Seoul National University

M1522.001400-001 Introduction to Data Mining

Homework 1: Finding Similar Items (Chapter 3)

2017-18538 Hwang SunYoung

1.

1) Jaccard similarity is not appropriate in situations where order is important. For data such as vertors and matrices, it is not appropriate to calculate the similarity with Jaccard similarity because when the order differs, it becomes completely different data.

2) Jaccard similarity can find textually similar documents, but it is difficult to find documents with similar meanings. So Jaccard similarity is not suitable for finding documents in a similar context.

2.

1) the last ten 3-shingles are "of cases grow", "cases grow the", "grow the rate", "the rate at", "rate at which", "at which it", "which it grows", "it grows increases", "grows increases as", "increases as well"

2) the number of 5-shingles is 40.

3.

1) ↓ permutation

| element | $S_1$ | $S_2$ | $S_3$ | $S_4$ | 2x+1mod6 | 3x+2mod6 | 5x+2mod6 |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 | 1 | 2 | 2 |
| 1 | 0 | 1 | 0 | 0 | 3 | 5 | 1 |
| 2 | 1 | 0 | 0 | 1 | 5 | 2 | 0 |
| 3 | 0 | 0 | 1 | 0 | 1 | 5 | 5 |
| 4 | 0 | 0 | 1 | 1 | 3 | 2 | 4 |
| 5 | 1 | 0 | 0 | 0 | 5 | 5 | 3 |

Min-hash signature matrix

| $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|
| 5 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 |

| | | | |
|---|---|---|---|
| 0 | 1 | 4 | 0 |

2) true permutation is $h_3$ function.

3)

Similarities:

| | 1-2 | 1-3 | 1-4 | 2-3 | 2-4 | 3-4 |
|---|---|---|---|---|---|---|
| Col/Col(true) | 0 | 0 | 0.25 | 0 | 0.25 | 0.25 |
| Sig/Sig | 0.33 | 0.33 | 0.67 | 0.67 | 0.67 | 0.67 |
| difference | 0.33 | 0.33 | 0.42 | 0.67 | 0.42 | 0.42 |

The estimated Jaccard similarities are all different from true Jaccard similarities.

4.

calculate the threshold for which false positive rate is less than 0.05, when r=4, b=10,

$$\int_0^p \{1-(1-p^4)^{10}\}dp < 0.05$$

After calculating, it follows p>0.49

The threshold is 0.49