02.12.17

Problem set 6

Exercise 3

geg.

El.	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

$h_1(x) = 2x+1 \mod 6$

$h_2(x) = 3x+2 \mod 6$

$h_3(x) = 5x+2 \mod 6$

Calculate minhash signatures for set 1

Ind.	S_1	h_1 Perm S_1	h_2 Perm S_1	h_3 Perm S_1
0	0			1
1	0	0,0		0
2	1		0,1,0	0
3	0	0,0		1
4	0			0
5	1	1,1	0,0,1	0

↓ minhash(S_1) = 5 ↓ minhash(S_1) = 2 ↓ minhash(S_1) = 0

Permutations

$h_1(0) = 1$	$h_2(0) = 2$	$h_3(0) = 2$
$h_1(1) = 3$	$h_2(1) = 5$	$h_3(1) = 1$
$h_1(2) = 5$	$h_2(2) = 2$	$h_3(2) = 0$
$h_1(3) = 1$	$h_2(3) = 5$	$h_3(3) = 5$
$h_1(4) = 3$	$h_2(4) = 2$	$h_3(4) = 4$
$h_1(5) = 5$	$h_2(5) = 5$	$h_3(5) = 3$

Calculate minhash signatures for set 2, set 3, set 4

set 2

Ind	S_2	h_1 Perm S_2	h_2 Perm S_2	h_3 Perm S_3
0	1			0
1	1	1,0		1
2	0		1,0,0	1
3	0	1,0		0
4	0			0
5	0	0,0	1,0,0	0

h_1 minhash$(S_2) = 1$

h_2 minhash$(S_2) = 2$

h_3 minhash$(S_2) = 1$

set 3

Ind.	S_3	h_1 Perm S_3	h_2 Perm S_3	h_3 Perm S_3
0	0			0
1	0	0,1		0
2	0		0,0,1	0
3	1	0,1		0
4	1			1
5	0	0,0	0,1,0	1

h_1 minhash$(S_3) = 1$

h_2 minhash$(S_3) = 2$

h_3 minhash$(S_3) = 4$

set 4

Ind.	S_4	h_1 Perm S_4	h_2 Perm S_4	h_3 Perm S_4
0	1			1
1	0	1,0		0
2	1		1,1,1	1
3	0	0,1		0
4	1			1
5	0	1,0	0,0,0	0

h_1 minhash$(S_1) = 1$

h_2 minhash$(S_4) = 2$

h_3 minhash$(S_4) = 0$

→ minhash Signatures

	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) Oney h_3 is a true permutation as the two other hash functions result in collisions.

The collisions can be seen in the Permutations table on the last page.

c) Jaccard similarity : $\text{sim}(S_1, S_1) = \frac{|S_1 \cap S_2|}{S_1 \cup S_2}$

$\text{sim}(S_1, S_2) = \frac{0}{4} = 0$
$\text{sim}(S_1, S_3) = 0/4 = 0$
$\text{sim}(S_1, S_4) = 1/4$
$\text{sim}(S_2, S_3) = 0/4 = 0$
$\text{sim}(S_2, S_4) = 1/4$
$\text{sim}(S_3, S_4) = 1/4$

Similarity of minhash signatures $\hat{=}$ fraction of hash functions in which they agree

$\text{sim}_h(S_1, S_2) = 1/3$
$\text{sim}_h(S_1, S_3) = 1/3$
$\text{sim}_h(S_1, S_4) = 2/3$
$\text{sim}_h(S_2, S_3) = 2/3$
$\text{sim}_h(S_2, S_4) = 2/3$
$\text{sim}_h(S_3, S_4) = 4/3$

⇒ The two similarity measures differ especially in $\text{sim}(S_2, S_3) = 0$ and $\text{sim}_h(S_2, S_3) = 2/3$. Obviously the similarity between the two sets should be zero, as they have nothig in common.

Thus it is important to use good hash functions.