1, Calculate the set of 2-shingles for each sentence and use matrix to represent the sentences, where the element is enumerated from 0.

|  |  |  |
| --- | --- | --- |
|  | Sentence | 2-shingles |
| S(I) | abbcba | {ab, bb, bc, cb, ba} |
| S(II) | ccbbca | {cc, cb, bb, bc, ca} |
| S(III) | bbaacb | {bb, ba, aa, ac, cb} |
| S(IV) | bbacab | {bb, ba, ac, ca, ab} |
| S(V) | cbbbac | {cb, bb, ba, ac} |

Generate the matrix:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Element | I | II | III | IV | V |
| (0)ab | 1 | 0 | 0 | 1 | 0 |
| (1)bb | 1 | 1 | 1 | 1 | 1 |
| (2)bc | 1 | 1 | 0 | 0 | 0 |
| (3)cb | 1 | 1 | 1 | 0 | 1 |
| (4)ba | 1 | 0 | 1 | 1 | 1 |
| (5)cc | 0 | 1 | 0 | 0 | 0 |
| (6)ca | 0 | 1 | 0 | 1 | 0 |
| (7)aa | 0 | 0 | 1 | 0 | 0 |
| (8)ac | 0 | 0 | 1 | 1 | 1 |

2, Compute the minhash signature for each column if we use the following three hash functions: h1(𝑥) = 2𝑥 + 1 mod 6; h2(𝑥) = 8𝑥 + 2 mod 10; h3(𝑥)=6𝑥+2mod10.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Hash Function | I | II | III | IV | V |
| h1 | 1 | 1 | 1 | 1 | 1 |
| h2 | 0 | 0 | 0 | 0 | 0 |
| h3 | 0 | 0 | 0 | 0 | 0 |

3, Which of these hash functions are true permutations?

None of above hash functions are true permutation

4, How close are the estimated Jaccard similarities for the 10 pairs of columns to the true Jaccard similarities?

|  |  |  |
| --- | --- | --- |
| Pair | Col/Col | Sig/Sig |
| 1-2 | 0.43 | 1 |
| 1-3 | 0.43 | 1 |
| 1-4 | 0.43 | 1 |
| 1-5 | 0.5 | 1 |
| 2-3 | 0.25 | 1 |
| 2-4 | 0.25 | 1 |
| 2-5 | 0.28 | 1 |
| 3-4 | 0.43 | 1 |
| 3-5 | 0.8 | 1 |
| 4-5 | 0.5 | 1 |