HW6 LiChen Liang

P1

a) Assume 2 kind of extreme cases of document: all unique letters, all repeated letters. Assume 2 kind of extreme k: $k=1$ and $k=N-1$

The max possible size is when $k=1$ and all unique letters, size $= N$

The min possible size is when all repeated letters despite of k value, size $= 1$

For multiset, both max and min are equal, size $= N - k + 1$

b) Since every value in set can only be connected to another unique value, there's only 4 kinds of document regardless of N value.

P2

a)

Element	h_1	h_2	h_3
0	1	2	2
1	3	5	1
2	5	2	0
3	1	5	5
4	3	2	4
5	5	5	3

initially

	S_1	S_2	S_3	S_4
h_1	∞	∞	∞	∞
h_2	∞	∞	∞	∞
h_3	∞	∞	∞	∞

	S_1	S_2	S_3	S_4
$h_1(0)$	∞	1	∞	1
$h_2(0)$	∞	2	∞	2
$h_3(0)$	∞	2	∞	2

⇒

$h_1(1)$	∞	1	∞	1
$h_2(1)$	∞	2	∞	2
$h_3(1)$	∞	1	∞	2

$h_1(2)$	5	1	∞	1
$h_2(2)$	2	2	∞	2
$h_3(2)$	0	1	∞	0

⇒

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

	$h_1(4)$	5	1	1	1
$\Rightarrow$	$h_2(4)$	2	2	2	2
	$h_3(4)$	0	1	4	0

signature matrix

	S_1	S_2	S_3	S_4
$h_1(5)$	5	1	1	1
$\Rightarrow$ $h_2(5)$	2	2	2	2
$h_3(5)$	1	1	4	0

2) h_3 is true permutation

3)

	1-2	1-3	1-4	2-3	2-4	3-4
col	0	0	1/4	0	1/4	1/4
sig	1/3	1/3	2/3	2/3	2/3	2/3

b) a can be random number
c must be prime numbe and $c > N$

c) If 2 columns has no element in common, then assuming there are no hash collisions, the possibility of hashed values are equal is 0.

P3

a) $P(sim(c_1, c_2)) = (0.9)^{10}$
false negative : $(1 - (0.9)^{10})^{10} = 0.0137$

b) 1) map : hash (signature) into bucket as key and element as value
reduce : bucket as key and list of all elements as value

P4

m = length of input S
$n = 6 * m$
$k = n/m \ln(2) = 4.16$

P5

Median = 296,397,483