Assignment 2

Locality Sensitive Hashing

(1)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
R_1	1	2	1	1	2	5	4
R_2	2	3	4	2	3	2	1
R_3	3	1	2	3	1	3	2
R_4	4	1	3	1	2	4	4
R_5	5	2	5	1	1	5	1
R_6	6	1	6	4	1	1	4

Given that LSH with three bands of 2 rows each

From Band 1, i.e rows R_1 and R_2

$C_1 = C_4$, $C_2 = C_5$ have same Candidate pairs

From Band 2, i.e rows R_3 and R_4

$C_1 = C_6$ have same Candidate pairs

From Band 3, i.e rows R_5 and R_6

$C_1 = C_3$, $C_4 = C_7$ have same Candidate pairs

(2) Given,

N pairs of signatures that are 50% similar

M pairs of signatures that are 20% similar

Given bands b are 1, 2, 3, 4, 6, 8, 12, 24

r	$b = \frac{\text{Hash}}{r}$	FP $1-(1-s^r)^b$	FN $(1-s^r)^b$
1	24	0.99	$5.9\ e^{-8}$
2	12	0.38	0.031
3	8	0.06	0.34
4	6	0.009	0.678
6	4	0.0002	0.928
8	3	$7.67\ e^{-6}$	0.988
12	2	$8.19\ e^{-9}$	0.9995
24	1	0	0.9999

if H : N is the ratio 1 : 1 then choose $r = 3$ 0.40

10 : 1 then choose $r = 2$ 0.38

100 : 1 then choose $r = 1$ 0.99

1000 : 1 then choose $r = 1$ 0.99

(3) Given, ABRACADABRA

BRICABRAC

How many two shingles does ABRACADABRA have?

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

Total 7

How many two shingles does BRICABRAC have?

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

Total 7

How many 2-shingles do they have in common?

Common = 5

Jaccard Similarity = $\frac{5}{7+7-5} = \frac{5}{9}$

(4)

	C_1	C_2	C_3	C_4
R_1	0	1	1	0
R_2	1	0	1	1
R_3	0	1	0	1
R_4	0	0	1	0
R_5	1	0	1	0
R_6	0	1	0	0

Jaccard Similarity $JS = \frac{b_{11}}{b_{10} + b_{01} + b_{11}}$

$JS(C_1, C_3) = 0$

$JS(C_1, C_2) = 0.5$

$JS(C_1, C_4) = 0.33$

$JS(C_2, C_3) = 0.167$

$JS(C_2, C_4) = 0.25$

$JS(C_3, C_4) = 0.2$

(5) Given order of rows $R_4, R_6, R_1, R_3, R_5, R_2$

	C_1	C_2	C_3	C_4
R_4	0	0	R_4	0
R_6	0	R_6	R_4	0
R_1	0	R_6	R_4	0
R_3	0	R_6	R_4	R_3
R_5	R_5	R_6	R_4	R_3
R_2	R_5	R_6	R_4	R_3

Minhash value of

$C_1 = R_5$

$C_2 = R_6$

$C_3 = R_4$

$C_4 = R_3$