

HIROSHIMA UNIVERSITY広島大学

課題２Finding 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. *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. *Jaccard Similarity of Documents*

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

|  | *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.* | | |

* From Table 1, which shows the Characteristic Matrix of Set 1 and 2, the Jaccard Similarity can be calculated as follows:

**第２問の答え:**

1. *Jaccard Similarity of Each Pair*
   * + 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
  + Step 1: Pick min(Sig(i, j), hx(x)) for R4 (x=3) which updates C3 as it has 1.
  + Step 2: Similarly, for R6 (x=5), only C2 has 1.
  + Step 3: For R1 (x=0), C2 and C3 have 1. (Update C3 and h2 of C2)
  + Step 4: For R3 (x=2), C2 and C4 have 1. (Update only C4)
  + Step 5: For R5 (x=4), C1 and C3 have 1. (C3 is unchanged)
  + Step 6: Finally, for R2 (x=1), C1, C3, and C4 have 1. (Update h(1,1), h(1,4), and h(2,4)
    - 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.

Where *x* are the rows where both columns had 1, and *y* are the rows where either had 1 and the other had 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)}