BIG DATA 課題２（解答付）(Solutions to Homework 2)

第１問: 以下の文書１と文書２が与えられたとする.

文書 1:

HIRODAIHERO

文書２:

BIGDATAHERO

1-1. 各文書における 2 シングル（2-shingles）の集合を求めよ.

文書 1

D1 = {HI, IR, RO, OD, DA, AI, IH, HE, ER}

文書２

D2 = {BI, IG, GD, DA, AT, TA, AH, HE, ER, RO}

1-2. ２つの集合に対して Jaccard 類似度を計算せよ

$$\mathrm{SIM}(D1, D2) = \frac{|D1 \cap D2|}{|D1 \cup D2|} = \frac{4}{15}$$

$$|D1 \cap D2| = |\{RO, DA, HE, ER\}| = 4$$

$$|D1 \cup D2| = |\{HI, IR, RO, OD, DA, AI, IH, HE, ER, BI, IG, GD, AT, TA, AH\}| = 15$$

第２問: 以下の 6×4 特徴行列が与えられたとする（C1 は 第１列, R1 は第１行):

	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

2-1. 各列の対に対して Jaccard 類似度を計算せよ.

SIM(C1, C2) = 0

SIM(C1, C3) = 2/4

SIM(C1, C4) = 1/3

SIM(C2, C3) = 1/6

SIM(C2, C4) = 1/4

SIM(C3, C4) = 1/5

2-2. 次の行の並べ替えを用いて R4, R6, R1, R3, R5, R2, ミンハッシュ（minhash）を計算せよ.

	C1	C2	C3	C4
R4	0	0	1	0
R6	0	1	0	0
R1	0	1	1	0
R3	0	1	0	1
R5	1	0	1	0
R2	1	0	1	1

h(C1) = 5, h(C2) = 2, h(C3) = 1, h(C4) = 4

第 3 問: 以下の行列はミンハッシュシグネチャー行列である.

.

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

LSH を r = 2, b = 3 で用いた場合，すべての候補対を求めよ.

(C1, C4), (C2, C5)

(C1, C6)

(C1, C3), (C4, 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