IS 2012. Dandy DENG

DB (Q1)

(a) select distinct sname
from STUDENT, ENROLLED, OFFERING, PROFESSOR
where S.sid = E.sid and E.offid = O.offid and O.pid = P.pid
and P.pname = "Michael" and O.semester = "Spring" and O.year = "2012"

(b) select cname
from CLASS, OFFERING
where C.cid = O.cid and O.semester = "Spring" and O.year = "2012"
group by cid having count (distinct pid) >= 2

DB (Q2)

The selectivity of age > 55 AND city = "BOSTON" on CUSTOMERS is 30% × 20% = 6%

The selectivity of city = "Boston" on BRANCHES is 10%

The selectivity of balance > 100,000 on ACCOUNTS is 30%

The estimated size of $\sigma_{age>55 \wedge city=\text{"BOSTON"}}$ CUSTOMERS is 10,000 × 6% = 600 tuples

The estimated size of $\sigma_{city=\text{"Boston"}}$ BRANCHES is 2000 × 10% = 200 tuples.

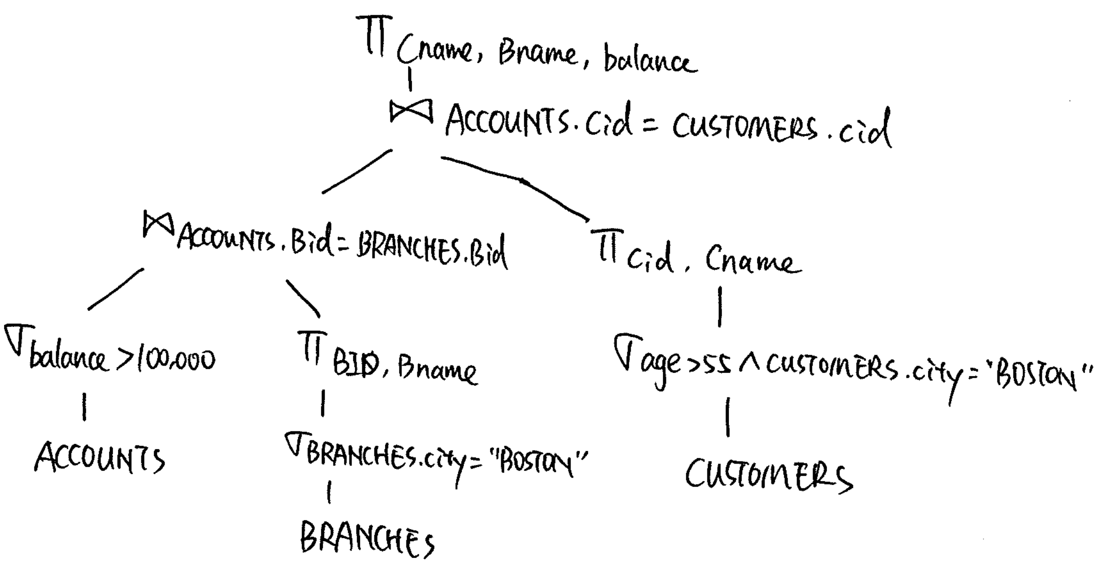
The estimated size of $\sigma_{balance>100,000}$ ACCOUNTS is 12,000 × 30% = 3600 Tuples

The estimated size of CUSTOMERS ⋈ ACCOUNTS is (600 × 3600)/600 = 3600 tuples

The estimated size of CUSTOMERS ⋈ BRANCHES is 600 × 200 = 120000 tuples

The estimated size of ACCOUNTS ⋈ BRANCHES is (200 × 3600)/200 = 3600 tuples

Thus, we can join CUSTOMERS and ACCOUNTS first or BRANCHES and ACCOUNTS.



IS2012 (cont.)

DM(Q3) (a) $O(n^2)$ is required to compute the distance matrix. After that, there are n-1 iterations involving steps 3 and 4 because there are n clusters at the start and two clusters are merged during each iteration.

(b) The distance from each cluster to all other clusters are stored as a sorted list, it can reduce the cost of finding the two closest clusters to $O(n-i+1)$.

OS(Q6) (a)

Need

	A	B	C	D
P0	0	0	1	2
P1	1	6	8	0
P2	2	3	5	6
P3	0	6	5	2
P4	0	6	5	6

∴ The system is a safe state

Possible safe sequences:

(1) P0 → P3 → P1 → P2 → P4

(2) P0 → P3 → P1 → P4 → P2

(3) P0 → P3 → P4 → P1 → P2

(b) V = (1, 6, 2, 2)

① P0 → V = (1, 6, 5, 4)

② P3 → V = (1, 6, 8, 6)

③ { P1 → V = (2, 6, 8, 6) (i)
   { P4 → V = (1, 6, 9, 10) (ii)

④ (i) { P2 → V = (3, 9, 13, 10) (iii)
       { P4 → V = (2, 6, 9, 10) (iv)

(ii) P1 → V = (2, 6, 9, 10) (v)

⑤ (iii) P4 (iv) P2 (v) P2

IS2013

DB(Q1)

(a) (1) $\Pi_{tname, year}(\sigma_{pname=\text{"Michael Jordan"}}(PLAYER \bowtie TEAM \bowtie REGISTER))$

(2) $\Pi_{pname}(\sigma_{tname=\text{"Heat"} \wedge year=2013}(PLAYER \bowtie TEAM \bowtie REGISTER))$

(b) (1)
```
select pid, MIN(year), MAX(year)
from REGISTER
group by pid
```

(2)
```
select pid from REGISTER
where year >= 1996 and year <= 2005
group by pid
having count(distinct year) = 10
```

IS2013 (cont.)

DB(Q2) (a) R: $4 \times 10 \times 20000 / 4000 = 200$ pages

S: $2 \times 10 \times 36000 / 4000 = 180$ pages

(b) Step 1: b_R (read) + $b_R (+n_h)$ (write) = 400 (410)

Step 2: b_S (read) + $b_S (+n_h)$ (write) = 360 (370)

Step 3: $b_R (+n_h)$ (write) + $b_S (+n_h)$ (write) = 380 (400)

Total: $400 + 360 + 380 (+40) = 1140$ (1180)

(c) $3 \times (9 \times \frac{180}{10} + 9 \times \frac{200}{10}) + \frac{180}{10} + \frac{200}{10} = 1064$

DM(Q3) (a) Frequent Itemsets: {A}, {B}, {C}, {D}, {E}, {AB}, {AC}, {AD}, {AE}, {BC}, {BD}, {BE}, {CD}, {CE}, {DE}, {ABC}, {ABD}, {ABE}, {ACD}, {ACE}, {ADE}, {BCD}, {BCE}, {BDE}, {CDE}. All are closed itemsets. All 3-itemsets are maximal.

(b) (1) $S(\text{"coffee} \rightarrow \text{milk"}) = \frac{1000}{5000} = 0.2$, $C(\text{"coffee} \rightarrow \text{milk"}) = \frac{1000}{3000} = \frac{1}{3}$

(2) $S(\text{"}\overline{\text{coffee}} \rightarrow \text{milk"}) = \frac{1500}{5000} = 0.3$, $C(\text{"}\overline{\text{coffee}} \rightarrow \text{milk"}) = \frac{1500}{2000} = 0.75$

(3) $\text{Interest (coffee, milk)} = \frac{P(\text{coffee, milk})}{P(\text{coffee}) \cdot P(\text{milk})} = \frac{0.2}{0.6 \times 0.5} = \frac{2}{3} < 1$. negatively correlated.

OS(Q6) (a)

1	1	1	1	4	4	4	4	1
	2	2	2	2	2	5	5	5
			3	3	3	3	2	2
F	F		F	F		F	F	F

7 page faults.

(b)

1	1	1	1	1	1	5	5	5
	2	2	2	4	4	4	4	1
			3	3	3	3	2	2
F	F		F	F		F	F	F

7 page faults

(c)

1	1	1	1	1	1	1	1	1
	2	2	2	2	2	2	2	2
			3	4	4	5	5	5
F	F		F	F		F		

5 page faults

IS2014

DB(Q1)

IS 2014 (cont.)

DB (Q2) (a) $br = 30000/30 = 1000$, $bs = 1000/5 = 200$. Cost : $br + bs = 1200$

(b) cost : $\lceil \frac{br}{M-2} \rceil \times bs + br = 201000$ (c) cost : $\lceil \frac{br}{M-2} \rceil \times bs + br = 11000$

(d) We can replace record scans with more efficient index lookups.

DM (Q3) (a) $\{A\}: 8$, $\{B\}: 7$, $\{C\}: 7$, $\{D\}: 6$, $\{AB\}: 5$, $\{AC\}: 5$, $\{AD\}: 4$,
$\{BC\}: 5$, $\{CD\}: 4$

(b) $S(A \to B) = \frac{5}{10} = 0.5$, $C(A \to B) = \frac{5}{8}$, $Lift(A \to B) = \frac{P(B|A)}{P(B)} = \frac{5/8}{7/10} = \frac{25}{28}$

DM (Q4) (a)

A	+	−
T	4	1
F	1	4

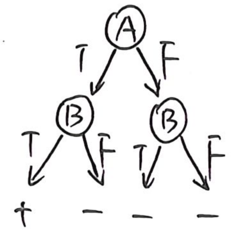
B	+	−
T	3	3
F	2	2

$GINI_A = \frac{1}{2}[1-(\frac{1}{5})^2-(\frac{4}{5})^2] + \frac{1}{2}[1-(\frac{1}{5})^2-(\frac{4}{5})^2] = \frac{8}{25}$

$GINI_B = \frac{3}{5}[1-(\frac{1}{2})^2-(\frac{1}{2})^2] + \frac{2}{5}[1-(\frac{1}{2})^2-(\frac{1}{2})^2] = \frac{1}{2}$

$\because GINI_A < GINI_B$ $\therefore$ A should be chosen.

(b)



(c) Confusion Matrix:

		Predicted	Class
Actual		+	−
Class	+	3	2
	−	0	5

Accuracy $= \frac{3+5}{10} = 0.8$

Precision $= \frac{3}{3+0} = 1$

Recall $= \frac{3}{3+2} = 0.6$

$F_1 = \frac{2RP}{R+P} = 0.75$

OS (Q5)

(a) Need

	A	B	C	D	E
P0	0	0	2	0	0
P1	0	7	5	0	1
P2	6	6	2	2	2
P3	2	0	0	2	3
P4	0	3	2	0	0
P5	1	0	1	2	6

(b) $V = (2,1,2,0,0)$

① $P0 \to V = (2,1,3,2,3)$

② $P3 \to V = (4,4,8,6,3)$

③ $P4 \to V = (4,7,11,8,3)$

④ $P1 \to V = (6,7,11,8,4)$

⑤ $P2 \to V = (6,7,14,12,6)$

⑥ $P5 \to V = (9,9,18,12,6)$

The system is in a safe state.
$P0 \to P3 \to P4 \to P1 \to P2 \to P5$

(c) Available resources become
$V = (2,0,0,0,0)$.

Allocation $P2 \Rightarrow (0,1,5,4,2)$
Need $P2 \Rightarrow (6,5,0,2,2)$

Currently, no process can be satisfied. So it is not safe to grant P2 immediately.

OS (Q6) (a) $250 + 250 = 500$ ns

(b) $(500+20) \times 10\% + (250+20) \times 90\% = 295$ ns

(c) $8 - 5 = 3$ bits $\Rightarrow 2^3 = 8$ entries

(d)

IS 2015

DS (Q1)

(a)
```
def find (T, x):
    if T is None:
        return False
    else if T.value = x:
        return True
    elif T.value > x:
        return Find(T.left, x)
    elif T.value < x:
        return find(T.right, x)
```
Time complexity: $O(\log n)$

(b)
```
def find-pair(T, z):
    tmp = set()
    if not findpairUtil(T, z, tmp):
        return None

def findpair(T, z, tmp):
    if T is None:
        return False
    if findpairUtil(T.left, z, tmp):
        return True
    if z - T.value in tmp:
        print("%s, %s" % (T.value, z-T.value)
        return True
    else:
        tmp.add(T.value)
    return findpairUtil(T.right, z, tmp)
```
Time complexity: $O(n)$

Space complexity: $O(n)$

DS (Q2)

OS (Q3)

(a) (1)

	1	1	1	1	4	4	4	4	3
		2	2	2	2	2	5	5	5
			3	3	3	3	3	1	1
	F	F	F		F		F	F	F (7)

(2)

	1	1	1	1	1	1	1	1	1
		2	2	2	4	4	4	4	3
			3	3	3	3	5	5	5 (6)
	F	F	F		F		F		F

(3)

	1	1	1	1	1	1	1	1	1
		2	2	2	4	4	5	5	5
			3	3	3	3	3	3	3
	F	F	F		F		F		(5)

(b) $210 \times 0.9 + 410 \times 0.1 = 230$

(c) ① Paging, which is transparent to the programmer, eliminates external fragmentation and thus provides efficient use of main memory.

② Because the pieces that are moved in and out of main memory are of fixed, equal size, it is possible to develop sophisticated memory management algorithms that exploit the behaviours of the program.

DB (Q4) (a) select category from Book group by category having count (bid) >= 2000

(b) select title from Books, Borrow where Books.bid = Borrow.bid
group by bid having count (distinct sid) >= all
(select count (distinct sid) from Borrow group by bid)

(b) (1) $br = 600$, $bs = 1000$, $br + bs = 1600$ (2) $\lceil \frac{br}{3-2} \rceil \times bs + br = 601000$

(c) The estimate size of Strategy 1 and 2 are $6000 \times 20000 = 120000000$ and $30000 \times 6000 / 6000 = 30000$, respectively. Thus, choose Strategy 2.

IS 2015 (Cont.)

DM (Q5) (a) $E_{orig} = 1 - \max(\frac{50}{100}, \frac{50}{100}) = 0.5$

A	+	-
T	25	0
F	25	50

B	+	-
T	30	20
F	20	30

C	+	-
T	25	25
F	25	25

$E_{A=T} = 1 - \max(\frac{25}{25}, \frac{0}{25}) = 0$

$E_{A=F} = 1 - \max(\frac{25}{75}, \frac{50}{75}) = \frac{1}{3}$

$\Delta A = E_{orig} - \frac{25}{100} E_{A=T} - \frac{75}{100} E_{A=F} = 0.25$

$\Delta B = 0.1$

$\Delta C = 0$

$\therefore$ Choose A.

(b) Since the instances are all + when A = T, we only split A = F.

B	+	-
T	25	20
F	0	30

C	+	-
T	0	25
F	25	25

$E_{B=T} = 1 - \max(\frac{25}{45}, \frac{20}{45}) = \frac{20}{45}$

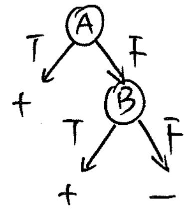
$E_{B=F} = 1 - \max(0, 1) = 0$

$\Delta B = E_{A=F} - \frac{45}{75} E_{B=T} = \frac{1}{15}$

$\Delta C = E_{A=F} - \frac{50}{75} E_{C=F} = 0$

$\therefore$ Choose B.

(c)



20 instances are misclassified.

(d) For C=T, $E_{C=T} = \frac{1}{2}$.

A	+	-
T	25	0
F	0	25

B	+	-
T	5	20
F	20	5

$\Delta A = \frac{1}{2}$

$\Delta B = \frac{3}{10}$

Choose A.

For C=F, $E_{C=F} = \frac{1}{2}$.

A	+	-
T	0	0
F	25	25

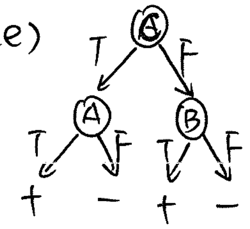
B	+	-
T	25	0
F	0	25

$\Delta A = 0$

$\Delta B = \frac{1}{2}$

Choose B.

(e)



0 instance is misclassified.

(f) The greedy heuristic does not necessarily lead to the best tree.

IR (Q6) (a) (1) Exercise 8.2 (2) Spam e-mail detection. ~~Since~~ We can sacrify some recall to obtain a better precision on detect the spam email, since we don't want to classify some important emails to be spam email.

(b) (1) TrainRocchio(C, D)

for each $c_j \in C$

do $D_j \leftarrow \{d : \langle d, c_j \rangle \in D\}$

$\vec{\mu}_j \leftarrow \frac{1}{|D_j|} \sum_{d \in D_j} \vec{v}(d)$

return $\{\vec{\mu}_1, \ldots, \vec{\mu}_j\}$

ApplyRocchio($\{\vec{\mu}_1, \ldots, \vec{\mu}_j\}, d$)

return $\arg\max_j \mathrm{CosSim}(\vec{\mu}_j, \vec{v}(d))$

(2) $\vec{w}_{ij} = \vec{\mu}(C_i) - \vec{\mu}(C_j)$

$b_{ij} = 0.5 * (|\vec{\mu}(C_i)|^2 - |\vec{\mu}(C_j)|^2)$

(c) Figure 14.5.

IS 2016

DS(Q1)

(a) Geeksforgeeks, BST, Ch41. (b) ⇐ the same

DS(Q2) Leetcode 064 Minimum Path Sum

OS(Q3)(a) Need

	A	B	C	D	E
P0	0	0	5	1	2
P1	0	2	1	5	2
P2	5	0	1	2	0
P3	1	2	0	0	2
P4	4	0	0	3	6

V = (1, 2, 0, 0, 3)

① P3 → V = (1, 2, 2, 6, 4)

② P1 → V = (2, 2, 5, 8, 4)

③ P0 → V = (6, 4, 5, 8, 8)

④ { P2 → V = (8, 9, 9, 11, 8) ⇒ P4
   P4 → V = (7, 5, 6, 8, 8) ⇒ P2

Safe.

P3 → P1 → P0 → P2 → P4

P3 → P1 → P0 → P4 → P2

(b) Seg # : 24 − 21 = 3 bits. Page# : 21 − 7 = 14 bits. Offset : 7 bits.

(c) (1) $\langle 8, 857 \rangle$ (2) $\langle 22, 989 \rangle$

DB(Q4) (a) (1) $\Pi_{cname}(COURSE) - \Pi_{cname}(ENROLL \bowtie COURSE)$

(2) $\Pi_{sname}(STUDENT \bowtie_{sid} G_{count(distinct\ cid) = G_{count(cid)}(\sigma_{pid="P123"}(TEACH))}(\sigma_{pid="P123"}(TEACH \bowtie ENROLL)))$

(b)(1) Select cname from STUDENT, ENROLL, COURSE
where S.sid = E.sid, E.cid = C.cid
group by cid having count (distinct dept) > 1

(2) select pname from PROFESSOR, ENROLL, TEACH
where P.pid = T.pid and T.cid = E.cid
group by cid having count (sid) >= all (select count (sid)
from ENROLL
group by cid)

(c) The estimated size :

$\sigma_{dept="SEEM"}$ STUDENT = 10000 × 20% = 2000

$\sigma_{credit=2}$ COURSE = 200 × 50% = 100

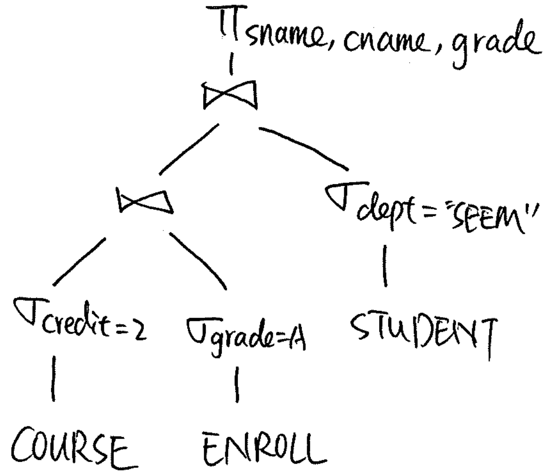
$\sigma_{grade=A}$ ENROLL = 100000 × 5% = 5000

STUDENT ⋈ ENROLL = 5000 × 2000 / 2000 = 5000

COURSE ⋈ ENROLL = 5000 × 100 / 100 = 5000

STUDENT ⋈ COURSE = 5000 × 100 = 500000

Thus, we should join STUDENT and ENROLL or COURSE and ENROLL first.



IS2016 (cont.)

DM(Q5) (a) Frequent itemsets: {M}:3, {C}:3, {O}:3, {Y}:4, {K}:6, {E}:4,
{M,K}:3, {C,K}:3, {O,K}:3, {O,E}:3, {E,K}:4, {Y,K}:4, {O,K,E}:3

Closed itemsets: {C,K}:3, {M,K}:3, {O,K,E}:3, {Y,K}:4, {E,K}:4, {K}:6

(b) $C(E \to C) = \frac{1}{4}$ $\quad lift(E \to C) = \frac{P(C|E)}{P(C)} = \frac{1/4}{1/2} = 0.5 < 1$. Negatively associated.

(c) $\because$ If $h(\{a_1, a_2, \dots, a_k\}) \geq min\_h$, then $h(\{a_1, a_2, \dots, a_{k-1}\}) \geq min\_h$.

$\therefore$ It is anti-monotone

(d)

IR(Q6) (a) (1) Minimum: $\min\{M_1, M_2\} + 1$, Maximum: $M_1 + M_2$

(2) Exercise 1.10. (b) ~~Exercise~~ Figure 14.5

(C) (1) Basic idf value is 0. (Exercise 6.9) (2) Exercise 6.12.

IS2017

DS(Q1)(a) Leetcode 131. Palindrome Partitioning (b) Leetcode 128. Longest Consecutive Seq.

OS(Q2) (a) (1) "full" is used to keep track of the number of items in the buffer. "empty" is used to keep track of the number of empty spaces. "mutex" is used to enforce mutual exclusion.

(2) In the Producer Process, "signal(empty)" should be "signal(full)".

(b) The system spends most of its time swapping pieces rather than executing instructions.

(c)

1	1	1	4	4	4	4	4	4	7	7	7	7	7	7	7	7	5	5	5	5	5
	2	2	2	5	5	5	1	1	1	8	8	8	8	8	8	8	8	4	4	4	4
		3	3	3	3	3	3	6	6	6	6	6	9	9	9	9	9	9	9	9	2
F	F	F	F	F			F	F	F	F			F				F	F			F

13 page faults.

DB(Q3) (a) $5000 \times 20000 / 5000 = 20000$, $1000 \times 5000 = 5000000$

(b) If CUSTOMER is outer relation, cost $= \left\lceil \frac{5000/100}{4-2} \right\rceil \times \frac{20000}{50} + \frac{5000}{100} = 10050$

Otherwise, cost $= \left\lceil \frac{400}{4-2} \right\rceil \times 50 + 400 = 10400$. $\therefore$ CUSTOMER should be outer relation.

(c) Cost $= b_r + n_r \times (h+1) = 50 + 5000 \times (3+1) = 20050$

IS2017 (cont.)

DM (Q4)

(a) 1. {1100, 1600} ⇒ {0, 200, 300, 900, 1100}, {1600}

2. {500, 1600} ⇒ {0, 200, 300, 900}, {1100, 1600}

3. {350, 1350} ⇒ {0, 200, 300}, {900, 1100, 1600} 4. $\{\frac{500}{3}, 1200\}$

(b) $SSE = (0-\frac{500}{3})^2 + (200-\frac{500}{3})^2 + (300-\frac{500}{3})^2 + (900-1200)^2 + (1100-1200)^2 + (1600-1200)^2$

$BSS = 3\times(\frac{500}{3}-\frac{2050}{3})^2 + 3\times(1200-\frac{2050}{3})^2$

(c) $SC(200) = 1-a/b = 1-[(200+100)/2]/[(700+900+1400)/3] = 0.85$

$SC(1100) = 1-a/b = 1-[(200+500)/2]/[(1100+900+800)/3] = 0.625$

(d) ①

	P1	P2	P3	P4	P5	P6
P1	0	200	300	900	1100	1600
P2	200	0	100	700	900	1400
P3	300	100	0	600	800	1300
P4	900	700	600	0	200	700
P5	1100	900	800	200	0	500
P6	1600	1400	1300	700	500	0

②

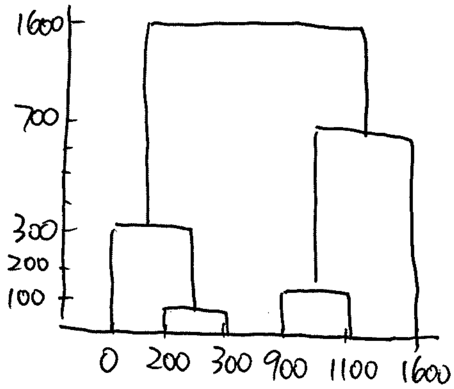
	P1	P2∪P3	P4	P5	P6
P1	0	300	900	1100	1600
P2∪P3	300	0	700	900	1400
P4	900	700	0	200	700
P5	1100	900	200	0	500
P6	1600	1400	700	500	0

③

	P1	P2∪P3	P4∪P5	P6
P1	0	300	1100	1600
P2∪P3	300	0	900	1400
P4∪P5	1100	900	0	700
P6	1600	1400	700	0

④

	P1∪P2∪P3	P4∪P5	P6
P1∪P2∪P3	0	1100	1600
P4∪P5	1100	0	700
P6	1600	700	0



IR (Q5) (a) (1) ① $\vec{u}_1 = \frac{1}{2}(\vec{d}_1+\vec{d}_2)$, $\vec{u}_2 = \frac{1}{2}(\vec{d}_3+\vec{d}_4)$ ② $\arg\min_j |\vec{u}_j - \vec{d}_t|$

(2) The set of points with equal distance from the two centroids. (3) Figure 14.5

(b)

IS2018

DS (Q1) (a) Trees. Ch78 + Ch80 (b)

OS (Q2) (a) (1) $2^{32} = 4GB$ (2) $2^{32}/2^{16} \cdot 2^8 = 2^{24}$, 24 bits (3) 0xEAC3E2F7 (4) 0x6F78C2

(b)

(c) FIFO: (8 page faults)

1	1	1	1	4	4	4	2	2	2
	3	3	3	3	1	1	1	1	6
		2	2	2	2	5	5	5	5
F	F	F		F	F	F	F		F

LRU: 8 page faults.

1	1	1	1	4	4	4	2	2	2
	3	3	3	3	3	5	5	5	5
		2	2	2	1	1	1	1	6
F	F	F		F	F	F	F		F

IS2018 (cont.)

DB(Q3) (a) $b_R = 3\times10\times18000/3000 = 180$, $b_S = 3\times10\times48000/3000 = 480$

(b) $\text{height} = \lceil \log_{\lceil \frac{20}{2} \rceil} 18000 \rceil = 5$, $\text{cost} = b_R + n_R \times (h+1) = 180 + 18000\times(5+1) = 108180$

(c) Step 1: $b_R(\text{read}) + b_R(+n_h)(\text{write}) = 360\ (366)$

Step 2: $b_S(\text{read}) + b_S(+n_h)(\text{write}) = 960\ (966)$

Step 3: $b_R(+n_h) + b_S(+n_h)(\text{write}) = \cancel{1320\ (1332)}\ 660\ (672)$

Total: $360 + 960 + 660\ (+24) = \cancel{2040\ (2064)}\ 1980\ (2004)$

(d) $180/6 < 36. \Rightarrow \text{cost} = 3\times(5\times\frac{180}{6} + 5\times\frac{480}{6}) + \frac{180}{6} + \frac{480}{6} = 1760$

DM (Q4) (a) ~~Eorig~~

A	+	−
T	4	0
F	1	5

B	+	−
T	2	3
F	3	2

$GINI_{A=T} = 1 - 1^2 - 0^2 = 0$

$GINI_{A=F} = 1 - (\frac{1}{6})^2 - (\frac{5}{6})^2 = \frac{10}{36}$

$GINI(A) = \frac{4}{10}\times0 + \frac{6}{10}\times\frac{10}{36} = \frac{1}{6}$

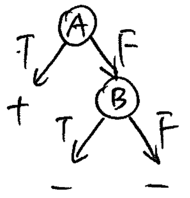
$GINI(B) = \frac{12}{25}$

$\therefore$ A should be chosen.

(b) Since instances are all + when A=T, we only split A=F.

B	+	−
T	0	3
F	1	2

$\Rightarrow$



(c) Confusion Matrix

		Predicted Class	
		+	−
Actual Class	+	4	1
	−	0	5

$\text{Accuracy} = \frac{4+5}{10} = 0.9$

$\text{Precision} = \frac{4}{4} = 1$

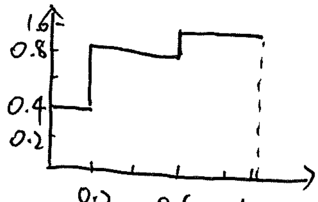
$\text{Recall} = \frac{4}{5} = 0.8$

$F_1 = \frac{2\times0.8}{1.8} = \frac{8}{9}$

(d)

class	−	−	+	−	−	+	+	−	+	+	
~~thre~~>=	0.35	0.40	0.45	0.5	0.7	0.75	0.8	0.82	0.85	0.9	1.0
TP	5	5	5	4	4	4	3	2	2	1	0
FP	5	4	3	3	2	1	1	1	0	0	0
TN	0	1	2	2	3	4	4	4	5	5	5
FN	0	0	0	1	1	1	2	3	3	4	5
TPR	1	1	1	0.8	0.8	0.8	0.6	0.4	0.4	0.2	0
FPR	1	0.8	0.6	0.6	0.4	0.2	0.2	0.2	0	0	0

ROC:



$AUC: \frac{0.2\times0.4 + 0.4\times0.8 + 0.4\times1}{1} = 0.8$

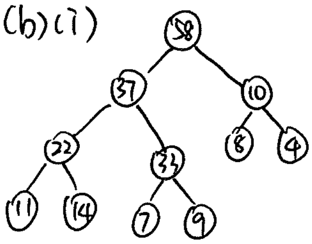
IR (Q5) (a) Doc1, Doc3. (b) P32. (c) x.

IS2019

DS(Q1)

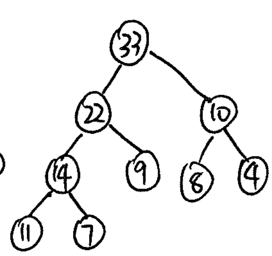
(a) BST. Ch01.

(b)(i)



O(log n)

(b)(ii)



(c)

47		2	10	9	12	19

16: 2, 10, 9, 12, 19, 47

47: 12, 19, 47

IS2019 (cont.)

OS (Q2) (a)

1	1	1	4	4	4	4	4	6	6	6	5	5	5	2	2	2	2	2	2
	2	2	2	5	5	5	5	5	~~8~~	7	7	7	9	9	9	5	5	5	9
		3	3	3	1	1	1	1	7	8	8	8	8	8	4	4	4	4	4
F	F	F	F	F	F			F	F	F	F		F	F	F	F			F

15 page faults.

(b) Need

	A	B	C	D	E
P1	3	0	2	1	4
P2	0	4	6	0	0
P3	0	0	7	3	1
P4	2	7	0	0	0
P5	3	2	2	0	0

$V = (3,4,3,1,1)$

① $P5 \rightarrow V = (4,5,6,1,2)$

② $P2 \rightarrow V = (5,5,9,3,2)$

③ $P3 \rightarrow V = (5,7,10,6,4)$

④ $\begin{cases} P1 \rightarrow V = (7,10,10,6,6) \rightarrow P4 \\ P4 \rightarrow V = (9,10,10,9,7) \rightarrow P1 \end{cases}$

Safe.

1°. P5, P2, P3, P1, P4

2°. P5, P2, P3, P4, P1

(c) $220 \times 0.8 + 420 \times 0.2 = 260$ ns

DB (Q3) (a) (i) $\Pi_{AName}(\sigma_{Gender=Female \wedge Profit>1000000}(ACTOR \bowtie ROLE \bowtie MOVIE))$

(ii) $_{AID}G_{count(MID),\,sum(Pay)}(\sigma_{Year>=2009 \wedge Year<=2019 \wedge Gender=Male}(ACTOR \bowtie ROLE \bowtie MOVIE))$

(b)(i) Select AName from ACTOR, ROLE, MOVIE
where ACTOR.AID = ROLE.AID and ROLE.MID = MOVIE.MID
and ACTOR.gender = Female and MOVIE.profit > 1000000.

(ii) Select AID, count(MID), sum(Pay) from ACTOR, ROLE, MOVIE
where ACTOR.AID = ROLE.AID and ROLE.MID = MOVIE.MID
and Gender = Male and Year >= 2009 and Year <= 2019
group by AID

(c)(i) ACTOR: 5000/100 = 50, ROLE: 100000/100 = 1000

If ACTOR is used as the outer relation, cost $= \lceil \frac{50}{4-2} \rceil \times 1000 + 50 = 25050$.

Otherwise, cost $= \lceil \frac{1000}{4-2} \rceil \times 50 + 1000 = 26000$. ∴ ACTOR should be used.

(ii) cost $= \lceil \frac{50}{60-2} \rceil \times 1000 + 50 = 1050$. $+ \; 2 \times \lceil \frac{50}{60-2} \rceil = 2$ seeks.

DM (Q4) (a) (i) Frequent itemsets: {a}:3, {b}:4, {c}:5, {d}:7, {e}:5, {b,d}:4, {b,e}:3, {c,d}:4, {d,e}:5, {b,d,e}:3. Closed: {a}, ~~{b}~~, {c}, {d}, {c,d}, {d,e}, {b,d}, {b,d,e}. Maximal: {b,d,e}, {c,d}, {a}

(ii) $s(b \rightarrow de) = \frac{3}{8}$ $\quad c(b \rightarrow de) = \frac{3}{4}$ $\quad lift(b \rightarrow de) = \frac{3/4}{5/8} = 1.2$

(b) (i) ① {10}, {20}, {30, 40, 50, 60} ⇒ (10, 20, 45)

② {10}, {20, 30}, {40, 50, 60} ⇒ (10, 25, 50)

IS2019 (cont.) DM(Q4)

(b)(iii) $SSE = 5^2+5^2+10^2+10^2 = 250$ $\quad BSS = 1\times25^2 + 2\times10^2 + 3\times15^2$

IR(Q5) (a)(i) $P = 2/5 = 0.4$, $R = 2/4 = 0.5$

(ii) $MAP = \frac{1}{4}\times(1+\frac{2}{3}+\frac{3}{9}+\frac{4}{10}) = 0.6$

(b)(i) Correct a single query term at a time

(ii) Compute the edit distance from q_e to each string in V, then select the string(s) of minimum edit distance.

(c)