BIG DATA 課題 2 （Homework 2）

提出締切（Submission deadline）: 2025 年 1 月 7 日 23:50

第 1 問: 以下の文書 1 と文書 2 が与えられたとする．(The following two "documents" are given.)

文書 1 (Document 1):

HIRODAIHERO

文書 2 (Document 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．(Find the set of 2-shingles for each of the "documents".)

1-2. ２つの集合に対して Jaccard 類似度を計算せよ．(What is the Jaccard similarity between the two “documents", i.e. between their 2-shingles sets?)

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): (The following 6×4 matrix is given, where C1 means column 1, R1 means row 1, etc.)

	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．(Compute the Jaccard similarity between each pair of columns of the matrix.)

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. (Perform a minhashing of the data, with the order of rows: R4, R6, R1, R3, R5, R2.)

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．(The following is a matrix representing the signatures of seven columns, C1 through 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

LSH を r = 2, b = 3 で用いた場合，すべての候補対を求めよ．(Suppose we use locality-sensitive hashing with 3 bands of 2 rows each. Find all the candidate pairs.)