HIROSHIMA UNIVERSITY 広島大学

課題 2 Finding Similar Items (Homework 2)

Big Data KA218001

ビッグデータ KA218001

Submission Information

Date	Student ID	Name
4/1/2025 (DD/MM/YYYY)	C240424	Yousef Ibrahim Gomaa Mahmoud Mabrouk

第 1 問の答え:

1.1. K-Shingles Sets (K=2)

- 文書 1 (Document 1): HIRODAIHERO
    - Set of 2-shingles: {HI, IR, RO, OD, DA, AI, IH, HE, ER}
- 文書 2 (Document 2): BIGDATAHERO
    - Set of 2-shingles: {BI, IG, GD, DA, AT, TA, AH, HE, ER, RO}

1.2. Jaccard Similarity of Documents

$$\text{Jaccard Similarity } = Sim(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

- Set 1: {HI, IR, RO, OD, DA, AI, IH, HE, ER}
- Set 2: {BI, IG, GD, DA, AT, TA, AH, HE, ER, RO}

- From Table 1, which shows the Characteristic Matrix of Set 1 and 2, the Jaccard Similarity can be calculated as follows:
  - $Sim(Set1, Set2) = \frac{|\text{Set} \cap \text{Set2}|}{|\text{Set1} \cup \text{Set2}|} = \frac{\text{RO+DA+HE+E}}{\text{HI+IR+RO+OD+DA+AI...+AH}} = \frac{4}{15} \approx 0.267$

	Set 1	*Set 2*
HI	*1*	*0*
IR	*1*	*0*
RO	*1*	*1*
OD	*1*	*0*
DA	*1*	*1*
AI	*1*	*0*
IH	*1*	*0*
HE	*1*	*1*
ER	*1*	*1*
BI	*0*	*1*
IG	*0*	*1*
GD	*0*	*1*
AT	*0*	*1*
TA	*0*	*1*
AH	*0*	*1*

Table 1.

第 2 問の答え:

	$C1$	$C2$	$C3$	$C4$
$R1$	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. Jaccard Similarity of Each Pair

- $\text{Sim}(\text{C1}, \text{C2}) = \frac{0}{R1+R2+R3+R5+R6} = 0$
- $\text{Sim}(\text{C1}, \text{C3}) = \frac{R2+R5}{R1+R2+R4+R5} = \frac{2}{4} = 0.5$
- $\text{Sim}(\text{C1}, \text{C4}) = \frac{R2}{R2+R3+R5} = \frac{1}{3} = 0.33$
- $\text{Sim}(\text{C2}, \text{C3}) = \frac{R1}{R1+R2+R3+R4+R5+R6} = \frac{1}{6} = 0.167$
- $\text{Sim}(\text{C2}, \text{C4}) = \frac{R3}{R1+R2+R3+R6} = \frac{1}{4} = 0.25$
- $\text{Sim}(\text{C3}, \text{C4}) = \frac{R2}{R1+R2+R3+R4+R5} = \frac{1}{5} = 0.2$
  - Note that Sim(C3, C1) is the same as Sim(C1, C3), this applies to the other columns as well.

- *Minhash (row order: R4, R6, R1, R3, R5, R2 ~ 3, 5, 0, 2, 4, 1)*
  - Let there be two hash functions and replace the rows with integers. (0 → k-1)
    - h1(x) = x+1 mod k, h2(x) = 2x+1 mod k, where k = row count

x	$C1$	$C2$	$C3$	$C4$	$h1(x)$	$h2(x)$
0	0	1	1	0	1	1
1	1	0	1	1	2	3
2	0	1	0	1	3	5
3	0	0	1	0	4	1
4	1	0	1	0	5	3
5	0	1	0	0	0	5

$Sig.Matrix$:

	$C1$	$C2$	$C3$	$C4$
$h1$	∞	∞	∞	∞
$h2$	∞	∞	∞	∞

- Step 1: Pick min(Sig(i, j), hx(x)) for R4 (x=3) which updates C3 as it has 1.

$$Sig.Matrix: \quad \begin{array}{c|cccc} & C1 & C2 & C3 & C4 \\ h1 & \infty & \infty & 4 & \infty \\ h2 & \infty & \infty & 1 & \infty \end{array}$$

- Step 2: Similarly, for R6 (x=5), only C2 has 1.

$$Sig.Matrix: \quad \begin{array}{c|cccc} & C1 & C2 & C3 & C4 \\ h1 & \infty & 0 & 4 & \infty \\ h2 & \infty & 5 & 1 & \infty \end{array}$$

- Step 3: For R1 (x=0), C2 and C3 have 1. (Update C3 and h2 of C2)

$$Sig.Matrix: \quad \begin{array}{c|cccc} & C1 & C2 & C3 & C4 \\ h1 & \infty & 0 & 1 & \infty \\ h2 & \infty & 1 & 1 & \infty \end{array}$$

- Step 4: For R3 (x=2), C2 and C4 have 1. (Update only C4)

$$Sig.Matrix: \quad \begin{array}{c|cccc} & C1 & C2 & C3 & C4 \\ h1 & \infty & 0 & 1 & 3 \\ h2 & \infty & 1 & 1 & 5 \end{array}$$

- Step 5: For R5 (x=4), C1 and C3 have 1. (C3 is unchanged)

$$Sig.Matrix: \quad \begin{array}{c|cccc} & C1 & C2 & C3 & C4 \\ h1 & 5 & 0 & 1 & 3 \\ h2 & 3 & 1 & 1 & 5 \end{array}$$

- Step 6: Finally, for R2 (x=1), C1, C3, and C4 have 1. (Update h(1,1), h(1,4), and h(2,4)

$$Sig.Matrix: \quad \begin{array}{c|cccc} & C1 & C2 & C3 & C4 \\ h1 & 2 & 0 & 1 & 2 \\ h2 & 3 & 1 & 1 & 3 \end{array}$$

  - According to this signature matrix, columns C1 and C4 are identical. C2 and C3 are identical in only half the rows in this case as well.
  - However, this is not entirely true due to the small sample used.

$$Sim(C1, C4) = \frac{x}{x + y} = \frac{1}{1 + 2} = \frac{1}{3}$$

Where x are the rows where both columns had 1, and y are the rows where either had 1 and the other had 0.

- $Sim(C1, C2) = \frac{0}{0+5} = 0$, which is the same as the result from the signature. This proves that Minhash can be useful to remove dissimilar features effectively.

第 3 問の答え:

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

- The signature matrix is divided into 3 bands, and each has 2 rows.

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

- Since a hashing function would put identical pairs in the same bucket, and assuming other pairs have a low chance of appearing in said same bucket.
- Then the candidate pairs for each band are:
  - Band 1: (RED)
    - (C1,C4), (C2, C5)
  - Band 2: (BLUE)
    - (C1,C6)
  - Band 3: (GREEN)
    - (C1,C3), (C4, C7)
- The aggregated candidate list is: {(C1,C4), (C2, C5), (C1,C6), (C1,C3), (C4, C7)}