therefore $\frac{A \cap B}{A \cup B}$ is small which means that the Jaccard Distance is big.

2.2 Question 3.2

Decreasing the k to 1 means that we have pretty much all shingles in common, which makes the Jaccard Distance small.

In case we increase the shingle size to 15, both strings dont́ have any shingles in common, so the Jaccard Distance is very big.

3 Question 5

3.1 Question 5.1

Only in the word touchdown, removing the spaces affects the similarity of shingles. Therefore the Jaccard distance will decrease, although it is only a little bit.

Minhashing

4 Question 2

4.1 Question 2.1

First we create our signature matrix

	s1	s2	s3	s4
a (0)	1	0	0	1
b (1)	0	0	1	0
c (2)	0	1	0	1
d (3)	1	0	1	1
e (4)	0	0	1	0

Table 1: Signature matrix

	s1	s2	s3	s4
h_1	∞	∞	∞	∞
h_2	∞	∞	∞	∞

Table 2: Initialization

$$h_1(0) = 1 \mod 5 = 1 \tag{1}$$
$$h_2(0) = (3 \cdot 0 + 1) \mod 5 = 1 \tag{2}$$

	s1	s2	s3	s4
h_1	1	∞	∞	1
h_2	1	∞	∞	1

Table 3: Row 0

$$h_1(1) = 2 \mod 5 = 2 \tag{3}$$
$$h_2(1) = (3 \cdot 1 + 1) \mod 5 = 4 \tag{4}$$

	s1	s2	s3	s4
h_1	1	∞	2	1
h_2	1	∞	4	1

Table 4: Row 1

$$h_1(2) = 3 \mod 5 = 3 \tag{5}$$
$$h_2(2) = (3 \cdot 2 + 1) \mod 5 = 2 \tag{6}$$

	s1	s2	s3	s4
h_1	1	3	2	1
h_2	1	2	4	1

Table 5: Row 2

$$h_1(3) = 4 \mod 5 = 3 \tag{7}$$
$$h_2(3) = (3 \cdot 3 + 1) \mod 5 = 0 \tag{8}$$

$$h_1(4) = 5 \mod 5 = 0 \tag{9}$$
$$h_2(4) = (3 \cdot 4 + 1) \mod 5 = 3 \tag{10}$$

	s1	s2	s3	s4
h_1	1	3	2	1
h_2	0	2	0	0

Table 6: Row 3

	s1	s2	s3	s4
h_1	1	3	0	1
h_2	0	2	0	0

Table 7: Row 4

5 Question 4

5.1 Question 4.1

The result from our script satisfies the result as found in Question 2.1

Locality Sensitive Hashing

6 Question 2

6.1 Question 2.1

Column segments from other bands can hash to the same hashcode. This means that when selecting candidates these sets will be set as candidate however the match is a match in completely different column segment.

7 Question 3

7.1 Question 3.1

If column segments in the signature matrix are really similar, the possibility of column segments Min Hash Signature being the same is pretty big. If column segments aren't the same however, the probability of the column segments (definitely if the bands are small) being the same are small. However probability is still there that column segments are the same. If two column segments are the same in only 1 case (in the many rows), this means both sets are immediately selected as candidates. This results in the fact that sets which are pretty much the same, are always selected, but sets which differ a lot, might also be selected rarely.

7.2 Question 3.2

Iteration 1 Result: (0,3)

Iteration 2 Result: (0,3)

Iteration 3 Result: (0,3)

Iteration 4 Result: (0,3), (0,2)

Iteration 5 Result: (0,3), (0,2)

Iteration 6 Result: (0,3), (0,2), (1,3), (2,3)

Iteration 7 Result: (0,3), (1,3)

Iteration 8 Result: (0,3), (0,2), (1,3)

Iteration 9 Result: (0,3)

Iteration 10 Result: (0,3)

If we calculate the Jaccard distances we find:
Jaccard distance (0,2) = 0,75
Jaccard distance (0,3) = 0,33
Jaccard distance (1,3) = 0,66
Jaccard distance (2,3) = 0,8

If we calculate the relative frequencies in our 10 samples, we find:
1 - Relative frequency (0,2) = 1 - (4/10) = 0,6
1 - Relative frequency (0,3) = 1 - (10/10) = 0
1 - Relative frequency (1,3) = 1 - (3/10) = 0,7
1 - Relative frequency (2,3) = 1 - (1/10) = 0,9

We see that the Jaccard distances fit the relative distance quite well.

7.3 Question 3.3

If the data set is too large to look compare against all data, we first have to determine possible candidates and compare our sample with this data.

7.4 Question 3.4

If the number of buckets is small, this means that the modulo with which we calculate the hash is small. This means that the chance of collisions between the column segments of the signature matrix will get bigger. Of course we donť want this, since we only want valid candidates in the same bucket.

7.5 Question 3.5

If the number of rows per band is small, this means that column segments can easily be the same for different columns, which means we get a lot of candidates/false positives. If the number of rows per band is the same as the length of the signature, this means that only sets which are exactly the same will be put in the same bucket.
The bigger the number of candidates, the smaller the effect of this algorithm, since we are using this

algorithm to divide the complete set into smaller sets. In the latter case, we put every set in its own bucket, which doesnt́ give any candidates at all.

References