Big Data Analytics

ESSEC

Olga Klopp

Home work 3 : Finding Similar Items, part 2

1. (**Exercise 3.2.3 MMDS book**) 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. (In UTF-8 encoding each letter occupies 1 byte(8 bits).)
   Solution: The number of k-shingles = the number of characters $-k+1 = n-k+1$

2. (**Exercise 3.3.2 MMDS book**) Using the data from Fig. 3.4, add to the signatures of the columns the values of the following hash functions:

   - $h_3(x) = 2x + 4 \bmod 5$
   - $h_4(x) = 3x - 1 \bmod 5$

Row	S_1	S_2	S_3	S_4	$x+1 \bmod 5$	$3x+1 \bmod 5$
0	1	0	0	1	1	1
1	0	0	1	0	2	4
2	0	1	0	1	3	2
3	1	0	1	1	4	0
4	0	0	1	0	0	3

Figure 3.4 Hash functions computed for the matrix of Fig. 3.2

Solution: ˇ

Rows	$2x + 4 \bmod 5$	$3x - 1 \bmod 5$
0	4	4
1	1	2
2	3	0
3	0	3
4	2	1

3. (**Exercise 3.3.3 MMDS book**) In Fig. 3.5 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

   - Compute the minhash signature for each column if we use the following three hash functions: $h_1(x) = 2x + 1 \bmod 6$; $h_2(x) = 3x + 2 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$.
   - Which of these hash functions are true permutations?
   - How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

Solution: ˇ

Rows	$2x + 1 \bmod 6$	$3x + 2 \bmod 6$	5x + 2 mod 6
0	1	2	2
1	3	5	1
2	5	2	0
3	1	5	5
4	3	2	4
5	5	5	3

h_3 is a true permutation.

To compute the signatures: Let $SIG(i, c)$ be the element of the signature matrix for the ith hash function and column c. Initially, set $SIG(i, c)$ to ∞ for all i and c. We handle row r by doing the following:

(a) Compute $h_1(r), h_2(r), ..., h_n(r)$.

(b) For each column c do the following:

   i. If c has 0 in row r, do nothing.

   ii. However, if c has 1 in row r, then for each $i = 1, 2, ..., n$ set $SIG(i, c)$ to the smaller of the current value of $SIG(i, c)$ and $h_i(r)$.

Applying this algorithm we get:

	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

- $\mathrm{Sim}(S_1, S_2) = 0$, estimated 1/3
- $\mathrm{Sim}(S_1, S_3) = 0$, estimated 1/3
- $\mathrm{Sim}(S_1, S_4) = 1/4$, estimated 2/3
- $\mathrm{Sim}(S_2, S_3) = 0$, estimated 2/3
- $\mathrm{Sim}(S_2, S_4) = 1/4$, estimated 2/3
- $\mathrm{Sim}(S_3, S_4) = 1/4$, estimated 2/3