TI2736-C Assignment 1

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

February 18, 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 arenť 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 isnť 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 donť 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

5 Question 4

5.1 Question 4.1

Locality Sensitive Hashing

6 Question 2

6.1 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

$$h_1(3) = 4 \mod 5 = 3 \tag{7}$$

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

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

Table 6: Row 3

$$h_1(4) = 5 \mod 5 = 0 \tag{9}$$

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

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

Table 7: Row 4

7 Question 3

7.1 Question 3.1

7.2 Question 3.2

7.3 Question 3.3

7.4 Question 3.4

7.5 Question 3.5

References