

• - Billion rows ka list, 100 rows are enough for signature.

Date: _____

⇒ Min-Hashing (Koi bhi matrix ka 100 row)

Jitnay cols utnay boxes bana

Signature Matrix

∞ ∞ ∞ ∞

∞ 1 ∞ 1

2 1 2 1
 ∞ ∞ ∞ ∞

2 1 2 1
2 1 ∞ 1

↓
Bar ho to replace nai krna.
we can not write 3 as it is larger than 2.

2 1 2 1
2 1 4 1

Write Row number if it is smaller than previous one.

4	3	1	0	1	0
2	4	1	0	0	1
1	7	0	1	0	1
3	6	0	1	0	1
6	1	0	1	0	1
7	2	1	0	1	0
5	3	1	0	1	0

Signature $\rightarrow \frac{n(n-1)}{2}$

$= \frac{4(4-1)}{2} = \frac{12}{2} = 6$

$\text{Sim}(c_1, c_2) = 0$

$\text{Sim}(c_1, c_3) = 0.75$

$\text{Sim}(c_1, c_4) = 0.14$

0	3	Signature
0.67		
0		

A-2/02.

Date: _____

$$\begin{aligned} C_2 C_3 &= 0 \\ C_2 C_4 &= 0.75 \\ C_3 C_4 &= 0 \end{aligned}$$

Input Matrix

	D_1	D_2	D_3		
3	1	0	0	3	4
4	1	1	0	5	5
5	1	1	1	1	6
1	1	1	1	2	7
6	0	1	1	4	2
7	0	1	0	6	2
2	0	0	1	7	1

$$\frac{3(3-1)}{2}$$

$$\frac{3(2)}{2} = 3$$

1	1	1
1	1	1
4	1	1

$$\text{Sim}(C_1, C_2) = \frac{3}{6} = \frac{1}{2} = 0.5$$

$$\text{Sim}(C_1, C_3) = \frac{2}{6} = \frac{1}{3} = 0.33$$

$$\text{Sim}(C_2, C_3) = \frac{3}{6} = \frac{1}{2} = 0.5$$

Signature form

$$\frac{2}{1} = 2$$

$$\frac{2}{1} = 2$$

$$\frac{3}{3} = 1$$