# Big Data Analytics

**ESSEC**

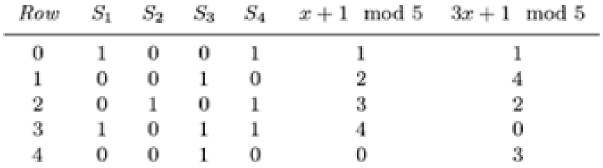
Mohamed Ndaoud

Homework 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).)

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:

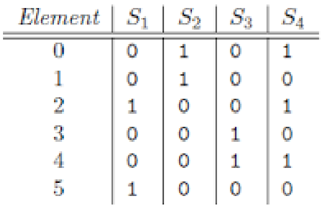
   - $h_3(x) = 2x + 4 \text{ mod } 5$
   - $h_4(x) = 3x - 1 \text{ mod } 5$



| *Row* | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $x+1$ mod 5 | $3x+1$ mod 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

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: $h_1(x) = 2x + 1 \text{ mod } 6$; $h_2(x) = 3x + 2 \text{ mod } 6$ ; $h_3(x) = 5x + 2 \text{ mod } 6$.
   - 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?

4. (**On Python**) Evaluate the S-curve $1 - (1 - s^r)^b$ for $s = 0.1, 0.2, \ldots, 0.9$, for the following values of $r$ and $b$:

   - $r = 3$ and $b = 10$.
   - $r = 6$ and $b = 20$.
   - $r = 5$ and $b = 50$.