

EE412 Foundation of Big Data Analytics, Fall 2018

HWx

Name: 김경만

Student ID: 20150073

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

Answer to Problem 1

Exercise 2.2.1

- (a) significant skew 가 있을 것이다. combiner 가 없다면 key에 따른 data 양이 1개
크게 불균등하게 분포되고, 이에 따라 skew 가 생긴다.
- (b) combiner의 영향으로 reducer들이 10개 정도의 reduce task를 맡게 되면
비슷한 처리시간을 갖게 되어 skew가 줄어든다. 10000개의 reduce task로
합치는 경우에는 조금 더 크게 불균등한 처리시간을 갖게 되므로 더욱 skew가 줄어든다.
- (c) key 별로 1개의 value를 갖게 되기 때문에 reducer에서 처리할 때
비슷한 처리시간이 요구되어 skew가 줄어든다.

Exercise 2.3.3

- (a) bag union $\begin{cases} \text{map: } R \text{과 } S \text{의 tuple } t \text{를 } (t, 1) \text{로 만든다.} \\ \text{reduce: 같은 key 값을 갖는 pair를 합한 후, 각 key의 tuple } t \text{를 value 만큼 생성한다.} \end{cases}$
- (b) bag intersection $\begin{cases} \text{map: } R \text{과 } S \text{의 tuple } t \text{를 } R \text{이면 } (t, (1, 1)), S \text{면 } (t, (0, 1)) \text{로 만든다.} \\ \text{reduce: 같은 key 값을 갖는 pair를 합한 후, 각 key의 tuple } t \text{를 value 2개 중 최솟값만큼 생성한다.} \end{cases}$
- (c) bag difference $\begin{cases} \text{map: } R \text{과 } S \text{의 tuple } t \text{를 } R \text{이면 } (t, 1), S \text{면 } (t, -1) \text{로 만든다.} \\ \text{reduce: 같은 key 값을 갖는 pair를 합한 후, value가 양수인 경우만 각 key의 tuple } t \text{를 value 만큼 생성한다.} \end{cases}$

Exercise 2.4.1

task 가 속도 P 의 실행으로 실패하기 때문에 총 t 초가 걸리는 task 가 실패하는 실행은 Pt 이다. 복구에 $10t$ 초의 overhead 가 있으므로 $10t \cdot Pt$ 만큼의 추가 시간이 필요할 수 있다. 따라서 task 에 필요한 총 시간은 $(t + 10t \cdot Pt)$ 초이며 n 개의 task 가 있으므로 total expected execution time 은 $nt(1 + 10Pt)$ 초이다.

Answer to Problem 2

Exercise 6.1.1

- (a) 100 보다 작거나 같은 수의 약수로 5번 이상 나오야 함으로 $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ 이다.
- (b) 위의 집합 중 (a, b) 라면 a와 b의 최소공배수가 20 보다 작거나 같아야 한다. 그러므로
 $\{(1, 2), (1, 3), \dots, (1, 20), (2, 3), (2, 4), \dots, (2, 10), (2, 12), (2, 14), (2, 16), (2, 18), (2, 20),$
 $(3, 4), (3, 5), (3, 6), (3, 9), (3, 12), (3, 15), (3, 18), (4, 5), (4, 6), (4, 8), (4, 10), (4, 12),$
 $(4, 16), (4, 20), (5, 10), (5, 15), (5, 20), (6, 9), (6, 12), (6, 18), (7, 14), (8, 16), (9, 18), (10, 20)\}$
 이다.
- (c) 모든 basket 의 각 item 의 총 합은 모든 basket에서 각 item의 총 등장 횟수의 합과 같다.
 1은 100번, 2는 50번, 3은 33번, 4는 25번, 5는 20번, 6은 16, 7은 14, 8은 12, 9는 11,
 10은 10, 11은 9, 12는 8, 13~14는 7, 15~16은 6, 17~20은 5, 21~25는 4,
 26~33은 3, 34~50은 2, 51~100은 1번 나오므로 다 더하면 482 이다.

Exercise 6.2.3

- (a) Item이 I개면 $I C_2$ 개의 pairs 가 있으므로 $4 \times I C_2 = 2 \cdot I \cdot (I-1)$ bytes 필요하다.
- (b) 모든 basket에서 $K C_2$ 개의 pairs가 있으므로 만약 모든 서로 다른 pairs들이 한번씩 등장한다면
 largest possible number of pairs 가 되어 $K C_2 \times B$ 개의 pairs가 있을 것이다.
 $\therefore \frac{B \cdot K \cdot (K-1)}{2} \geq H$
- (c)
 (c) triple $\Rightarrow \frac{B \cdot K \cdot (K-1)}{2} \cdot 12 = 6 \cdot B \cdot K \cdot (K-1)$ bytes,
 triangular $\Rightarrow 2 \cdot I \cdot (I-1)$ bytes 이므로 $3 \cdot B \cdot K \cdot (K-1) < I \cdot (I-1)$ 이므로
 triple이 memory를 더 적게 쓴다.

Exercise 6.2.7

각 item number \leftrightarrow number for frequent item 과 hash table 을 위한 memory 를 무시하면,
 triangular 는 $4 \cdot N C_2 = 2 \cdot N \cdot (N-1)$ bytes 가 필요하며 hash table 을 사용 하면
 frequent pairs 10^6 개와 frequent 하진 않지만 frequent item으로 이루어진 쌍 M 개가 있으므로
 $12 \cdot (10^6 + M)$ bytes 가 필요하다. 그러므로 필요한 공간은 $\min(2N(N-1), 12(10^6 + M))$ 이다

Answer to Problem 3

Exercise 3.3.2

Row	s_1	s_2	s_3	s_4	$7x+1 \bmod 5$	$3x+1 \bmod 5$	$2x+4 \bmod 5$	$3x-1 \bmod 5$
0	1	0	0	1	1	1	4	4
1	0	0	1	0	2	4	1	2
2	0	1	0	1	3	2	3	0
3	1	0	1	1	4	0	0	3
4	0	0	1	0	0	3	2	1

2. 0 1 2 3 4 Row 0 1 2 3 4
1 2 3 4 0

h_i	s_1	s_2	s_3	s_4		s_1	s_2	s_3	s_4		s_1	s_2	s_3	s_4		s_1	s_2	s_3	s_4
h_1	∞	∞	∞	∞		1	∞	∞	1		1	∞	2	1		1	3	2	1
h_2	∞	∞	∞	∞	Row 0	1	∞	∞	1	Row 1	1	∞	4	1	3	1	2	4	1
h_3	∞	∞	∞	∞	\rightarrow	4	∞	∞	4	\rightarrow	4	∞	1	4	4	4	3	1	3
h_4	∞	∞	∞	∞		4	∞	∞	4		4	∞	2	4	4	4	0	2	0
Row 3	s_1	s_2	s_3	s_4		s_1	s_2	s_3	s_4		s_1	s_2	s_3	s_4		s_1	s_2	s_3	s_4
\rightarrow	1	3	2	1	Row 4	1	3	0	1	\therefore	h_1	1	3	0	1				
	0	2	0	0		0	2	0	0		h_2	0	2	0	0				
	0	3	0	0		0	3	0	0		h_3	0	3	0	0				
	3	0	2	0		3	0	1	0		h_4	3	0	1	0				

Exercise 3.4.2

$$1 - (1 - s^r)^b = \frac{1}{2} \text{ or } s = \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{b}}\right)^{\frac{1}{r}} \text{ or } \text{etc.}$$

$$\cdot r=3, b=10 \Rightarrow s = \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{10}}\right)^{\frac{1}{3}} = 0.4061, \left(\frac{1}{10}\right)^{\frac{1}{3}} = 0.4642 \text{ or } \text{etc.}$$

$$\cdot r=6, b=20 \Rightarrow s = \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{20}}\right)^{\frac{1}{6}} = 0.5694, \left(\frac{1}{20}\right)^{\frac{1}{6}} = 0.6070 \text{ or } \text{etc.}$$

$$\cdot r=5, b=50 \Rightarrow s = \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{50}}\right)^{\frac{1}{5}} = 0.4244, \left(\frac{1}{50}\right)^{\frac{1}{5}} = 0.4573 \text{ or } \text{etc.}$$

Exercise 3.6.1

$$(a) p \rightarrow p^2, p^2 \rightarrow 1 - (1 - p^2)^3 \text{ or } \text{etc.} \therefore 1 - (1 - p^2)^3$$

$$(b) p \rightarrow 1 - (1 - p)^3, 1 - (1 - p)^3 \rightarrow (1 - (1 - p)^3)^2 \text{ or } \text{etc.} \therefore (1 - (1 - p)^3)^2$$

$$(c) p \rightarrow p^2, p^2 \rightarrow 1 - (1 - p^2)^2, 1 - (1 - p^2)^2 \rightarrow (1 - (1 - p^2)^2)^2 \text{ or } \text{etc.} \therefore (1 - (1 - p^2)^2)^2$$

$$(d) p \rightarrow 1 - (1 - p)^2, 1 - (1 - p)^2 \rightarrow (1 - (1 - p)^2)^2, (1 - (1 - p)^2)^2 \rightarrow 1 - (1 - (1 - (1 - p)^2)^2)^2, \\ 1 - (1 - (1 - (1 - p)^2)^2)^2 \rightarrow (1 - (1 - (1 - (1 - p)^2)^2)^2)^2 \text{ or } \text{etc.} \therefore (1 - (1 - (1 - (1 - p)^2)^2)^2)^2$$