Seoul National University

M1522.001400 Introduction to Data Mining

Spring 2016, Kang

Homework 2: Finding Similar Items (Chapter 3)

Due: Mar 30, 09:30 AM

Reminders

- The points of this homework add up to 100.
- Like all homeworks, this has to be done individually.
- Lead T.A.: Minsoo Jung (qtyp456987@gmail.com)
- Please type your answers *in English*. Illegible handwriting may get no points, at the discretion of the graders.
- If you have a question about assignments, please upload your question in eTL.
- If you want to use slipdays or consider late submission with penalties, please note that you are allowed one week to submit your assignment after the due date. That is, after Apr 6, we will NOT receive your submission for this assignment.

Remember that:

- Whenever you are making an assumption, please state it clearly.

Question 1

(Exercise 3.1.1) Compute the Jaccard similarities of each pair of the following three sets: $\{1, 2, 3, 4\}$, $\{2, 3, 5, 7\}$, and $\{2, 4, 6\}$. [15 points]

Let $S_1 = \{1, 2, 3, 4\}, S_2 = \{2, 3, 5, 7\}, and\ S_3 = \{2, 4, 6\}$

$\text{sim}(S_1, S_2) = \frac{1}{3}, \text{sim}(S_2, S_3) = \frac{1}{6}, sim(S_1, S_3) = \frac{2}{5}$

Question 2

(Exercise 3.2.1) What are the first ten 3-shingles in the following sentence? [15 points]

> The most effective way to represent documents as sets, for the purpose of identifying lexically similar documents is to construct from the document the set of short strings that appear within it.

※ Tokens are characters (excluding space).

{the, hem, emo, mos, ost, ste, tef, eff, ffe, fec}

Question 3

(Exercise 3.2.3) What is the largest number of k-shingles a document of n bytes can have? You may assume that

- the size of the alphabet is large enough that the number of possible strings of length k is at least as n, and
- each byte corresponds to a token.

[20 points]

The largest number of k-shingles: $n - k + 1$

Question 4

(Exercise 3.3.3) In Figure 3.5 there is a matrix with six rows.

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Figure 3.5: Matrix for Exercise 3.3.3

(a) Compute the minhash signature for each column if we use the following three hash functions: [10 points]

i. $h_1(x) = 2x + 1 \bmod 6.$

ii. $h_2(x) = 3x + 2 \bmod 6.$

iii. $h_3(x) = 5x + 2 \bmod 6.$

	h_1	h_2	h_3	S_1	S_2	S_3	S_4
0	1	2	2	0	1	0	1
1	3	5	1	0	1	0	0
2	5	2	0	1	0	0	1
3	1	5	5	0	0	1	0
4	3	2	4	0	0	1	1
5	5	5	3	1	0	0	0

The table below shows the minhash signature matrix.

	S_1	S_2	S_3	S_4
h_1	5	1	1	1
h_2	2	2	2	2
h_3	0	1	4	0

(b) Which of these hash functions are true permutations? [5 points]

The hash function $h_3(\mathrm{x})$ is a true permutation.

(c) Compute the true Jaccard similarities and the estimated Jaccard similarities for the six pairs of columns. [10 points]

	True	Estimated
$\mathrm{sim}(S_1, S_2)$	0	$\frac{1}{3}$
$\mathrm{sim}(S_1, S_3)$	0	$\frac{1}{3}$
$\mathrm{sim}(S_1, S_4)$	$\frac{1}{4}$	$\frac{2}{3}$
$\mathrm{sim}(S_2, S_3)$	0	$\frac{2}{3}$
$\mathrm{sim}(S_2, S_4)$	$\frac{1}{4}$	$\frac{2}{3}$
$\mathrm{sim}(S_3, S_4)$	$\frac{1}{4}$	$\frac{2}{3}$

Question 5

Suppose $\mathrm{r} = 3$ and $\mathrm{b} = 10$.

(a) (Exercise 3.4.1) Evaluate the S-curve $1 - (1 - s^r)^b$ for $\mathrm{s} = 0.1, 0.2, \ldots, 0.9$, for r and b. [15 points]

s	$1 - (1 - s^r)^b$
0.1	0.0100
0.2	0.0772
0.3	0.2394
0.4	0.4839
0.5	0.7369
0.6	0.9123
0.7	0.9850
0.8	0.9992
0.9	0.9999

(b) Compute the threshold for which false positive rate is less than 0.05(calculate down to three decimal points). [10 points]

$$\int (1 - (1 - s^3)^{10} ds < 0.05$$

$$s < 0.390181$$

The threshold is 0.390181.