ADM - Homework 5

2.3.1

(a)
```
map( key, value):
// key: Name of document, value: integer
        max = -∞
        for i in value:
                if max < i:
                        max = i
        emit(1, max)

reduce ( key, values):
// key: 1 ; values: list of maxes
        max = -∞
        for value in values:
                if value > max:
                        max = value
        emit(1, max)
```

(b) Average:
```
map( key, value):
// key: Filename, value: list of int
        sum = 0
        count = 0
        for i in value:
                sum = sum + i
                count = count + 1
        emit( sum, count)
```

```
reduce(key, values):
// key: list of sums, values: counts
    avg = 0
    for c1, c2 in combinators(value):
        avg = sum / count

    emit(avg)
```

c)
```
map(key, value):
// key: filename, value: list of int
    for i in value:
        emit(key, 1)
```

```
reduce(key, values):
// key: word in file int, values: count
    for i in value:
        if len(i) == 1
            emit(key, count)
```

d)
```
map(key, value):
// key: filename , value: list of int
    for i in value:
        emit(key, 1)

reduce(key, values):
    max_ct = 0
    for count in values:
        max_ct += count
    emit(max_ct)
```

4.3.2 Memory = n bits ; Set = S, no. of members in $S = m$,

Case 1: K- Hash functions

False positive = Probability that all the targets are missed, and undesired object comes through.

$P(\text{Bit converts to } 1) = \frac{1}{n} \Rightarrow P(\text{No bit converts to } 1) = \left(1-\frac{1}{n}\right)$

So, if 'm' of them are missed,

$P(\text{None of } m \text{ convert to } 1) = \left(1-\frac{1}{n}\right)^m$

Now, since there are k- hash functions, each of 'm' gives k values
So,

$$P(\text{None of } Km \text{ convert to } 1) = \left(1-\frac{1}{n}\right)^{km} \approx e^{-km/n} \quad \left(\because (1-\varepsilon)^{1/\varepsilon} = e^{-1}\right)$$

So, $P(\text{Atleast } 1 \text{ converts to } 1)$ = False positive rate
$= \left(1-e^{-km/n}\right)^k$

Case 2 :- K- arrays of 'n' and 1 hash function

So, n gets replaced by $\left(\frac{n}{k}\right)$ and km gets replaced by m.

By replacing,

$$P(\text{None converted to } 1) = \left(1-\frac{1}{(n/k)}\right)^m$$

$$= \left(\left(1-\frac{1}{(n/k)}\right)^{(n/k)}\right)^{\left(\frac{k}{n}\right)\cdot m} \approx e^{-km/n}$$

So, $P(\text{Atleast } 1 \text{ converts to } 1) = \left(1-e^{-km/n}\right) \times \left(1-e^{-km/n}\right) \ldots$
(since AND cond^n)

$\Rightarrow P(\text{Atleast } 1 \text{ converts to } 1) = \left(1-e^{-km/n}\right)^k$

Hence Proved.

4.3.3 $\quad f(n) = \left(1 - e^{-km/n}\right)^{k}$ minimize $f(n)$ w.r.t 'k'

Taking log on both sides,

$$\log f(n) = k \log\left(1 - e^{-km/n}\right)$$

Minimize wrt 'k' by differentiating both sides by 'k'.

So,

$$\frac{d}{dk}\left(\log f(n)\right) = \frac{d}{dk}\left(k \log\left(1 - e^{-km/n}\right)\right)$$

$$\Rightarrow \frac{1}{f(n)} \cdot f'(n) = k \cdot \left[\frac{1}{\left(1 - e^{-km/n}\right)} \cdot +e^{-km/n} \cdot \left(+\frac{m}{n}\right)\right] + \log\left(1 - e^{-km/n}\right)$$

$$\Rightarrow \frac{1}{f(n)} f'(n) = \left(\frac{km}{n}\right) \frac{e^{-km/n}}{1 - e^{-km/n}} + \log\left(1 - e^{-km/n}\right)$$

Let $\frac{km}{n} = t \Rightarrow \frac{m}{n} = \frac{dt}{dk}$

$$\therefore \frac{1}{f(n)} f'(n) = \frac{t \cdot e^{-t}}{1 - e^{-t}} + \log\left(1 - e^{-t}\right)$$

$$\therefore f'(n) = f(n) \cdot \left[\frac{te^{-t}}{1 - e^{-t}} + \log\left(1 - e^{-t}\right)\right]$$

Equating $f'(n)$ to '0',

Trivial solⁿ :- $f(n) = 0$ $\quad$ <u>Reject</u>

The equation is:- $d\left(\log\left(1 - e^{t}\right)\right) = \log\left(1 - e^{-t}\right)$

Optimal solution for the form is $t = \log(2)$

$$\therefore \quad K = \frac{n}{m} \log(2)$$

4.4.1 Stream = [3, 1, 4, 1, 5, 9, 2, 6, 5]

① $h(x) = 2x+1 \mod 32$

$h(3) = 7$; $h(1) = 3$; $h(4) = 9$; $h(1) = 3$

$h(5) = 11$; $h(9) = 19$; $h(2) = 5$; $h(6) = 13$; $h(5) = 11$

② $h(x) = 3x+7 \mod 32$

$h(3) = 16$; $h(1) = 10$; $h(4) = 19$; $h(1) = 10$

$h(5) = 22$; $h(9) = 2$; $h(2) = 13$; $h(6) = 25$

$h(5) = 22$

③ $h(x) = 4x \mod 32$

$h(3) = 12$; $h(1) = 4$; $h(4) = 16$; $h(1) = 4$

$h(5) = 20$; $h(9) = 4$; $h(2) = 8$; $h(6) = 24$

$h(5) = 20$

Binary values:-

① 7: 00111 , 3: 00011 , 9: 01001 , 3: 00011

11: 01011 , 19: 10011 , 5: 00101 , 13: 01101

11: 01011

[No trailing zeros]

$\Rightarrow$ Unique $= 2^0 = 1$

② 16: 10000 , 10: 01010 , 19: 10011 , 10: 01010 ,

22: 10110 , 2: 00010 , 13: 01101 , 25: 11001 ,

22: 10110

[4 trailing zero's]

$\Rightarrow$ unique $= 2^4 = 16$

③ 12: 01100 , 4: 00100 , 16: 10000 , 4: 00100 ,

20: 10100 , 4: 00100 , 8: 01000 , 24: 11000 ,

20: 10100

$2^4 = 16$

4 · 5 · 3

$[3, 1, 4, 1, 3, 4, 2, 1, 2]$

Starting posⁿ = 1

$\Rightarrow$ X_1 . ele = 3 , X_1 . val = 1
X_2 . ele = 1 , X_2 . val = 1
X_3 . ele = 4 , X_3 . val = 1
X_2 . ele = 1 , X_2 . val = 2
X_1 . ele = 3 , X_1 . val = 2
X_3 ele = 4 , X_3 . val = 2
X_4 . ele = 2 , X_4 . val = 1
X_2 . ele = 1 , X_2 . val = 3
X_4 . ele = 2 , X_4 . val = 2

Actual moment =

$$\sum m_i^2 = (2)^2 + (3)^2 + (2)^2 + (2)^2$$

$$= 12 + 9 = 21$$

Alon- Matias :-

$$E(n(2X \cdot val - 1))$$

$$= \frac{1}{n} \sum_{i=1}^{n} (n(2c_i - 1))$$

$$= \sum 2c_i - 1$$

$$= 3 + 5 + 3 + 3 = 14$$

Starting posⁿ 2:

$\Rightarrow$ X_1 ele = 1 , X_1 . val = 1
X_2 . ele = 4 , X_2 . val = 1
X_1 . ele = 1 , X_1 . val = 2
X_3 ele = 3 , X_3 . val = 1
X_2 . ele = 4 , X_2 . val = 2
X_4 . ele = 2 , X_4 . val = 1
X_1 . ele = 1 , X_1 . val = 3
X_4 . ele = 2 , X_4 . val = 2

Alon- Matias :-

$$\sum 2c_i - 1$$

$$5 + 3 + 1 + 3 = 12$$

Starting posⁿ 3 :-

X_1 . ele = 4 , X_1 . val = 1
X_2 . ele = 1 , X_2 . val = 1
X_3 . ele = 3 , X_3 . val = 1
X_4 . ele = 4 , X_1 . val = 2
X_4 . ele = 2 , X_4 . val = 1
X_2 . ele = 1 , X_2 . val = 2
X_4 . ele = 2 , X_4 . val = 2

$[3, 1, 4, 1, 3, 4, 2, 1, 2]$

$$\sum 2c_i - 1$$

$$= 3 + 3 + 1 + 3$$

$$= 10$$

Starting posⁿ 4 :-

X_1 . ele = 1 , X_1 . val = 1
X_2 . ele = 3 , X_2 . val = 1
X_3 . ele = 4 , X_3 . val = 1
X_4 . ele = 2 , X_4 . val = 1
X_1 . ele = 1 , X_1 . val = 2
X_4 . ele = 2 , X_4 . val = 2

$$\sum 2c_i - 1$$

$$= 3 + 1 + 1 + 3$$

$$= 8$$

Starting pos^n : 5 $[3, 1, 4, 1, 3, 4, 2, 1, 2]$

$X_1 \cdot ele = 3$, $X_1 \cdot val = 1$

$X_2 \cdot ele = 4$, $X_2 \cdot val = 1$

$X_3 \cdot ele = 2$, $X_3 \cdot val = 1$

$X_4 \cdot ele = 1$, $X_4 \cdot val = 1$

$X_3 \cdot ele = 2$, $X_3 \cdot val = 2$

$$\Sigma 2c_i - 1 = 1 + 1 + 3 + 1 = 6$$

Starting pos^n : 6

$X_1 \cdot ele = 4$, $X_1 \cdot val = 1$

$X_2 \cdot ele = 2$, $X_2 \cdot val = 1$

$X_3 \cdot ele = 1$, $X_3 \cdot val = 1$

$X_2 \cdot ele = 2$, $X_2 \cdot val = 2$

$$\Sigma 2c_i - 1 = 1 + 3 + 1 = 4$$

Starting pos^n : 7

$X_1 \cdot ele = 2$, $X_1 \cdot val = 1$

$X_2 \cdot ele = 1$, $X_2 \cdot val = 1$

$X_1 \cdot ele = 2$, $X_1 \cdot val = 2$

$$\Sigma 2c_i - 1 = 3 + 1 = 4$$

Starting pos^n : 8

$X_1 \cdot ele = 1$, $X_1 \cdot val = 1$

$X_2 \cdot ele = 2$, $X_2 \cdot val = 1$

$$\Sigma 2c_i - 1 = 1 + 1 = 2$$

starting pos^n : 9

$X_1 \cdot ele = 2$, $X_1 \cdot val = 1$ $\quad \Sigma 2c_i - 1 = 1$

3.2.1

First Ten 3-Shingles :-

1. The most effective
2. most effective way
3. effective way to
4. way to represent
5. to represent documents
6. represent documents as
7. documents as sets
8. as sets for
9. sets for the
10. for the purpose

3.3.3

Element	S_1	S_2	S_3	S_4	$h_1(x) = 2x+1 \mod 6$	$h_2(x) = 3x+2 \mod 6$	$h_3(x) = 5x+2 \mod 6$
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

Initial :-

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

Element 0 :

Hash	S_1	S_2	S_3	S_4
h_1	∞	1	∞	1
h_2	∞	2	∞	2
h_3	∞	2	∞	2

Element 1:

Hash	S_1	S_2	S_3	S_4
h_1	∞	1	∞	1
h_2	∞	2	∞	2
h_3	∞	1	∞	2

Element 2:

Hash	S_1	S_2	S_3	S_4
h_1	5	1	∞	1
h_2	2	2	∞	2
h_3	0	1	∞	0

Element 3:

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

Element 4:

Hash	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

Element 5:

Hash	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) $h_3(x) = (5x+2) \bmod 6$ provides us a random output, and therefore is the true permutation.

c)

Columns	$S_1 S_2$	$S_1 S_3$	$S_1 S_4$	$S_2 S_3$	$S_2 S_4$	$S_3 S_4$
Jaccard Similarity	0	0	0.3	0	0.3	0
Signature Matrix Similarity	0.3	0.3	0.67	0.3	0.67	0.67

3.3.6

	S_1	S_2
0	0	1
1	1	1
2	0	0

Case 1 :-

$S_1 = [0, 1, 0]$, $S_2 = [1, 1, 0]$

Let $h(x) = 3x + 1 \mod 2$

$h(S_1) = [1, 0, 1]$, $h(S_2) = [0, 0, 1]$

Case 2 :

$S_1 = [0, 0, 1]$, $S_2 = [0, 1, 1]$

$h(S_1) = [1, 1, 0]$, $h(S_2) = [1, 0, 0]$

Case 3 :

$S_1 = [1, 0, 0]$, $S_2 = [1, 0, 1]$

$h(S_1) = [0, 1, 1]$, $S_2 = [0, 1, 0]$

So, cyclically, Jaccard similarity differs with the seed value, be it the same data.

3.4.4

```
map ( key , value ) :
// key : Filename ; value : element of signature matrix
    split item to bands
    for band in bands :
        for signature in band :
            bucket = hash( signature )
        emit( bucket , value )

reduce ( key , values ) :
// key : bucket , values : elements of signature matrix corresponding to bucket
    emit ( key , values )
```