Assignment - 2

Question 1

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Band 1	1	2	1	1	2	5	4
	2	3	4	2	3	2	2
Band 2	3	1	2	3	1	3	2
	4	1	3	1	2	4	4
Band 3	5	2	5	1	1	5	1
	6	1	6	4	1	1	4

Band 1 = (C_1, C_4) , (C_2, C_5)

Band 2 = (C_1, C_6)

Band 3 = (C_1, C_3) , (C_4, C_7)

Question 2

M = 50% ⇒ false positive

N = 20% ⇒ false negative

R	b	FP(M)	FN(N)
1	24	0.9952	0.00000005
2	12	0.3872	0.0316
3	8	0.0622	0.343
4	6	0.0095	0.6789
6	4	0.002	0.9389
8	3	$7.67E^{-6}$	0.98832
12	2	$8.192E^{-9}$	0.9995
24	1	0	0.99999

r: rows
b: bins

$1-(1-s^r)^b$ → for FP

for FN without the 2

a) 1:1 → M:N →
0.995277
0.418966
0.705845
0. 688495
0.989334
0.99951
0.99999

~~b) 1:10 → M:10N →~~

1:10 M:10N	1:100 M:100N	1:1000 M:1000N
0.99527	0.99528	0.99533
0.70405	0.35492	32.06364
3.49932	32.4231	343.6711
6.79890	67.9029	678.9473
9.38975	93.8952	938.9499
9.88327	98.8327	988.3269
9.995117	99.9511	999.9517
9.99999	99.999	999.9997

Question 3

Find the set of 2 shingles for documents given

document 1 : ABRACADABRA

set (A) → shingle of size 2

↳ { AB, BR, RA, AC, CA, AD, DA, AB, BR, RA }

so ⇒ { AB, BR, RA, AC, CA, AD, DA }

document 2 : BRICABRAC

set B ⇒ { BR, RI, IC, CA, AB, BR, RA, AC }

~~Union = 7+7~~

① 2 shingles document 1 has ⇒ 7

② 2 shingles document 2 has ⇒ 7

③ 2 shingles they have in common

doc-1 ⇒ AB, BR, RA, AC, CA, AD, DA

doc-2 ⇒ BR, RI, CA, AB, BR, RA, AC

so 5 shingles

④ Jaccard similarity ⇒ $\frac{\text{Intersection}}{\text{Union}}$

$= \frac{5}{7+7-5} = 5/9$

Question 4:

Compute the Jaccard similarity between each pair of columns.

	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

$J(C_1, C_2) = 0/5 = 0$

$J(C_1, C_3) = 2/4 = 0.5$

$J(C_1, C_4) = 1/3 = 0.33$

$J(C_2, C_3) = 1/6 = 0.167$

$J(C_2, C_4) = 1/4 = 0.25$

$J(C_3, C_4) = 1/5 = 0.2$

Question 5:

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

Minhash value of C_1 = R_5

Minhash value of C_2 = R_6

Minhash value of C_3 = R_4

Minhash value of C_4 = R_3