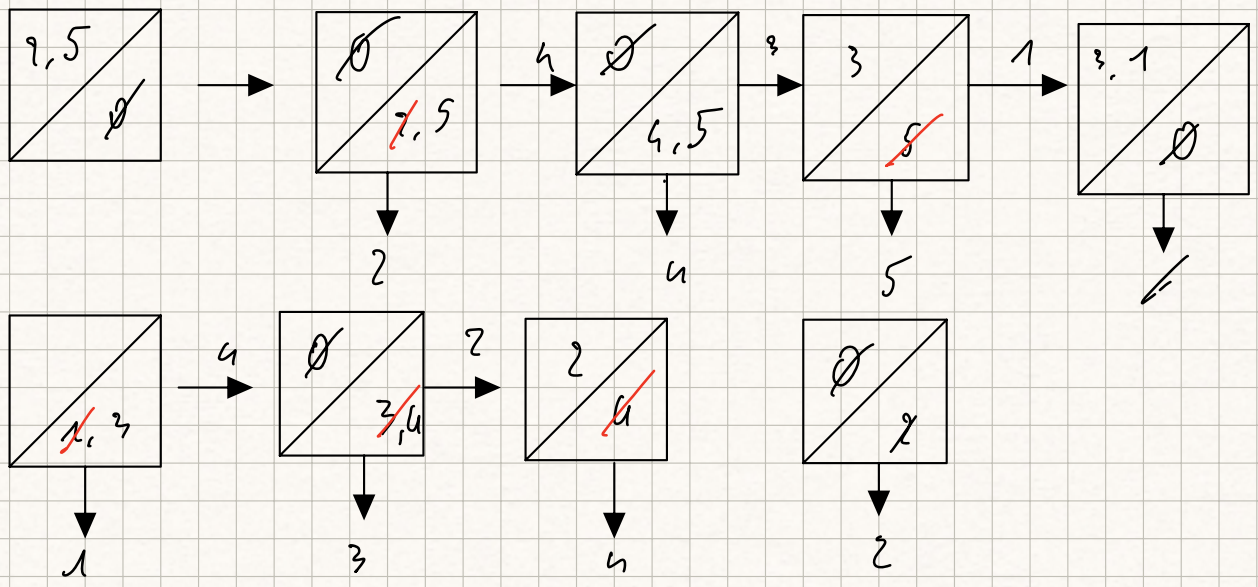
Simulate the algorithm SnowPlow over the sequence 2,5,4,3,1,4,2, and show which sorted blocks it forms with a memory of size M=2.



Given the ordered set of strings S = {AA, AC, BB, CC}. Compute the Minimal Ordered Perfect Hash for S by assuming the following two hash functions:

h1(xy) = x+y mod 7 and h2(xy) = x+2*y mod 7

in which x (resp. y) is the code of the first (resp. second) letter of a string of S, and the codes are: A=1, B=2, and C=3.

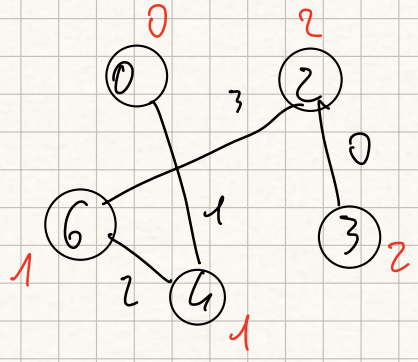
As an example, if the string is AC, then x=1 and y=3.

	h	h_1	h_2
AA	0	2	3
AC	1	4	0
BB	2	4	6
CC	3	6	2

$m' = 7$

$\#keys = 4$

$7 \geq 4$ ok



mod 4

$y = [0\ 0\ 2\ 2\ 1\ 0\ 1]$

$0\ 1\ 2\ 3\ 4\ 5\ 6$

Given the sequence a,b,c,d,e,f,g,h,i,l simulate:

- The sampling algorithm for m=2 which knows the sequence length n=10,

assuming probabilities for the parameter p = [0.5, 0.5, 0.5, 1, 1, 0.1, 0.5, 1, 0.1, 1]

- The sampling algorithm for m=2 which does not know the sequence length,

assuming values for the parameter h = [1, 3, 4, 2, 1, 5, 4, 6]

1)

	1	2	3	4	5	6	7	8	9	10
S =	a	b	c	d	e	f	g	h	i	l
p =	0.5	0,5	0,5	1	1	0,1	0,5	1	0,1	1

checked if $P \leq \frac{m-s}{n-j+1}$

$s=0 \; j=1 \Rightarrow \frac{2-0}{10-1+1} = \frac{1}{5}$ ✗

$s=0 \; j=2 \quad \frac{2-0}{10-2+1} = \frac{2}{9}$ ✗

$s=0 \; j=3 \quad \frac{2-0}{10-3+1} = \frac{2}{8}$ ✗

$s=0 \; j=4 \quad \frac{2-0}{10-4+1} = \frac{2}{7}$ ✗

$s=0 \; j=5 \quad \frac{2-0}{10-5+1} = \frac{2}{6} = \frac{1}{3}$

$s=0 \; j=6 \quad \frac{2-0}{10-6+1} = \frac{2}{5}$ ✓

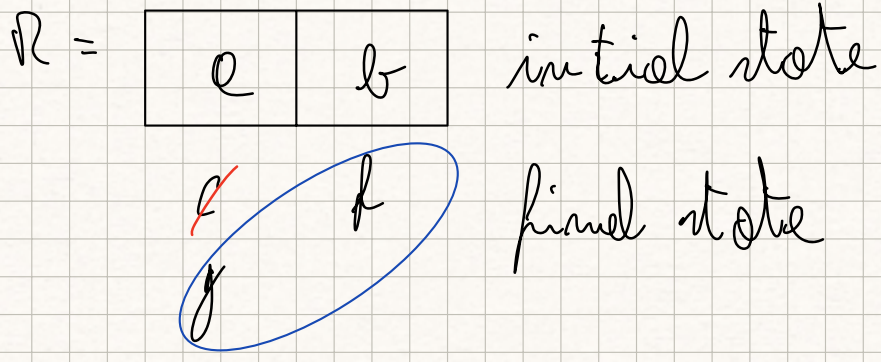
$s=1 \; j=7 \quad \frac{2-1}{10-7+1} = \frac{1}{4}$ ✗

$s=1 \; j=8 \quad \frac{2-1}{10-8+1} = \frac{1}{3}$ ✗

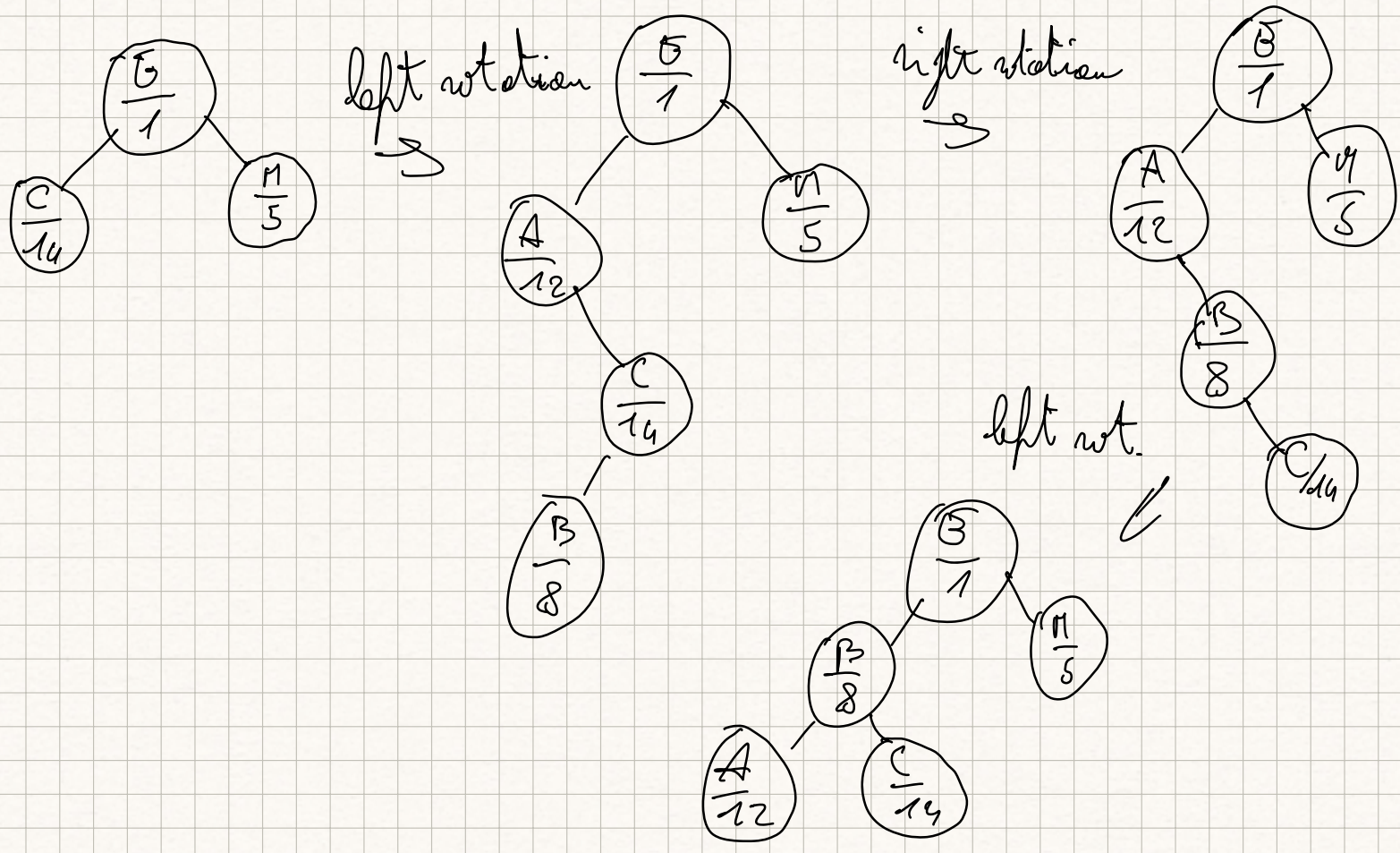
$s=1 \; j=9 \quad \frac{2-1}{10-9+1} = \frac{1}{2}$ ✓

Minimum length

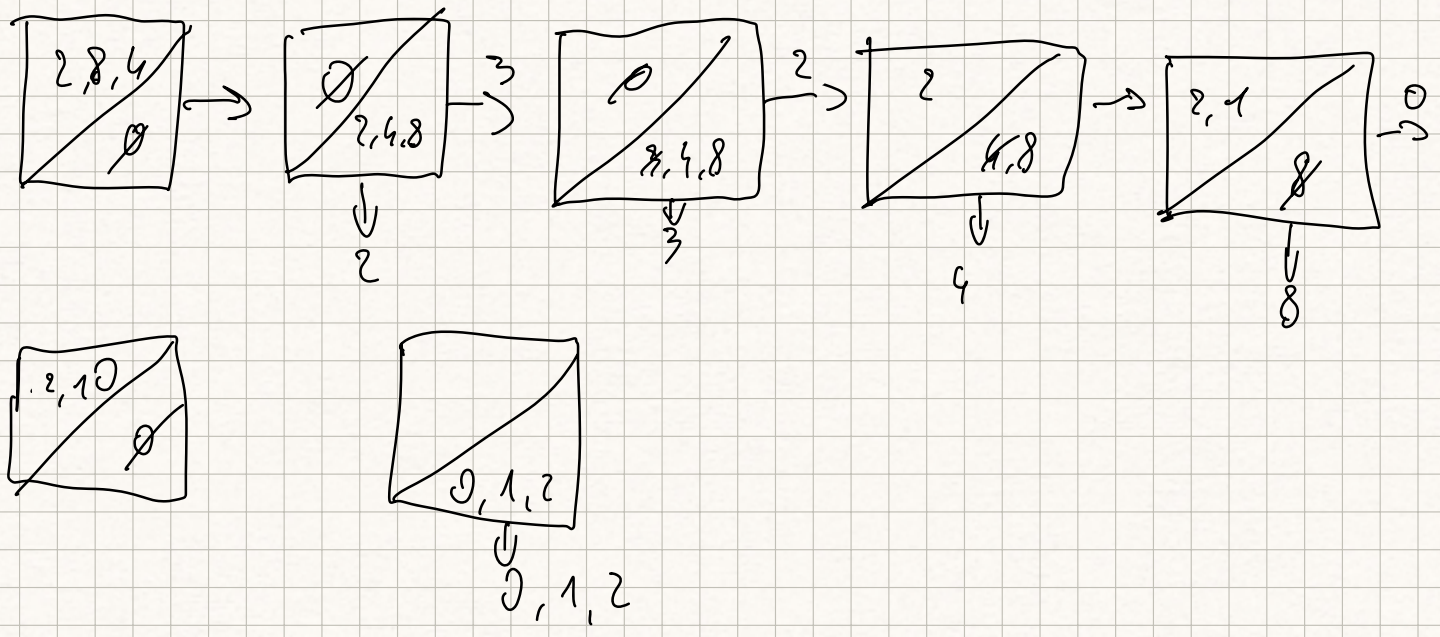
	1	2	3	4	5	6	7	8	9	10
S =	a	b	c	d	e	f	g	h	i	l
h =	1	3	4	2	1	5	4	6		



Build a Treap by inserting the following sequence of pairs <key,priority> and assuming that the MIN priority is in the root (the order among the keys is the alphabetic one):
<E,1><C,14><M,5><A,12><B,8>

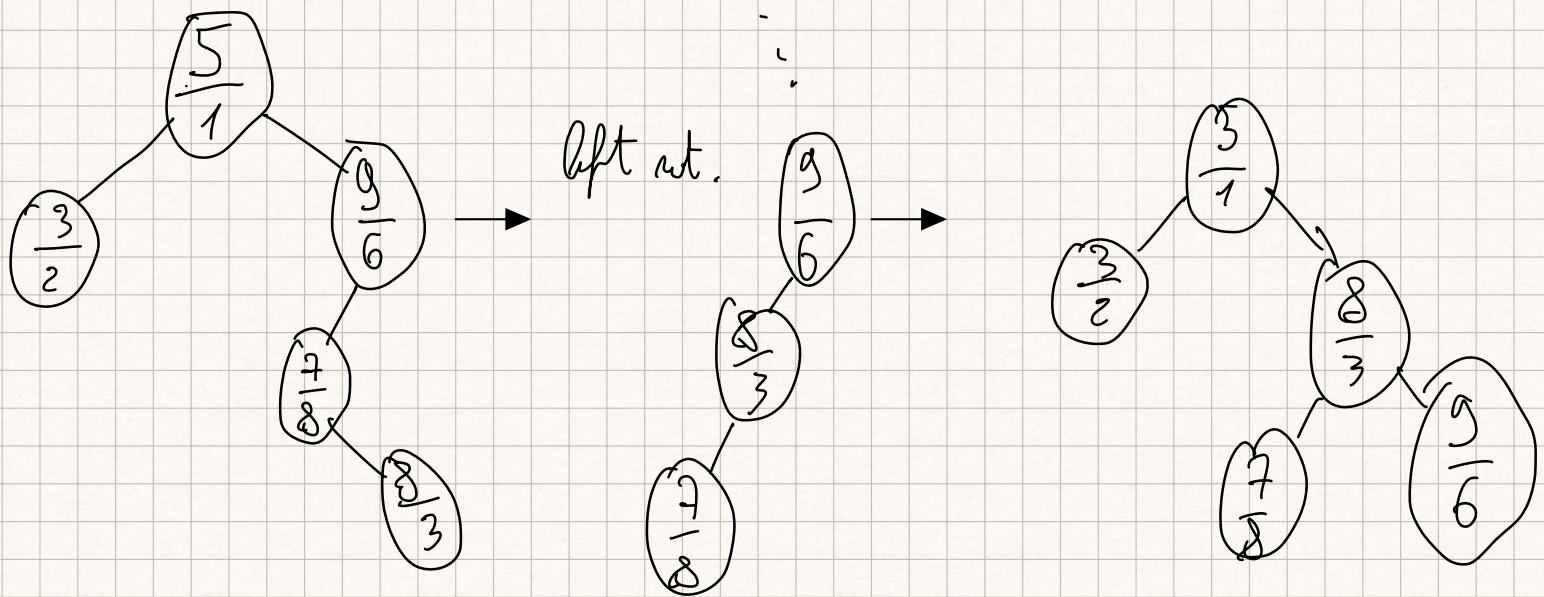


Simulate the algorithm Snow Plow on the sequence: (2, 8, 4, 3, 2, 1, 0) by assuming a memory size M=3.



You are given a set of pairs (key, priority) to insert in a Treap according to the following order: (5,1), (9,6), (3,2), (7,8).

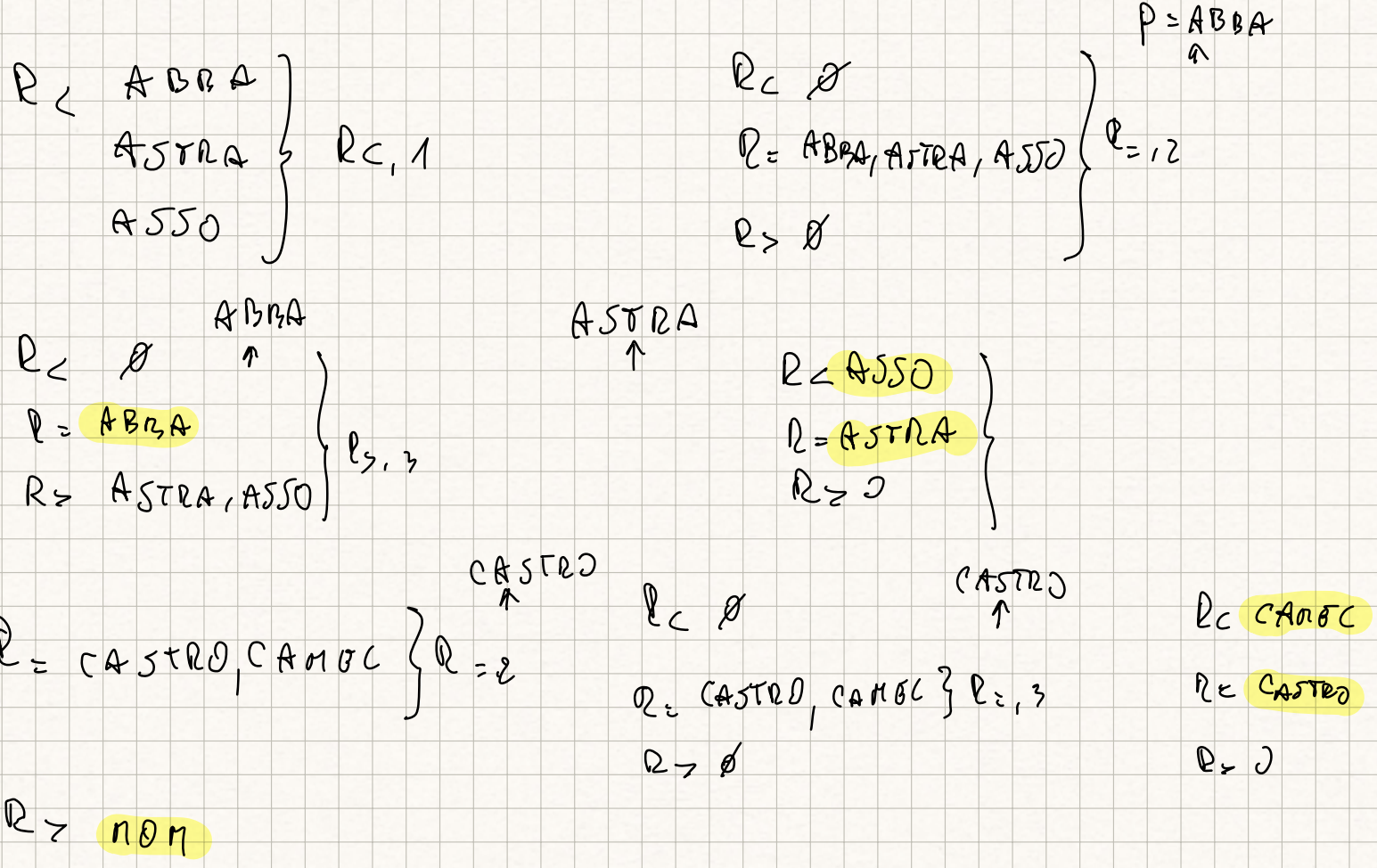
- Show the Treap resulting from the insertion of every key above
- Insert then the pair (8,3)



Sort the following strings via Multi-key Quicksort

S = { castro, abba, mom, camel, astra, asso } by using as pivot always the first string of the set to sort.

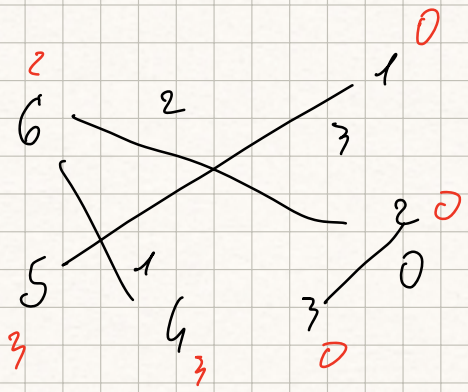
pivot = castro i = 1



Given the following set of strings, each consisting of two digits:
S={11,22,33,44}
Compute the Minimal Ordered Perfect Hash for S by assuming the following two hash functions: h1(xy) = x+y mod 7 and h2(xy) = x+2*y mod 7
in which x (resp., y) is the first (resp., second) digit of a string of S. As an example, if the string is 11, then x=1 and y=1.

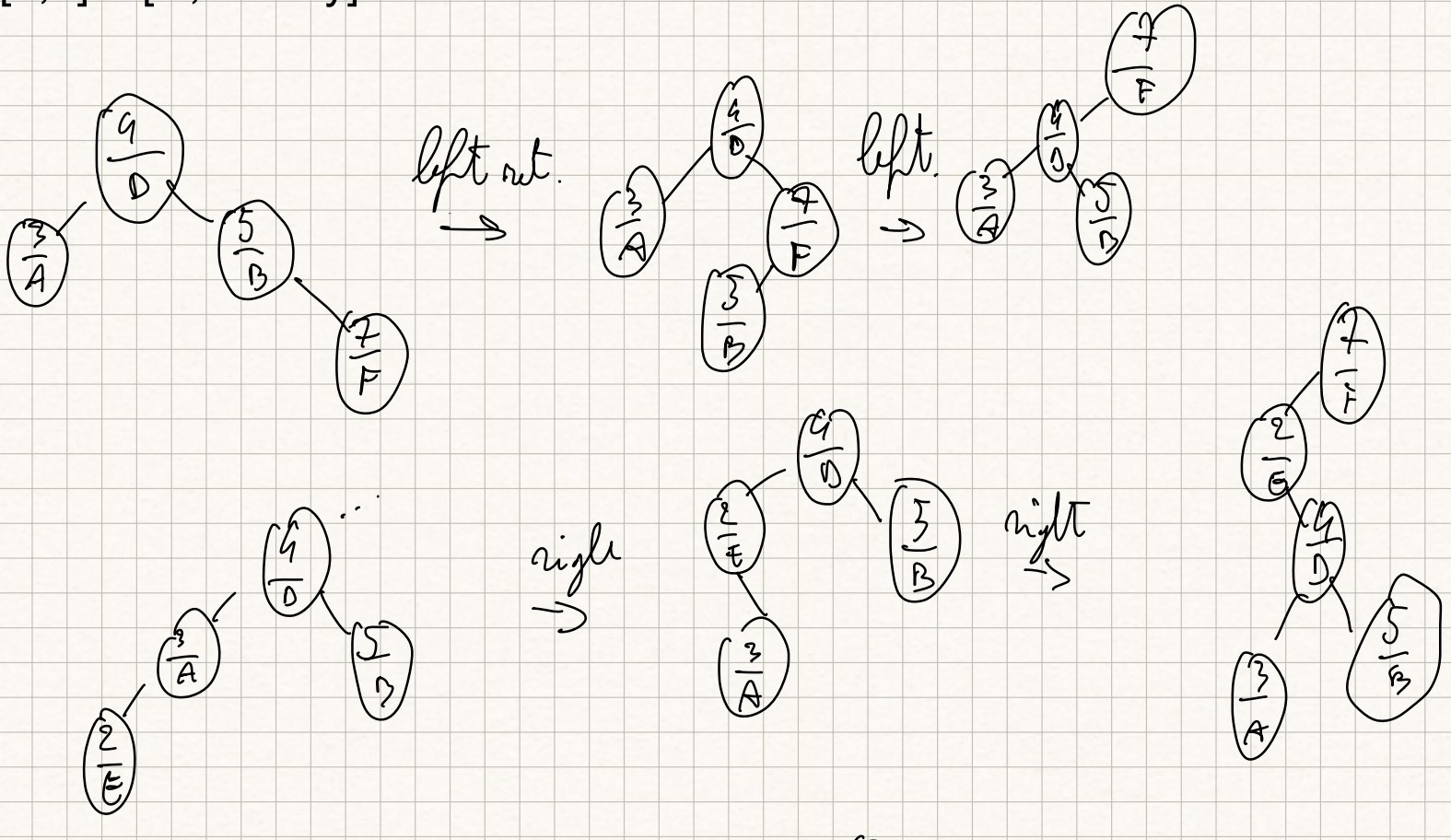
	h	h1	h2
11	0	2	3
22	1	4	6
33	2	6	2
44	3	1	5

$m' = 7 > m = 4$ ok

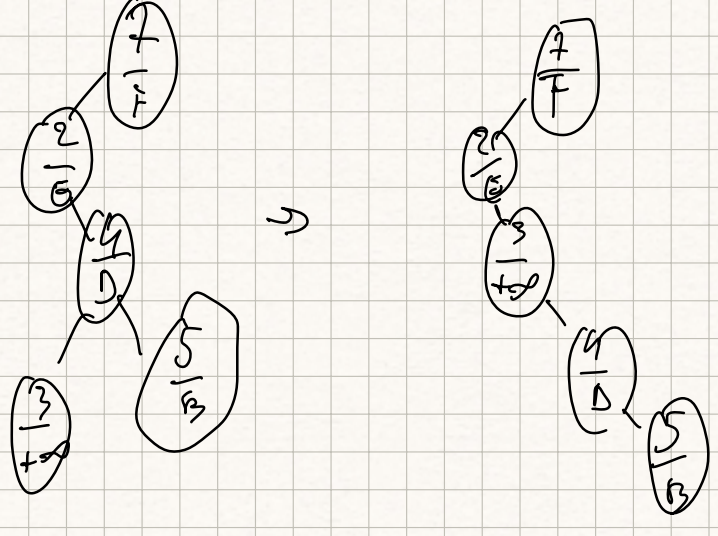


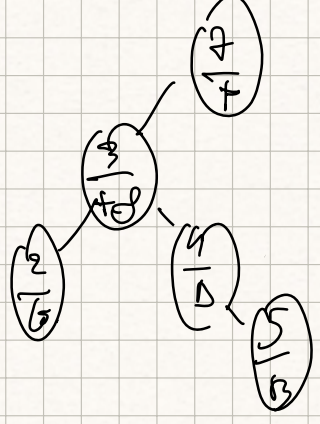
Given the set S of pairs {<4,D>, <5, B>, <3,A>, <7,F>, <2,E>}, where the first component is the key and the second component is the priority:

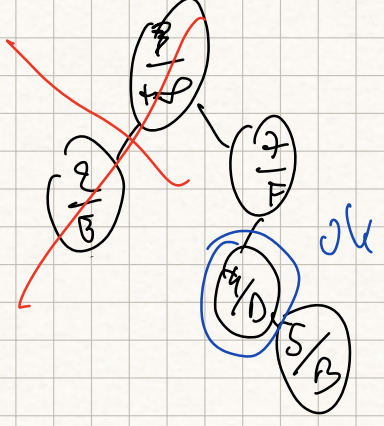
- Build a TREAP data structure by inserting the pairs in that order (you can assume that it is a MAX heap and letters on the Y-axes are sorted from bottom to top as A, B, C, ...).
- Show and comment how to solve the three-sided range query [3,6] x [C, + infty]



2) · Split (3)
· drop left subtree (item < 3)

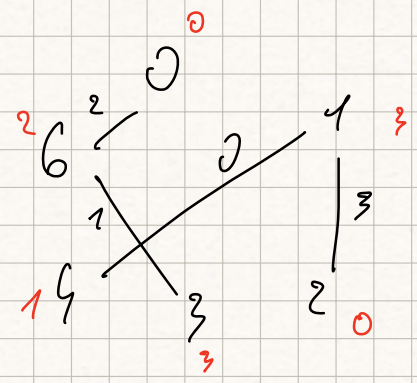






Assume you are given a set of 4 strings {aa, ac, bc, cc} and you wish to construct a minimal ordered perfect hash function (MOPHF), where the order is the alphabetic one. Assume that rank(x) =2; 3; 4 for the characters x = a; b; c; respectively. Given a string xy of two characters, we let the two random functions required by the design of MOPHF be h1(xy) = (3 * rank(x) + rank(y)) mod 7 and h2(xy) = (rank(x) + rank(y)) mod 7.

	h	h1	h2
AA	0	1	4
AC	1	3	6
BC	2	6	0
CC	3	2	1



Given the sequence of 6 items S = (a, b, c, d, e, f) use the random sampling algorithm with known sequence length n=6, to compute the m=2 extracted items given the random values p = (1/2, 1/10, 3/4, 3/4, 0, 1)

j =	1	2	3	4	5	6
S =	a	b	c	d	e	f
p =	1/2	1/10	3/4	3/4	0	1

checked if: $p \leq \frac{m-s}{n-j+1}$

$S=0\ j=1 \quad \frac{2-0}{6-1+1} = \frac{1}{3}$ X

$S=0\ j=2 \quad \frac{2-0}{6-2+1} = \frac{2}{5} > 1/1$ OK

$S=1\ j=3 \quad \frac{2-1}{6-3+1} = \frac{1}{4}$ X

$S=1\ j=4 \quad \frac{2-1}{6-4+1} = \frac{1}{3}$ X

$S=1\ j=5 \quad \frac{2\cdot 1}{6-5+1} = \frac{1}{2} > 0$ ✓

$S=1\ j=6 \quad \frac{1}{1} = 1$ OK

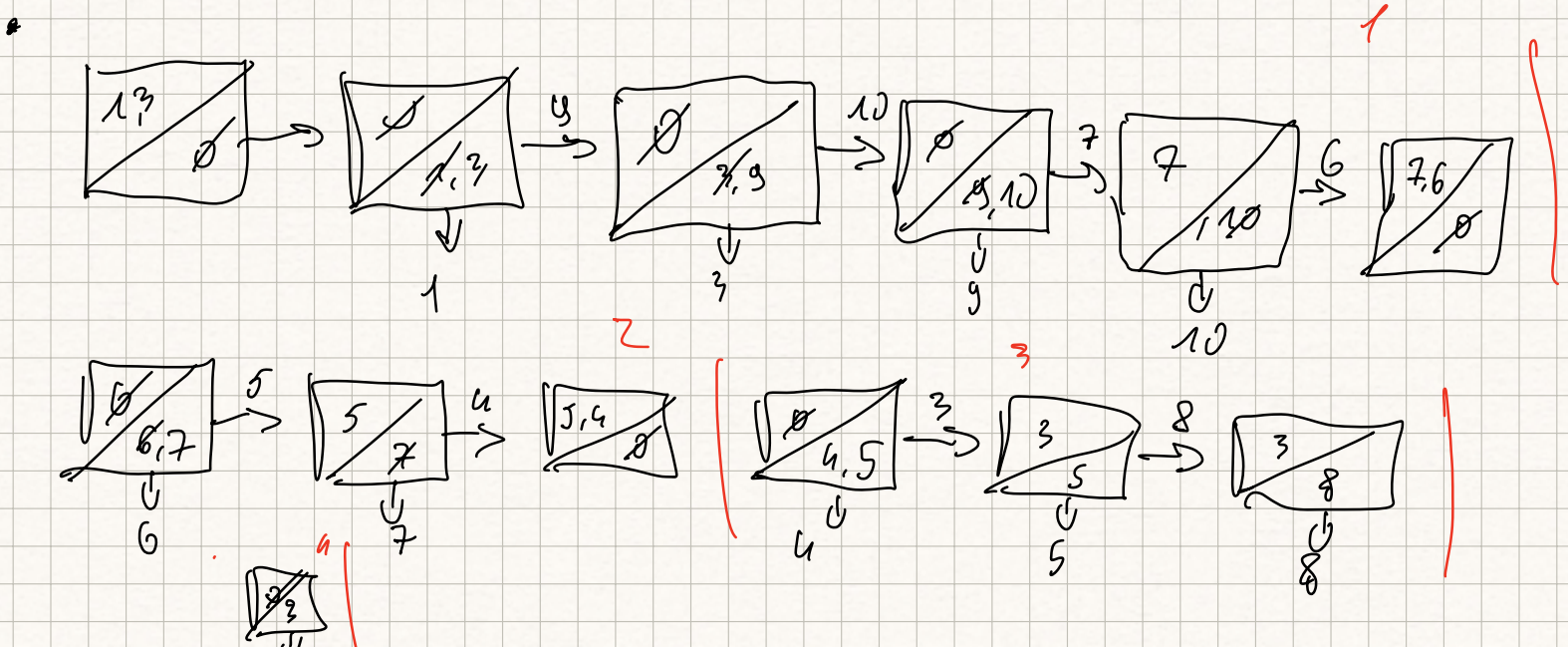
Consider the Snow Plow technique with memory M=2.

• Simulate Snow Plow over the sequence S = (1, 3, 9, 10, 7, 6, 5, 4, 3, 8).

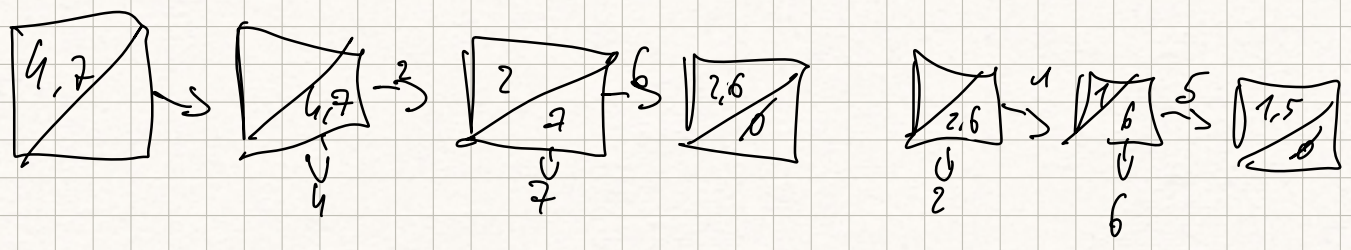
• Provide an example of sequence of length 10 that generates exactly 5 runs,
when the memory has size M=2.

• Show the average length of the run produced by Snow Plow, if we assume
that the probability that an item goes to the Heap is 3⁄4 rather than 1⁄2.

M = 2



- $S = (4, 7, 2, 6, 1, 5, 3, 11, 9, 20)$



- $T/4 = M$ $\quad t = 4M$ $\quad T$ #item read in a phase

P(item goes to memory) $= 1/4$

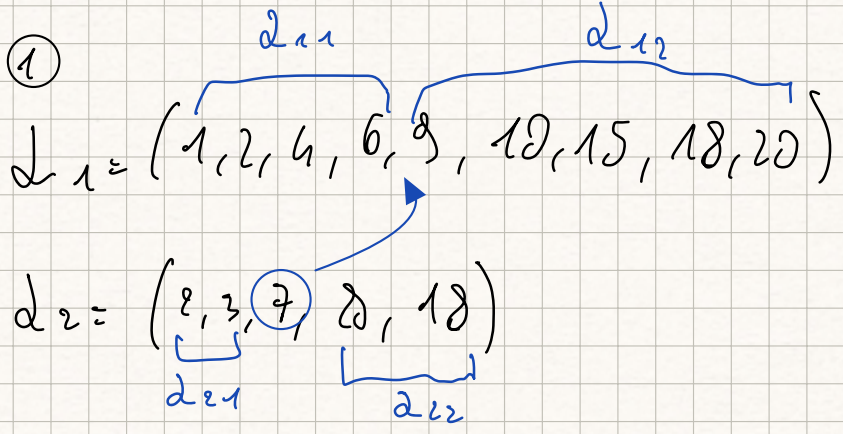
$L_1 = (1, 2, 4, 6, 9, 10, 15, 18, 20)$

$L_2 = (3, 3, 7, 8, 18)$

① Compute $L_1 \cap L_2$ by mutual partitioning

② " " by 2-level storage approach $B = 3$

①



$L_{11} \cap L_{21}$ $\quad$ 2 $\quad$ 4 $\quad$ 6 $\quad$ print 2

2 3

$d_{12} \cap d_{22}$ 9 10 15 18 20

8 18

⇒ print 18 in the next step

②

	B_1	B_2	B_3
d_1 =	1 2 4	6 9 10	15 18 20
d_1' =	1 6 15		
d_2 =	2 3 7	8 18	

d_2' = 2 8 not used here

$\{1,2,4\} \cap \{2,3\} = 2$

$\{6,9,10\} \cap \{7,8\} = \emptyset$

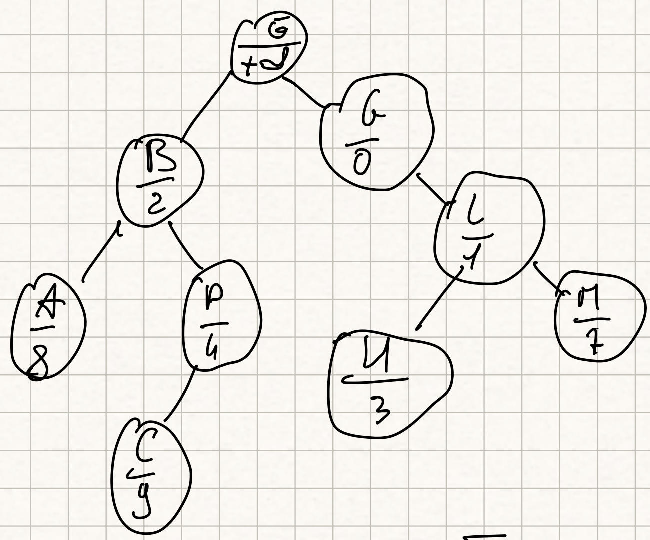
$\{15,18,20\} \cap \{18\} = 18$

Merge d_1' and d_2'. In this case we do not skip any block

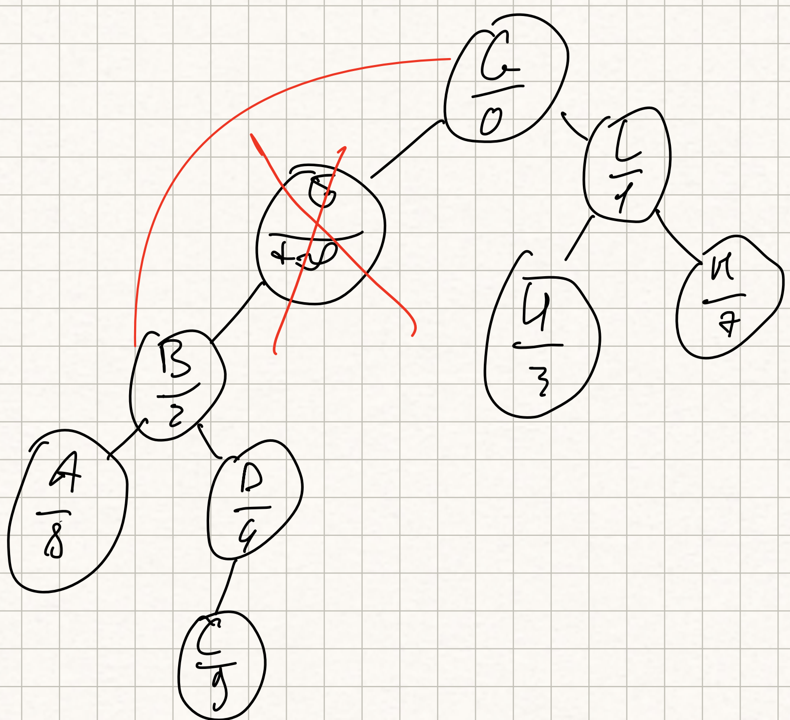
$T_1 = \langle A, 8 \rangle, \langle B, 2 \rangle, \langle C, 9 \rangle \ \langle D, 4 \rangle$

$T_2 = \langle H, 3 \rangle, \langle M, 7 \rangle \ \langle G, 0 \rangle, \langle L, 1 \rangle$

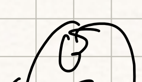
} merge

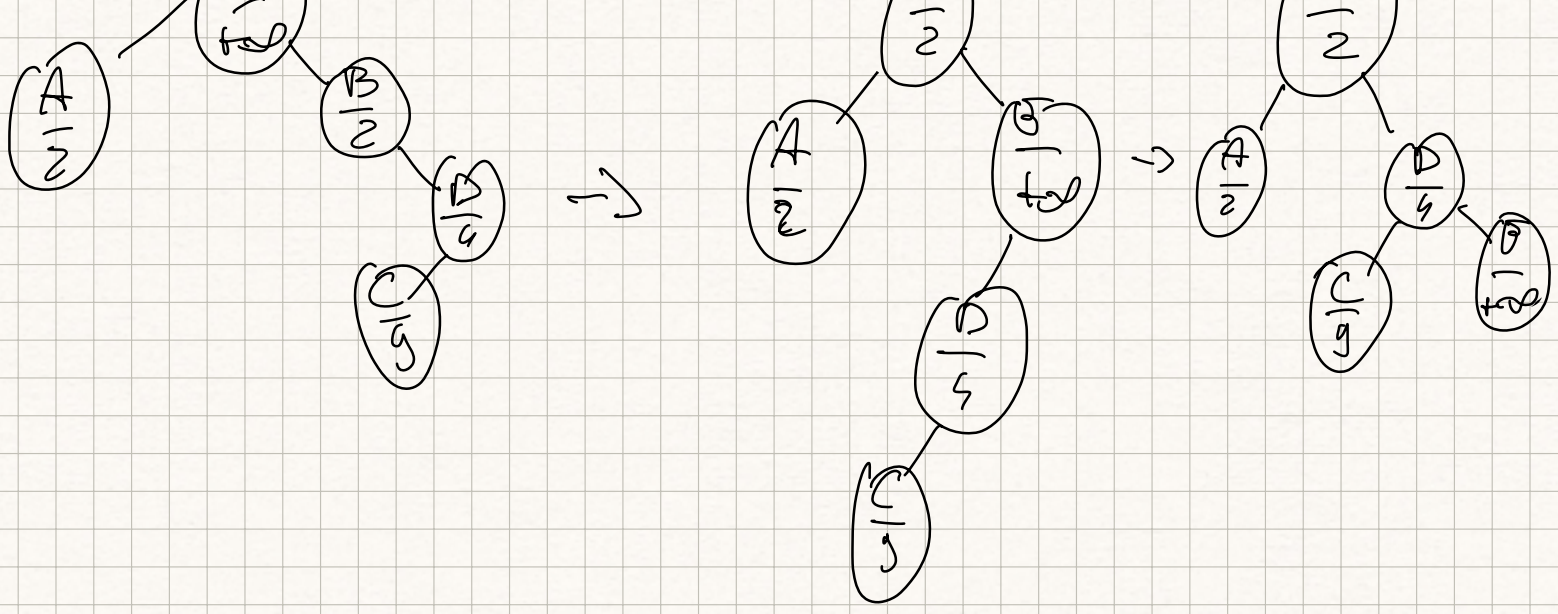


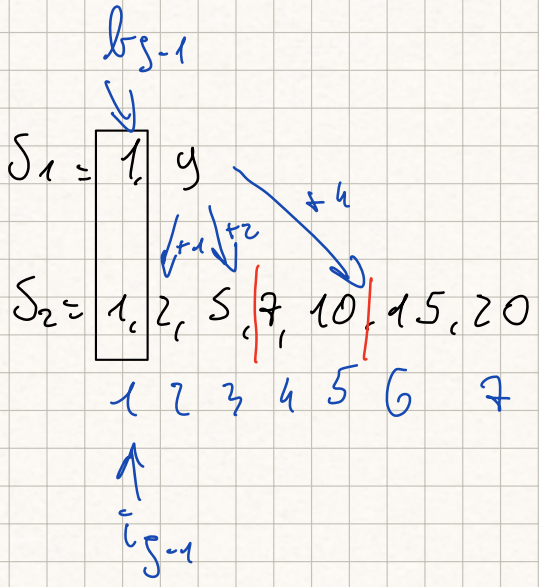
T_1 T_2



if I would continue







Exponential search

Mutual partitioning

$S_1 = (1), 9$

$S_2 = (1), 2, 5, 7, 10, 15, 20$

9

$2, 5, 7, 10, 15, 20$

$k=3$, $a=2$, $n=15$

sampled items $= k(a+1)-1 = 8$

$S = 3, 1, 10, 9, 7, 6, 2, 50, 40, 14, 39, 21, 20, 5, 4$

$A = 1, 9, 6, 2, 14, 39, 21, 4$

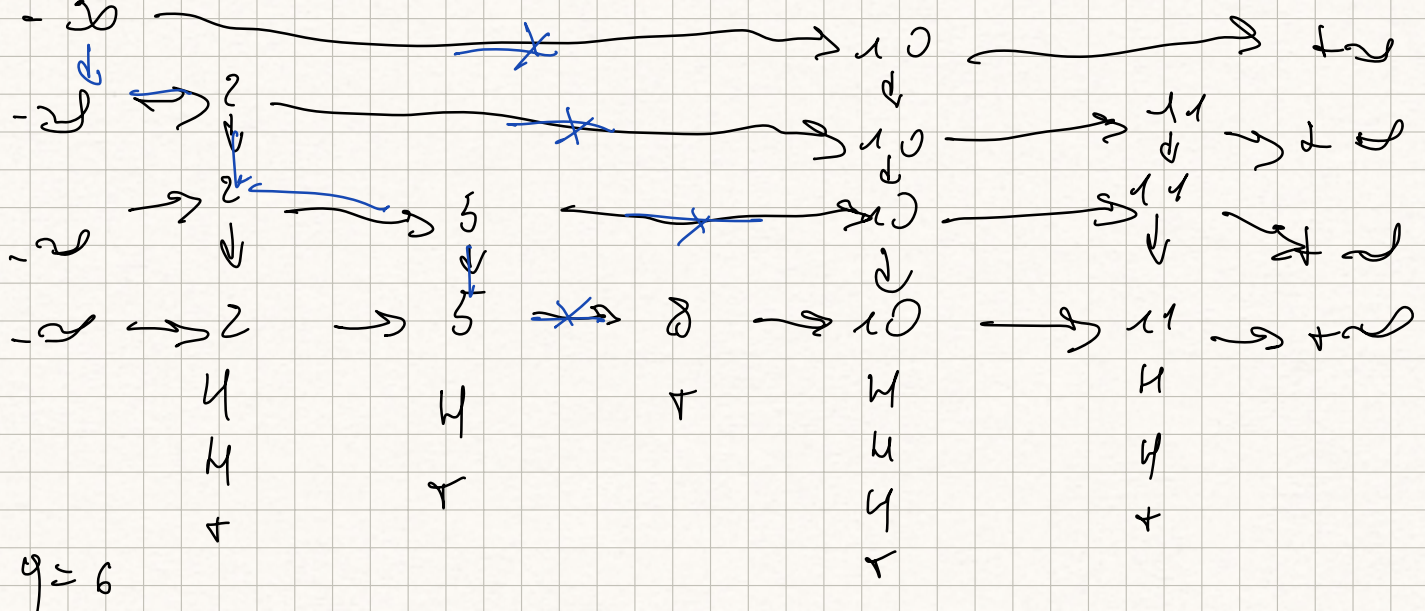
$= \underbrace{1, 2, 4}_{S_1}\ \underbrace{6, 9, 11}_{S_2}\ 21, 39$

$(-\infty, 4]\ (4, 11]\ (11, +\infty)$

$3, 1, 2, 4 \quad 10, 9, 7 \quad 50, 40, 39, 21, 20$

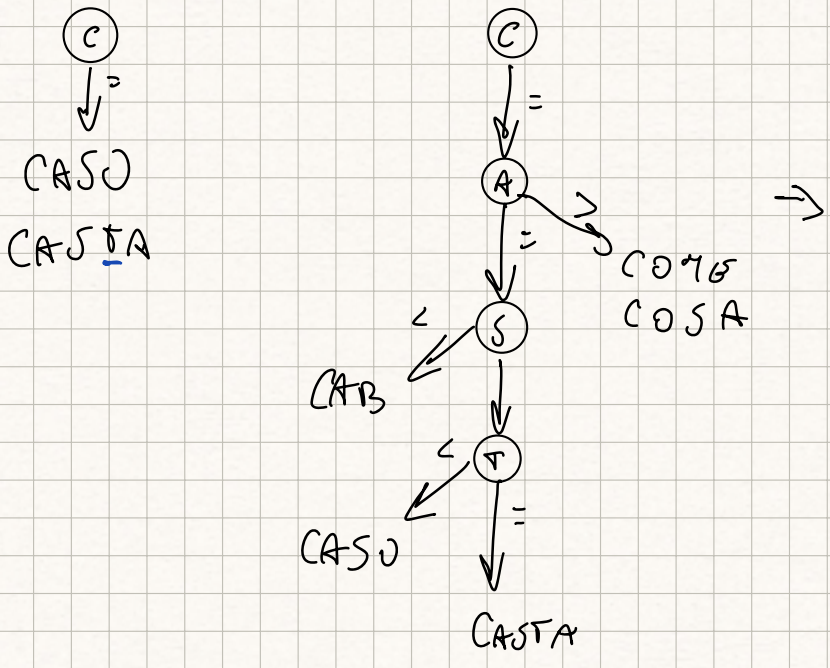
$6, 11, 5$

th. satisfied if bucket size is $\leq \frac{4n}{k} = 20$

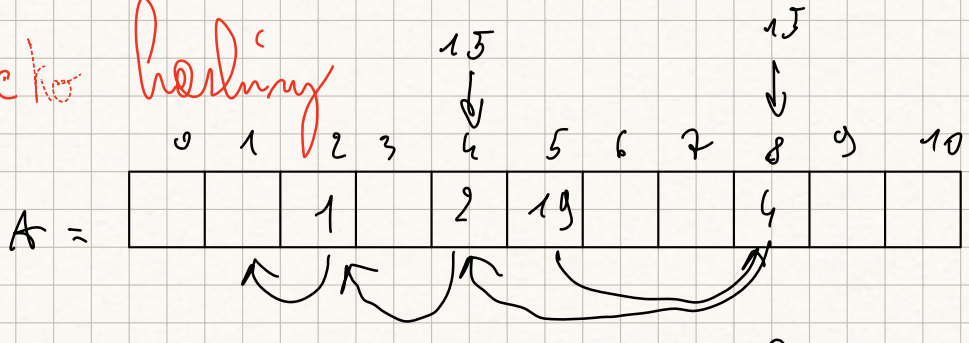


$q = 6$

$S = \{$ CASO, CASTA, CAB, COMO, COSA $\}$ $O(p + \log n)$ Time



Cuckoo hashing

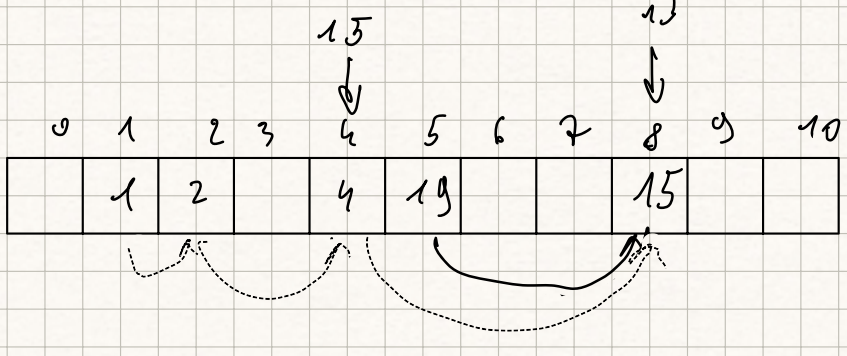


$m = 11$

$S = \{ 1, 4, 2, 19, 15$

$h_1(x) = 2x \bmod 11$

$h_2(x) = x \bmod 11$



Given the two sored lists L1 = (1, 2, 4, 6, 9, 10, 15, 18, 20) and L2 = (2, 3, 7, 8, 18) compute their intersection using the

1)- Mutual partitioning strategy

2)- Two-level storage approach, with block size b=3 for the list L1

1)

L_{11} L_{12}

$L_1 = 1, 2, 4, 6, 9, 10, 15, 18, 20$

$L_2 = 2, 3, 7, 8, 18$

L_{21} L_{22}

$L_{11} \; L_{21} = \{1, 2, 4, 6\}$ (L_{11}')

$\{2, 3\}$ (L_{21}')

$L_{12} \; L_{22} = \{9, 10, 15, 18, 20\}$ (L_{12}')

$\{8, 18\}$ (L_{22}')

$L_{11}' \; L_{21}' = \{4, 6\}$

$\{3\}$

$L_{12}' \; L_{22}' = \{9, 10, 15, 18, 20\}$

$\{18\}$

$L_1 = 1, 2, 4, 6, 9, 10, 15, 18, 20$

$L_2 = 2, 3, 7, 8, 18$

B_1 B_2 B_3

$L_1' = 1, 6, 15$

Merge L_2 and L_1'

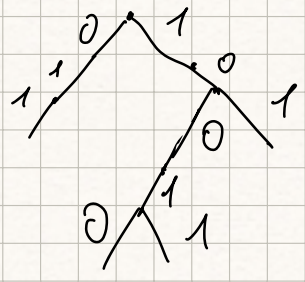
$B_1 \cap L_{11} = 2$

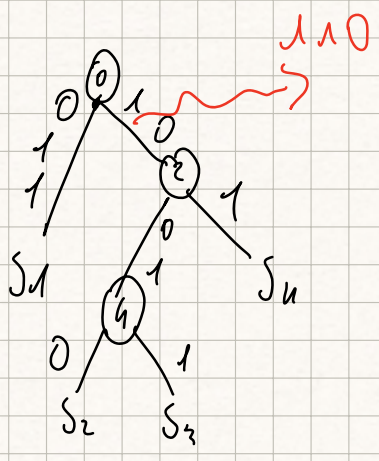
$B_2 \cap L_{12} = \emptyset$

$B_3 \cap L_{12} = 18$

Given the binary strings S={011, 10010, 10011, 101}.

- Build the Patricia Trie for S
- Show how to search for the lexicographic position of the string P=110 among the strings of S.





Given the sequence of 6 items S = (a, b, c, d, e, f) use the random sampling algorithm with known sequence length n=6, to compute the m=2 extracted items given the random values p = (1/2, 1/10, 3/4, 3/4, 0, 1)

S = a b c d e f

1 2 3 4 5 6

1/2 1/10 3/4 3/4 0 1

m = 2

$P < \frac{m-s}{n-s+1}$

$s=0\ s=1 \quad 1/2 < \frac{2-0}{6-1+1} = \frac{1}{3}$ ✗

$s=0\ s=2 \quad 1/10 < \frac{2-0}{6-2+1} = \frac{2}{5}$ ✓

$s=1\ s=3 \quad 3/4 < \frac{2-1}{6-3+1} = \frac{1}{4}$ ✗

$s=1\ s=4 \quad 3/4 < \frac{2-1}{6-4+1} = \frac{1}{3}$ ✗

$s=1\ s=5 \quad 0 < \frac{2-1}{6-5+1} = \frac{1}{2}$ ✓

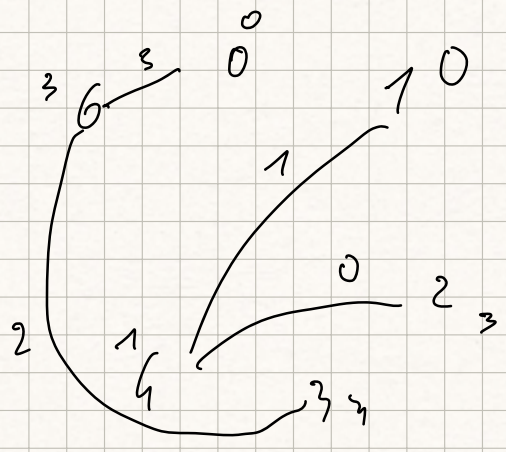
	h	h_1	h_2
AA	0	4	2
BB	1	1	4
BD	2	3	6
CD	3	6	0

$h_1 = 3 \cdot rank(x) + rank(y) \bmod 7$ ✓

$h_2 = rank(x) + rank(y) \bmod 7$ ✓

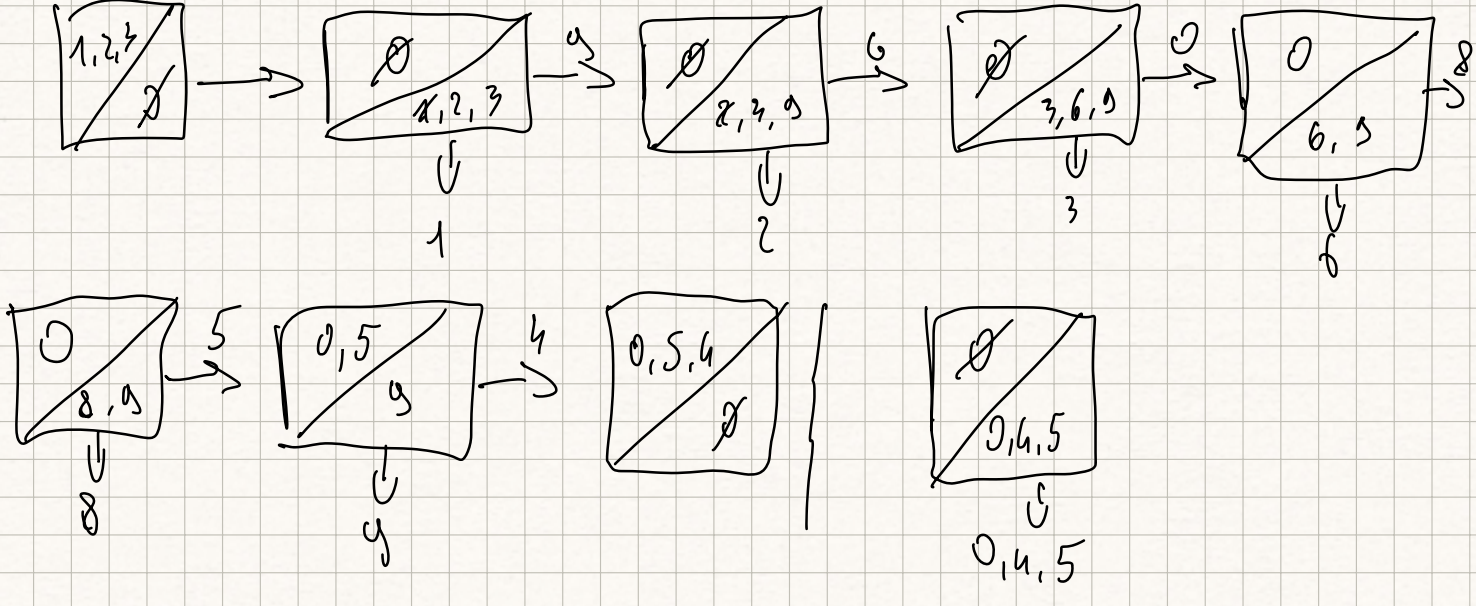
$rank(1,2,3,4)$

$h = g(h_1(1)) + g(h_2(1)) \bmod 4$

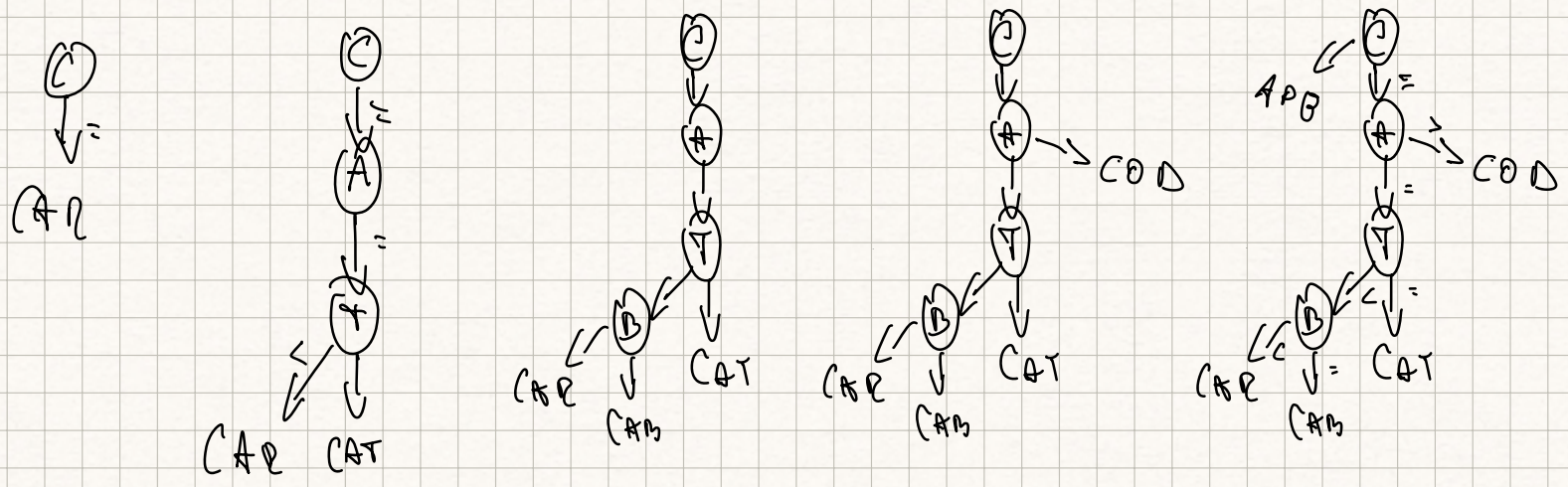


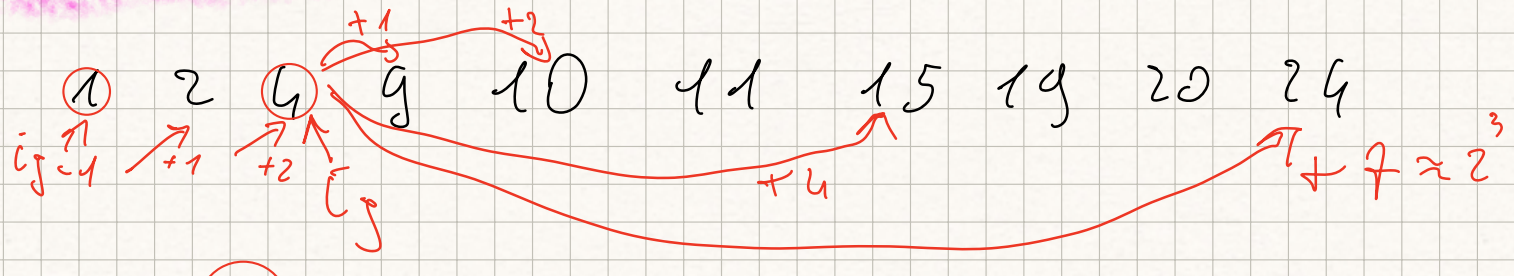
$g = 00\ 3\ 3\ 1\ 0\ 3$

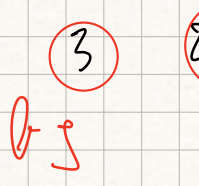
Snow Plow : 1, 2, 3, 9, 6, 0, 8, 5, 4 $B = 3$

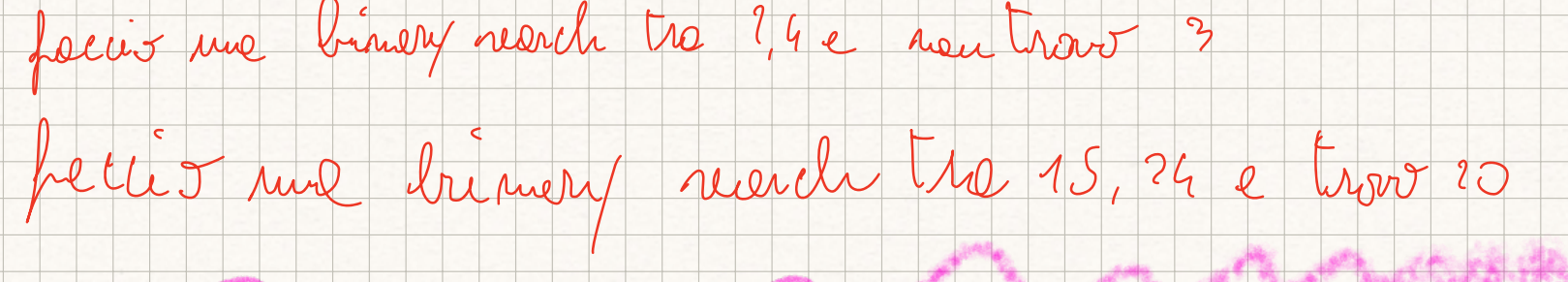


Build a Ternary Search Tree by inserting the following strings according to the given sequence (CAR, CAT, CAB, COD, APE)









facciamo una binary search tra 1, 4 e non trovo 3

facciamo una binary search tra 15, 24 e trovo 20

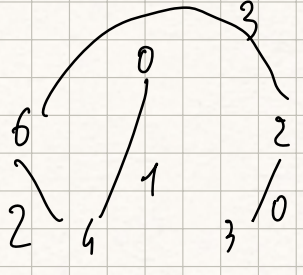
	h	h_1	h_2
AA	0	2	3
AC	1	4	0
BB	2	4	6
CC	3	6	2

$h_1 = x + y \bmod 7$

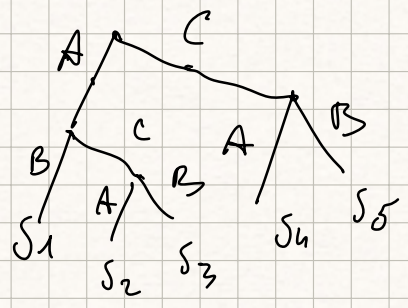
$h_2 = x + 2y \bmod 7$

$h = g(h_1) + g(h_2) \bmod 4$

	0	1	2	3	4	5	6
g =	0	0	2	2	1	0	1



AB, ACA, ACB, CA, CB



A = 3 5 6, 8 10 13, 14 16 17, B = 3

B = 5 7 8 14 15

A1 A2 A3

merge 3 5 7 8 8 14 14 15

A' = 3 8 14

Q5: CASTRO, ABBA, NOM, CAMIL, ASTRA, ASSO

R, ABBA, ASTRA, ASSO → i=1 R: ABBA, ASTRA, ASSO → i=2 P=ABBA

R = ABBA

R = ASTRA, ASSO → i=2 P=ASTRA

R = ASTRA, ASSO → i=3 R = ASTRA, ASSO

R, ASSO

R = CASTRO, CAMIL → i=2 P=CASTRO

R = CASTRO, CAMIL → i=3

R, CAMIL

R = CASTRO

R = NOM

$S = 1, 6, 15, 18, 21, 24, 30$

$\downarrow$

$S = 1, 5, 9, 3, 3, 3, 6$ gap encoding

γ - code

δ - code

Rice - code $k = 3$

$(1,3)$ - dense code

$\gamma\text{-code}(S') = 1\ 00101\ 0001001\ 011\ 011\ 011\ 00110$

$\delta\text{-code} = \gamma\,\text{code}(L)\ \text{bin}(x) = 11\ 011\,101\ 001\,0001001\ 0101\ 0101\ 0101\ 01110$

Rice - code:

$$q = \left\lfloor \frac{x-1}{2^k} \right\rfloor$$

$$r = x - 1 - 2^k \cdot q$$

	Unary	Binary
$R\text{-}c(5) \Rightarrow$	$q = 0$ (1)	$r = 4$ (100)
$R\text{-}c(9) \Rightarrow$	$q = 1$ (01)	$r = 0$ (000)
$R\text{-}c(3) \Rightarrow$	$q = 0$ (1)	$r = 2$ (010)
$R\text{-}c(6) \Rightarrow$	$q = 0$ (1)	$r = 5$ (101)

Dense - code:

	codify	
1	00	5
2	0100	c 5
3	1000	c 5
4	1100	c 5
5	01 01 00	c c 5
6	01 10 00	$3 \times 3 \times 1 = 9$
7	01 11 00	
8	10 01 00	
9	10 10 00	

$1 + 3 = 4 = 2^2$ ← b

00 5

01

10 c

11

c c c 5 = 27 > 25

$3 \times 3 \times 3 \times 1$

Elias - Fano:

$H = 0110\,101$

$L = \overbrace{10}^{0}\,\overbrace{10}^{1}\,\overbrace{100}^{2}\,\overbrace{0}^{3}\,\overbrace{110}^{4}\,\overbrace{110}^{5}\,\overbrace{0}^{6}\overbrace{0}^{7}$

$n = 7 = \#1$ in H

$l = u = \frac{|L|}{n}$

$h = 3$ $\#0$ in H

$b = 4$

	H	L
0	0 0 0	0
3	0 0 1	1
5	0 1 0	1
8	1 0 0	0
9	1 0 0	1
10	1 0 1	0
11	1 0 1	1

1, 3, 4, 5, 9, 16, 23, 27, 28, 31, 40

	H				L	
1	0	0	0	0	0	1
3	0	0	0	0	1	1
4	0	0	0	1	0	0
5	0	0	0	1	0	1
9	0	0	1	0	0	1
16	0	1	0	0	0	0
23	0	1	0	1	1	1
27	0	1	1	0	1	1
28	0	1	1	1	0	0
31	0	1	1	1	1	1
40	1	0	1	0	0	0

$n = 11$

$u = 41$

$l = \lceil \log_2 \frac{u}{n} \rceil = 2$

$b = \lceil \log_2 u \rceil = 6$

$h = 6 - 2 = 4$

$C = 01110001010100111100110\,0$

$H = 110110100101010101101100010000000$

$Access(4) = 0001|01$ $\quad Select_1(4) = 5 - 4 = 1$

$GEQ(8) = ?$

$GEQ(22) = ?$

$Access(7) = 0101|11$ $\quad Select_1(7) = 12 - 7 = 5$

$BBQ_2(8) = 0010100$ $\quad$ $Select_0(2) + 1 = 6 + 1 = 7$ $\quad$ select the right bracket

$7 - 2 = 5$

$Access(5) = 9$

$BBQ_2(32) = \underbrace{1000100}_{8}$ $\quad$ $p = select_0(8) + 1 = 18 + 1 = 19$

$p - 8 = 11$

$Access(11) = 40$

Interpolative code

$S = 11, 14, 16, 18, 20, 21, 22$

$l = 1$

$r = 7$

$low = 11$

$high = 22$

$m = \lceil \frac{l + r}{2} \rceil = 4$ $\quad$ $S_m = 18$

$c = low + m - l = 11 + 4 - 1 = 14$

$b = high + m - r = 22 + 4 - 7 = 19$

encode $S_m - c = 5$ in $\lceil \log_2 b - c + 1 \rceil = 3$ $\quad (101)_2$

---

right

$l = 5, r = 7, low = 20, hi = 22 \Rightarrow m = 6, S_m = 21$

$c = 20 + 6 - 5 = 21$

$b = 22 + 6 - 7 = 21$

} encode 0 in 0 bits

$\lceil \log_2 1 \rceil$

left

$l = 1, 2 = 3,$ low $= 11,$ high $= 18$

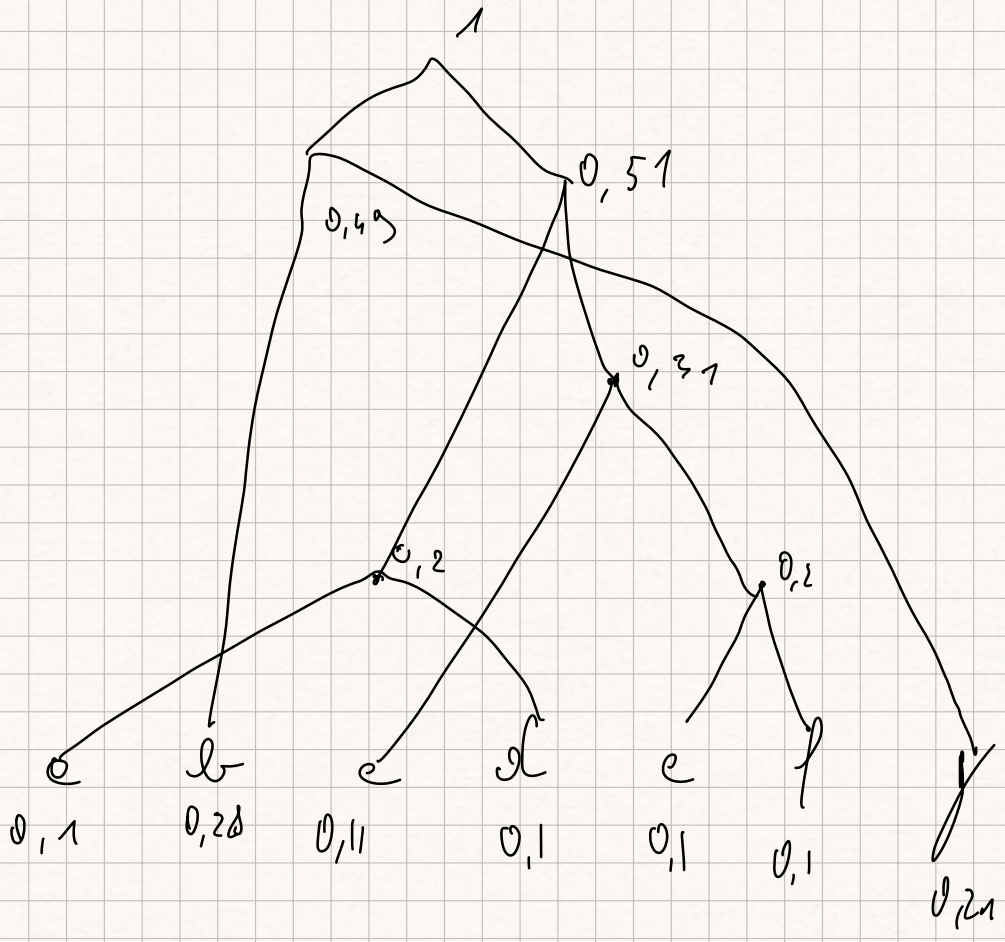
$m = 2 \quad S_m = 14$

$a = 11 + 2 - 1 = 12$

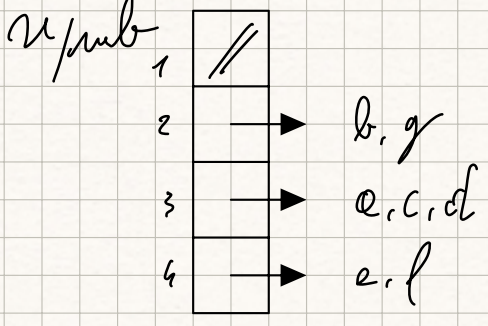
$b = 18 + 2 - 3 = 17$

encode 2 in $\lceil \log 17 - 12 + 1 \rceil = 3$ $(010)_2$

- Huffman tree
- Canonical Huffman tree
- Decompress 11001 using Canonical H.T.



num [1,4] = [0, 2, 3, 2]
1 2 3 4



FC [1,4] = [2, 2, 1, 0]

FC [4] = 0

FC [3] = $\frac{FC[4] + num[4]}{2} = 1$

FC [2] = $\frac{FC[3] + num[3]}{2} = 2$

FC [1] = $\frac{FC[2] + num[2]}{2} = 2$

Decompressor():

```
v = next_bit();
l = 1;
while (v < FC[l])
    v = 2·v + next_bit();
    l++
return symb[l, v - FC[l]]
```

1 1 0 0 1

v = 1 l = 1 v = 1 < FC [1] = 2

v = 3 l = 2 v = 3 > FC [2] = 2

symb [2, 1] = g

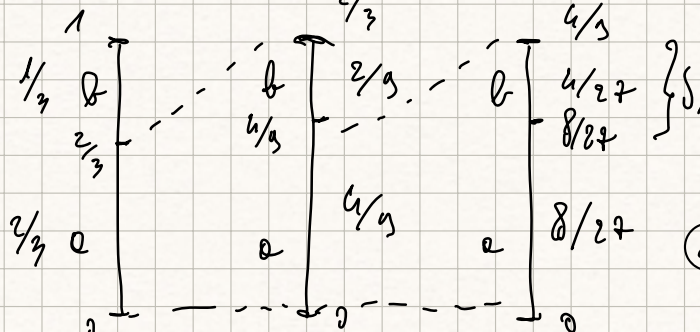
v = 0 l = 1 v = 0 < FC [1] = 2

v = 0 l = 2 v = 0 < FC [2] = 2

v = 2·0 + 1 = 1 l = 3 v = 1 < FC [3] = 1

symb [3, 1-1] = a

T = a a b $P(a) = \frac{2}{3}$ $P(b) = \frac{1}{3}$



$S_n \left[\frac{8}{27}, \frac{4}{9}\right] \rightarrow \frac{8}{27} + \frac{1}{2} \cdot \frac{4}{27} = \frac{10}{27}$ ①

② $d = \left\lceil \log_2 \frac{2}{S_n} \right\rceil = \left\lceil \log_2 \frac{2}{4/27} \right\rceil = \left\lceil \log_2 \frac{27}{2} \right\rceil = 4$

$S_0 = 1$ $S_1 = 2/3$ $S_2 = 4/9$

$l_0 = 0$ $l_1 = 0$ $l_2 = 0$

③ Converter $\left(\frac{10}{27}, 4\right) = 0.104$

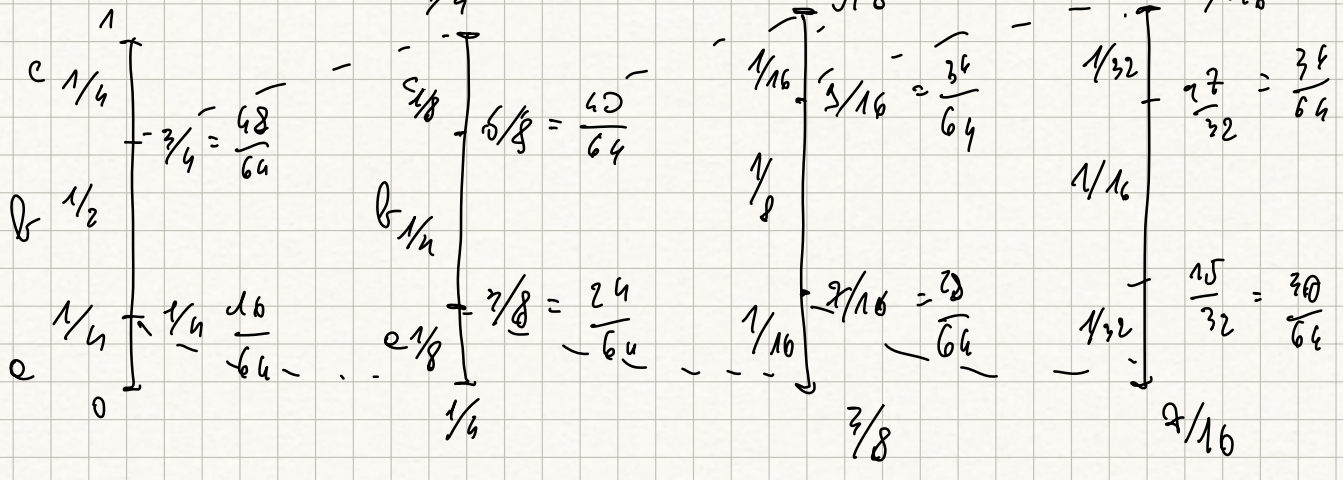
$$\frac{10}{27} \rightarrow \frac{20}{27} < 1 \rightarrow 0$$

$$\frac{20}{27} \rightarrow \frac{40}{27} > 1 \rightarrow 1 \rightarrow \frac{13}{27}$$

$$\frac{13}{27} \rightarrow \frac{26}{27} < 1 \rightarrow 0$$

$\langle 011111, 4 \rangle \; p(a) = p(c) = \frac{1}{4}, p(b) = 1/2 >$

$$\frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} = \frac{16+8+4+2+1}{64} = \frac{31}{64}$$



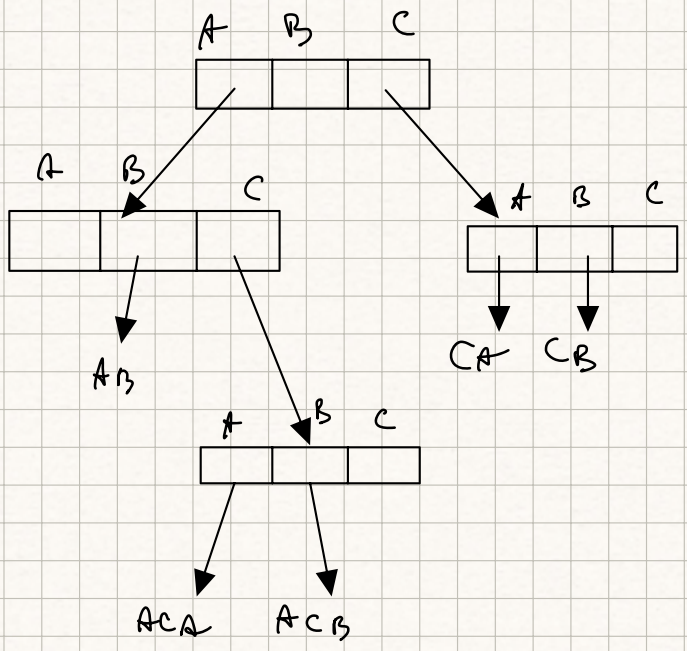
$S_1 = 1/2$

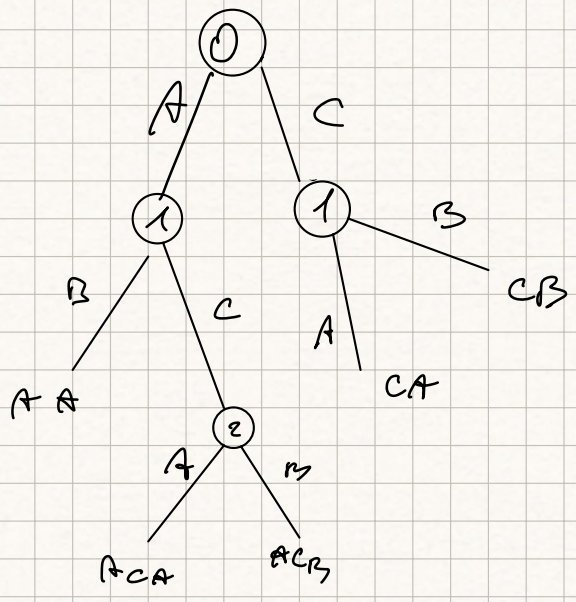
$l_1 = 1/4$

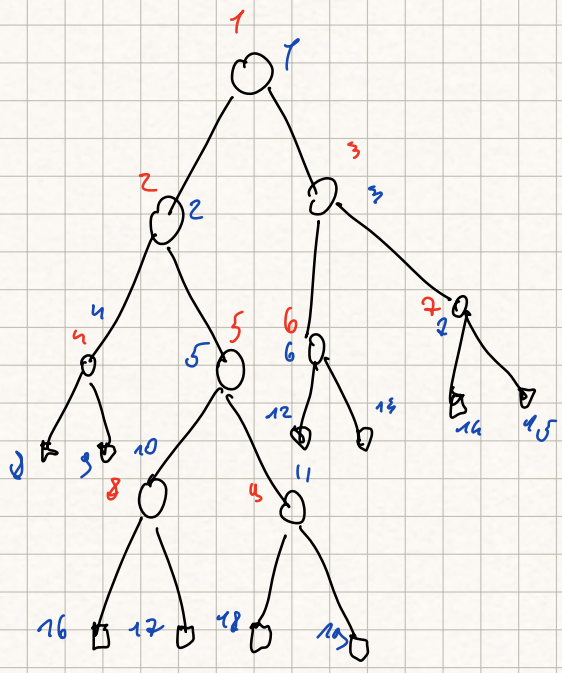
$S_2 = 1/4$

$l_2 = 3/8$

$D = \{AB, ACA, ACB, CA, CB\}$ Succinctly encode the tree







0 dummy nodes

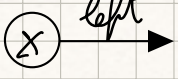
1 real nodes

$B =$ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

1 1 1 1 1 1 0 0 1 1 0 0 0 0 0 0 0 1

1 2 3 4 5 6 7 8 9

Pseudocode: return the length of the left most path



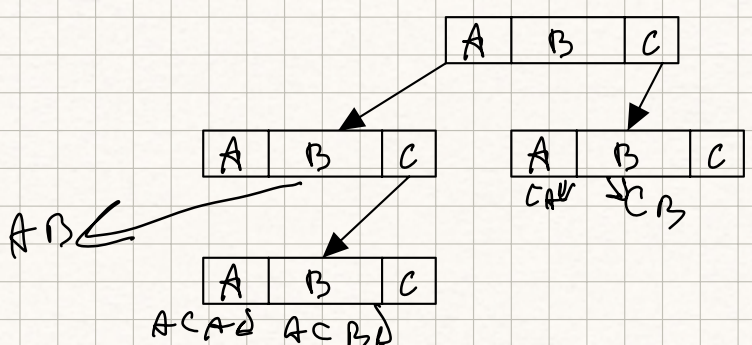
$B[2x] \neq 0$

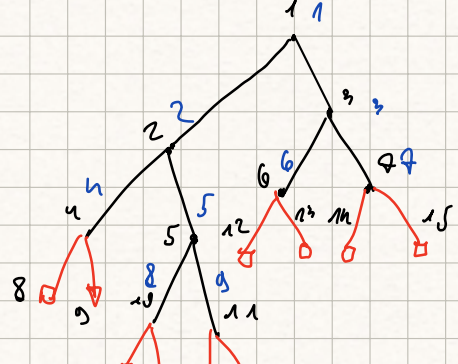
```
left = 0
x = 1
while (B[2x] ≠ 0) do
        left++;
        x = 2rank_1(2x)
return left
```

write a pseudo-code that check if the tree contains a path (A, B) 

Question #2 [scores 4+3+3] Given the ordered set of strings S={AB,ACA,ACB,CA,CB}

- Build the UNcompacted trie T for S by assuming an alphabet of 3 characters (S = {A,B,C}) and branching implemented via arrays.
- Show how to succinctly encode the structure of T in a binary array B, by assuming that pointers to strings are leaves of the tree T, and branching nodes are the internal nodes of T;
- Write a pseudo-code that, given a binary array succinctly encoding the structure of a binary tree (with its corresponding rank/select data structures), establishes the length of its left-only path (or, equivalently, the depth of its leftmost NULL pointer.





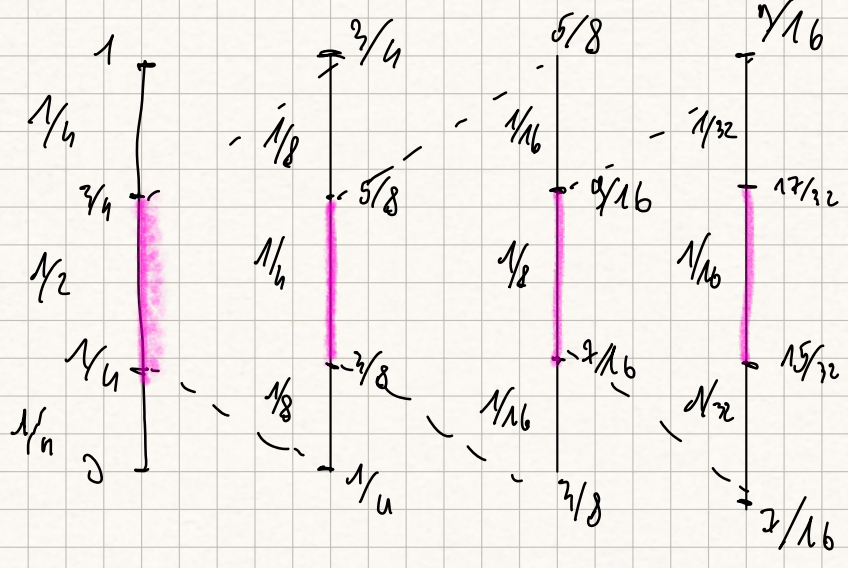
16 17 18 19

B = 1 1 1 1 1 1 1 0 0 1 1 0 0 0 0 0 0 0 0 0

Decode the compressed sequence <4, 011110> produced by arithmetic code, by assuming probabilities P[a]=P[c]=1/4 and P[b]=1/2.

0 1 1 1 1 0 ⇒ $\frac{1+2+4+8}{32} = \frac{15}{32} \simeq 1/2$

1 1/2 1/4 1/8 1/16 1/32 1/64



Show the first 8 codewords of the (s,c)-code with s=3 and c=1, hence s+c = 4 (briefly explain your calculations).

s+c = 4 = 2^b ⇒ b = 2

s = 00 01 10 c = 11

	codeword (3,1)
1	00
2	01
3	10
4	11 00
5	11 01
6	11 10
7	11 11 00
8	11 11 01

Decompress the 8th integer encoded via Elias-Fano in the two arrays:

L = 01|11|00|01|01|00|11|11|00|11|00 and H= 110 110 10 0 10 10 10 110 0 0 10 0 0 0 0 0 where the original encoding of the integers is in 6 bits. (hint: derive first the number of keys, and then the length of the low and high part)

Access(8) : take the 8-th group in L = 11 } h = 0110, l = 11 = 27

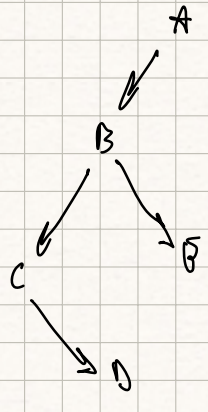
$Select_1(8) = 14 - 8 = (6)_2 = 110$

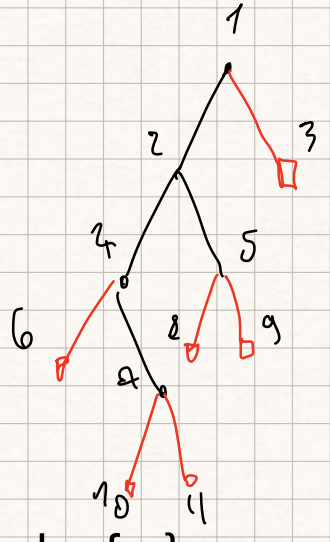
b = 6

l = L/#1 in H = 2

h = 4

[rank 4]. Show the binary succinct encoding of the tree T = { aàb (left child); bàc (left child); bàe (right child); càd (right child) } of root “a”.



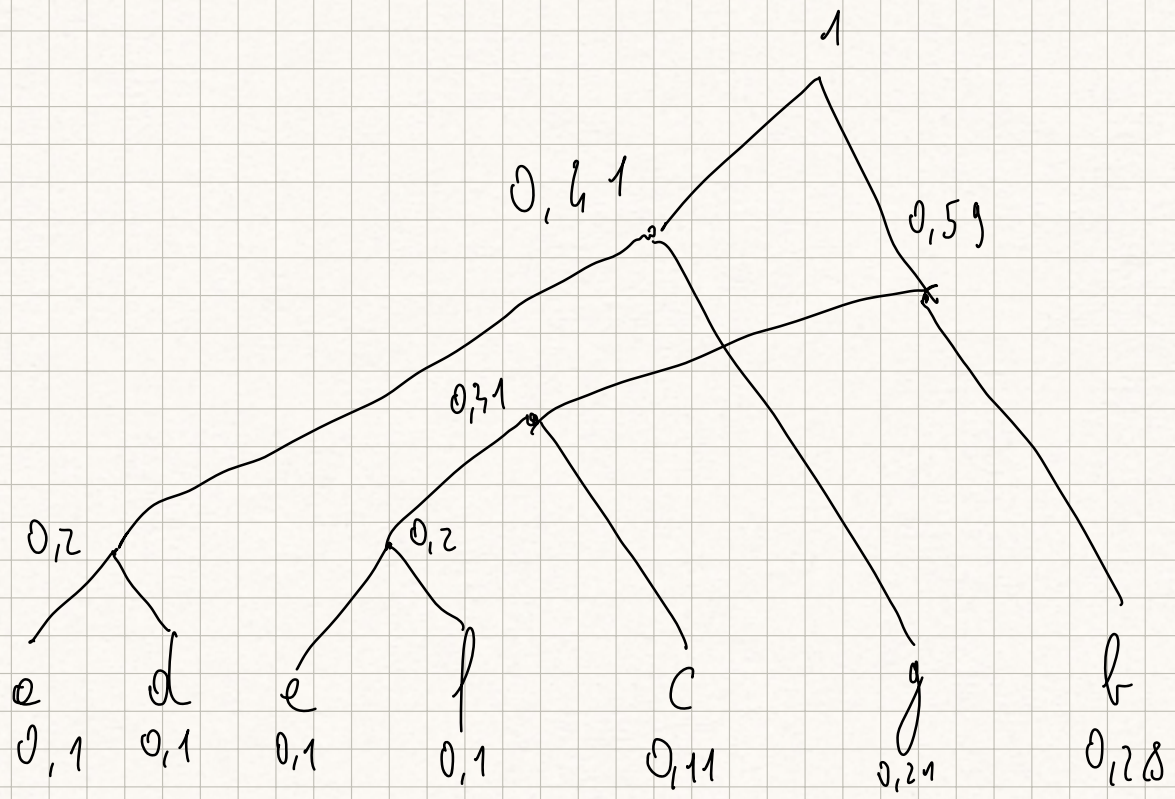


B = 11011010000

Given the symbols {a,b,c,d,e,f,g} occurring in a text with frequencies f(a) = f(d) = f(e) = f(f) = 0.1, f(b) = 0.28, f(c) = 0.11, f(g)= 0.21.

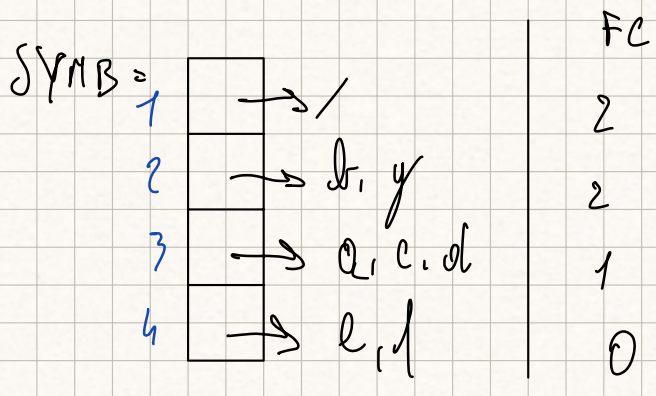
• Compute FC[] and SYMB[] tables of the Canonical Huffman code

• Decode the first 2 symbols of the compressed sequence:
11001....



num (1,4) = [0, 2, 3, 1]

FC [i] = $\frac{FC[i+1] + num[i+1]}{2}$

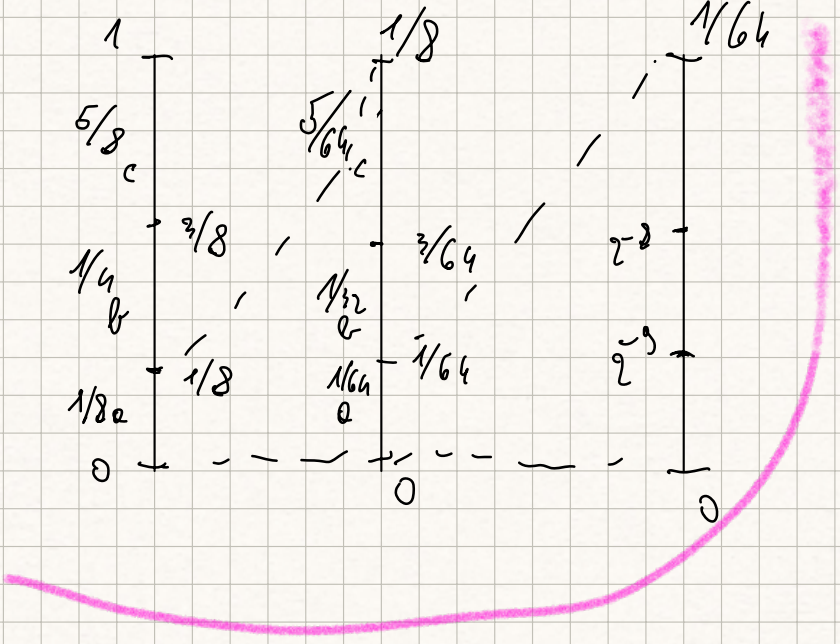


1 1 | 0 0 1

v = first_bit() l = 1
while v < FC[e]
  v = 2 · v + next_bit()
  l++
return symb [l, v - FC[l]]

v = 1 l = 1 1 < 2
v = 2 · 1 + 1 = 3 l = 2 3 ≥ 2
symb [2, v - FC[2]] = g
3 - 2
v = 0 l = 1 0 < 2 symb [3, 0] = e
v = 2 · 0 + 0 l = 2 0 < 2
v = 2 · 0 + 1 l = 3 1 ≥ 1

Given: p(a) = 1/8, p(b) = 1/4, p(c) = 5/8. Specify which is the length in bits of the text T = aabbaa, if it is compressed via Arithmetic coding. (Hint: work with negative powers of two.)

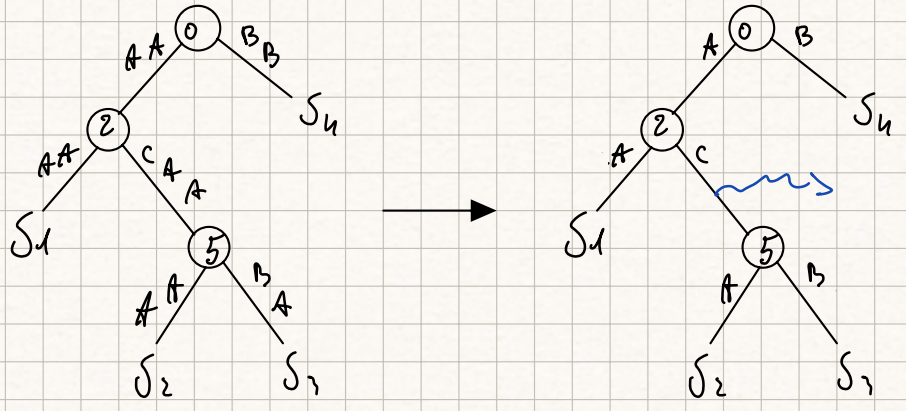


$$d = \lceil \log_2 \frac{2}{S_m} \rceil$$

$$S_m = \left(\frac{1}{8}\right)^4 \cdot \left(\frac{1}{4}\right)^2 = 2^{-12} \cdot 2^{-4} = 2^{-16}$$

$$d = \left\lceil \log_2 \frac{2}{2^{-16}} \right\rceil = \lceil \log_2 2^{17} \rceil = 17 \text{ bit}$$

Given the binary strings S={aaaaa, aacaaaa, aacaaba, bb}. Build the Patricia Tree for S and show how to search for the lexicographic position of the string P=aacbba among the strings of the set S.



P = A A c B B A

0 2 5

prendo S_2, calcolo $LCP[P, S_2] = AAC$, quindi cerco

P tra S_2 e S_3

Given the integer sequence S = (1, 2, 3, 4, 6, 8, 9), show how Interpolative Coding compresses the “first three” integers according to its algorithm.

$\langle l = 1, r = 7, low = 1, high = 9 \rangle$

$m = \left\lfloor \frac{l+r}{2} \right\rfloor = 4 \qquad S_m = 4$

$a = low + (m - l) = 1 + 3 = 4$

$b = high - (r - m) = 9 - 3 = 6$

$S_m - a = 4 - 4 = 0$ $\qquad (00)_2$

$\lceil \log_2 (b - a + 1) \rceil = 2$

$\langle l = 1, r = 3, low = 1, high = 3 \rangle$

$m = \left\lfloor \frac{1+3}{2} \right\rfloor = 2 \qquad S_m = 2$

$a = low + m - l = 1 + 1 = 2$

$b = high - (r - m) = 3 - 1 = 2$

$S_m - a = 0$ m $\lceil \log_2 (b - a + 1) \rceil = 1$

non emette simboli

$\langle l = 5, r = 7, low = 5, high = 9 \rangle$

$m = \left\lfloor \frac{l+r}{2} \right\rfloor = 6 \qquad S_m = 8$

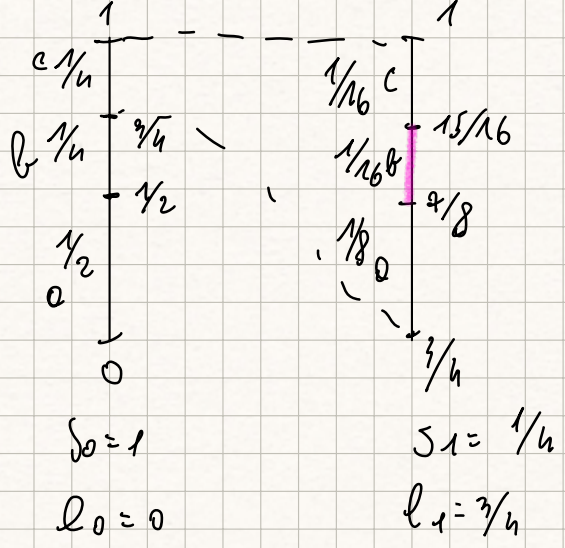
$a = low + m - l = 5 + 6 - 5 = 6$

$b = high - (r - m) = 9 - 1 = 8$

$S_m - a = 2$ m $\lceil \log (b - a + 1) \rceil = 2$ $\qquad (10)_2$

Let us given the probabilities: p(a) = 1/2, p(b)=p(c)=1/4. Decompress the first 2 symbols of the Arithmetic coded bit sequence: 111.

$\underset{1/2}{1}\ \underset{1/4}{1}\ \underset{1/8}{1} = \frac{7}{8}$



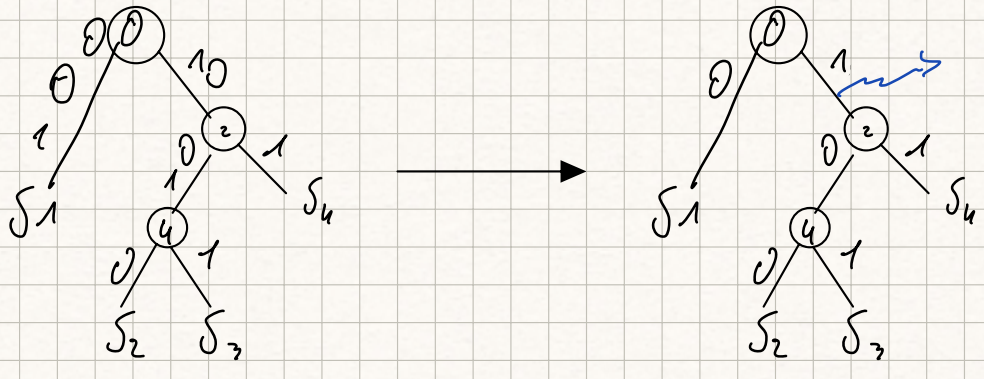
c b

$S_0 = 1$ $\quad S_1 = 1/4$

$\ell_0 = 0$ $\quad \ell_1 = 3/4$

Given the binary strings S={001, 10010, 10011, 101}.

- Build the Patricia Trie for S
- Show how to search for the lexicographic position of the string P=110 among the strings of the set S.
- Propose a succinct encoding of the Patricia Trie of S that allows navigation in constant time per traversed edge.
- Simulate the downward search for P=110 in this succinct encoding.



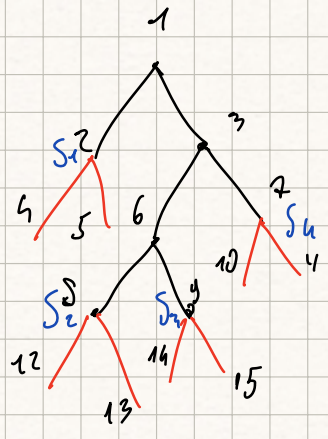
P = 110

0 2

prendo S_2

$LCP(S_2, P) = 1$

cerco a destra di S_4



B = 1|11|00|11|11|00|00 00

$\Sigma = [0, S_1, 2, 4, S_4, S_2, S_3]$

] Let us given the set of strings

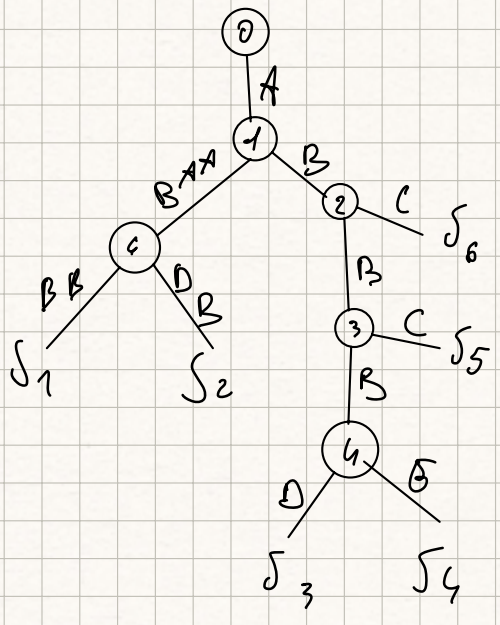
S={ baci, baco, boss, buono, buu, costi, costola}

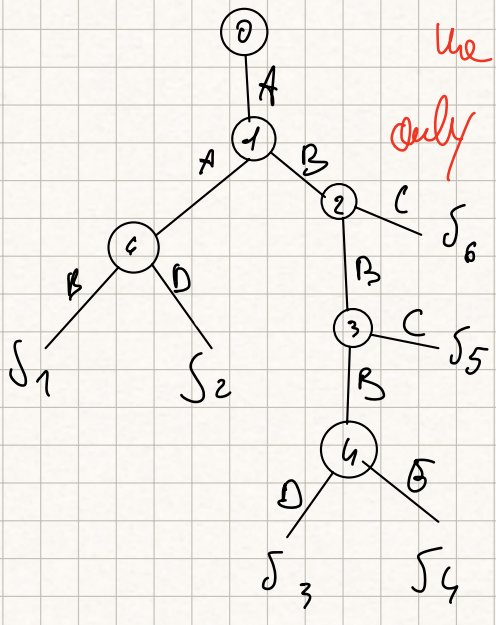
- Show the Front-coding compression of S
- Show the Locality-preserving front-coding compression of S, setting c=1 and counting (for simplicity) previous characters and digits for the “backward copy-check” in the LPFC-algorithm

<0,BACI><3,O><1,OSS><1,UONO><2,U><0,COSTI><4,OLA>

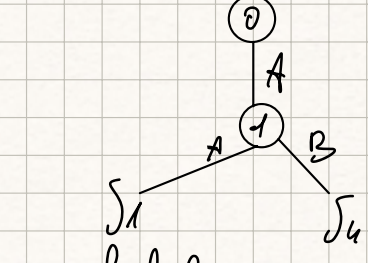
Let us given the set of strings S={ aaabbb, aaabdb, abbbd, abbbe, abbc, abc}

- Describe a two-level index that uses in internal memory a Patricia trie on 2 strings
- Show how it is searched the string “aaabcb” in that 2-level index

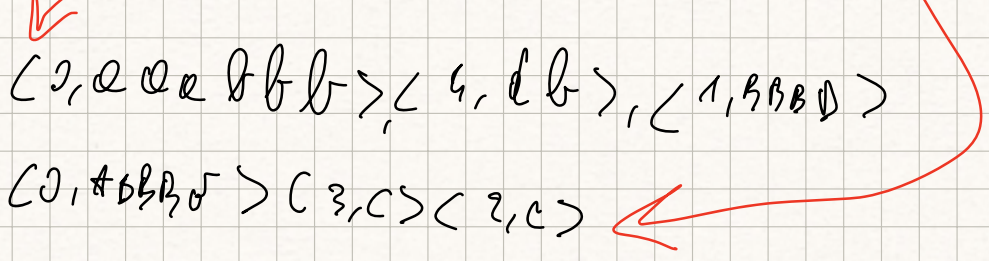




use least to save only 2 strings in internal memory



aaabbb abbbbe



Given a sequence of integers S = (1, 6, 15, 18, 19, 20, 21), encode them using (by motivating which sequence is compressed):

- Rice code with k=3
- Interpolative coding (the first 3 integers only: the ones in positions 2, 4, 6)

$x \geq 1$ $R_k(x)$

$q = \left\lfloor \frac{x-1}{2^k} \right\rfloor$ quotient → encoded as $U(q+1)$

$r = x - 1 - 2^k \cdot q$ remainder → encoded as bin (r) k bits

Jos encoding: 1, 5, 9, 3, 1, 1, 1

1) $k = 3$

$R_3(1) = \begin{cases} q = \left\lfloor \frac{1-1}{8} \right\rfloor = 0 \\ r = 1 - 1 - 8 \cdot 0 = 0 \end{cases}$ $U(1) = 1$ $R_3(1) = 1000$

000

$R_3(5) = \begin{cases} q = \left\lfloor \frac{5-1}{8} \right\rfloor = 0 \\ r = 5 - 1 - 8 \cdot 0 = 4 \end{cases}$ $U(1) = 1$ $R_3(3) = 1100$

100

$R_3(9) = \begin{cases} q = \left\lfloor \frac{9-1}{8} \right\rfloor = 1 \\ r = 9 - 1 - 8 \cdot 1 = 0 \end{cases}$ $U(2) = 01$ $R_3(9) = 01000$

000

$R_3(3) = \begin{cases} q = \left\lfloor \frac{3-1}{8} \right\rfloor = 0 \\ r = 3 \cdot 1 - 8 \cdot 0 = 2 \end{cases}$ $U(1) = 1$ $R_3(3) = 1010$

010

$1, 6, 15, 18, 19, 20, 21$

$l = 1, r = 7, low = 1, high = 21$

$m = \left\lfloor \frac{l + r}{2} \right\rfloor = 4 \qquad S_m = 18$

$a = low + m - l = 1 + 4 - 1 = 4$

$b = high - (r - m) = 21 - (7 - 4) = 18$

$S_m - a = 14$ codifico in $\lceil \log_2 b - a + 1 \rceil = 4$ bit $(1110)_2$

$\langle 1, 3, 1, 17 \rangle$

$m = \left\lfloor \frac{1+3}{2} \right\rfloor = 2 \qquad S_m = 6$

$a = 1 + 2 - 1 = 2$

$b = 17 - (3 - 2) = 16$

$S_m - a = 4$ in $\lceil \log_2 (16 - 2 + 1) \rceil = 4$ $(0100)_2$

$\langle 5, 7, 19, 21 \rangle$

$m = \left\lfloor \frac{5+7}{2} \right\rfloor = 6 \qquad S_m = 20$

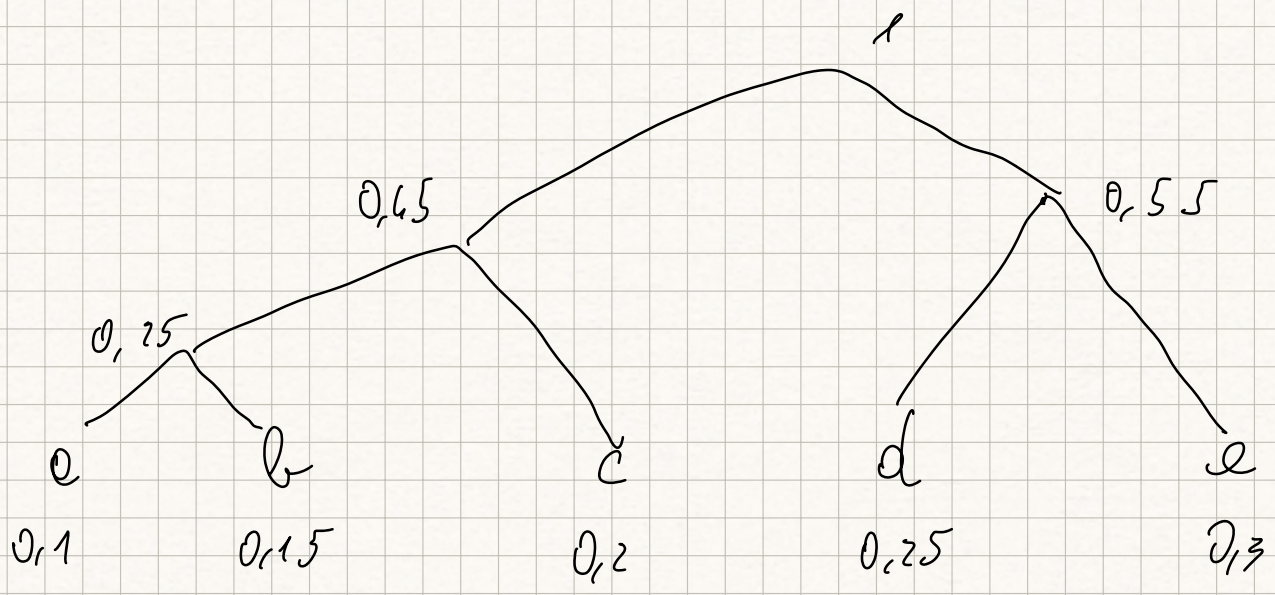
$a = 14 + 6 - 5 = 20$

$b = 21 - (7 - 6) = 20$

no bit emitted

Given the probabilities p(a)=0.1; p(b)=0.15; p(c)=0.2; p(d)=0.25; p(e)=0.3,

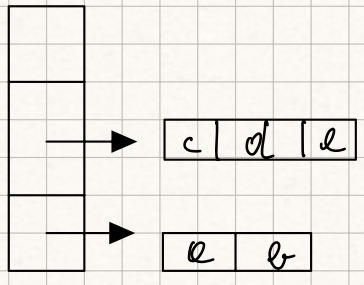
1. Construct the Canonical Huffman code, showing the steps followed by the algorithm.
2. Then use it to decode the bit sequence 1001001, showing each decoding step.



num(1,3) = [0, 3, 2]

FC[3] = 0

symb(1,3) =



$FC[i] = \frac{FC[i+1] + num[i+1]}{2} =$

$FC[2] = \frac{0+2}{2} = 1$

$FC[1] = \frac{1+3}{2} = 2$ we have now all in first level

1 0 | 0 0 1 | 0 0 1

v = 1 l = 1 v < FC[l]
1 < 2

v = 2·v + nextbit()
l++

v = 2 + 0 2 < 1
l = 2

symb[l, v - FC[l]] = d
2 1

v = 0 l = 1 1 < 2
v = 1 l = 2 1 < 1

symb[l, v - FC[l]] = c
2 0

$v = 0 \quad l = 1 \quad 0 < 2$

$v = 0 \quad l = 2 \quad 0 < 1$

$v = 1 \quad l = 3 \quad 1 < 0$ symbol $[l_3, v - PC[l_1]] = b$

Given the sequence of integers S = 1 2 3 4 6 10 11,

- compress the entire S via Interpolative Coding
- compress the gaps between the consecutive integers of S by PForDelta using b=2 bits, and base = 1.

$< l = 1, r = 7, low = 1, high = 11 >$

$m = \left\lfloor \frac{l+r}{2} \right\rfloor = 4 \qquad S_m = 4$

$a = low + (m - l) = 1 + 4 - 1 = 4$

$b = high - (r - m) = 11 - (7 - 4) = 8$

$4 \le 4 \le 8 \qquad S_m - a = 4 - 4 = 0$

$m \lceil \log_2 (b - a + 1) \rceil = 3 \qquad 000$

$< l = 1, r = 3, low = 1, high = 3 >$

$m = 2 \quad S_m = 2$

$a = 1 + 2 - 1 = 2$

no emission

$b = 3 - (3 - 2) = 2$

$< l = 5, r = 7, low = 5, high = 11 >$

$m = \frac{12}{2} = 6 \qquad S_m = 10$

$a = 5 + (6 - 5) = 6$

$b = 11 - (7 - 6) = 10$

$S_m - a = 10 - 6 = 4 \quad m \lceil \log_2 5 \rceil = 3$

100

$< l = r = 5, low = 5, high = 9 >$

$< l = r = 7, low = 11, high = 11 >$

$m = 5 \quad S_m = 6$

$a = 5 + (5 - 5) = 5$

$b = 9 - (5 - 5) = 9$

$S_m - a = 1 \quad m \lceil \log_2 5 \rceil = 3$

001

$m = 7 \quad S_m = 11$

$a = 11 + (7 - 7) = 11$ no bit emitted

$b = 11 - (7 - 7) = 11$

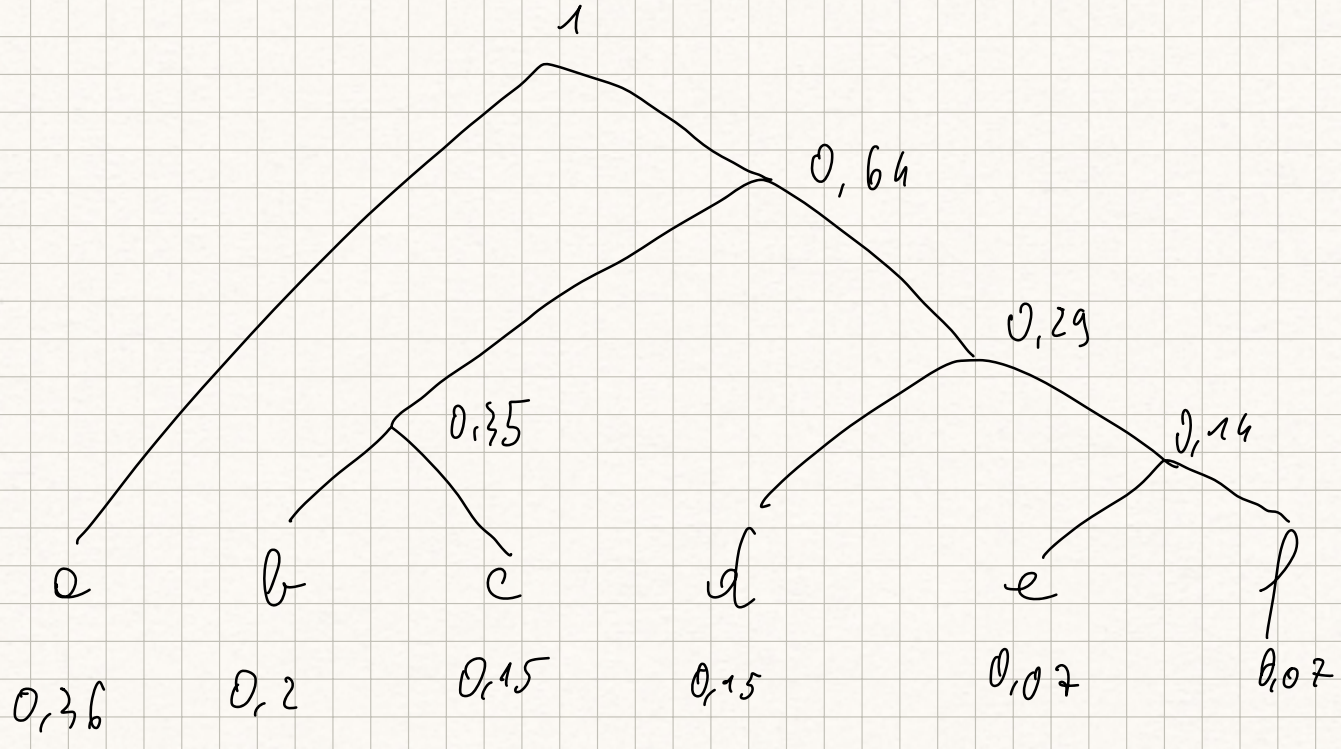
1 1 1 1 2 4 1
0 0 0 0 1 3 0

$b = 2 \quad bere = 1$

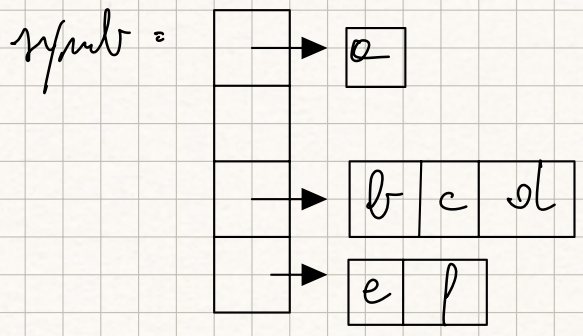
00 00 00 00 01 11 00

00 → 0
01 → 1
10 → 2
11 → esp

Let us given symbols and probabilities: p(f) = p(e) = 0.07, p(c)=p(d) = 0.15, p(b)=0.2, p(a)=0.36. Construct the Canonical Huffman code, showing how it is obtained step-by-step.



num [1, 4] = [1, 0, 3, 2]



FC[4] = 0

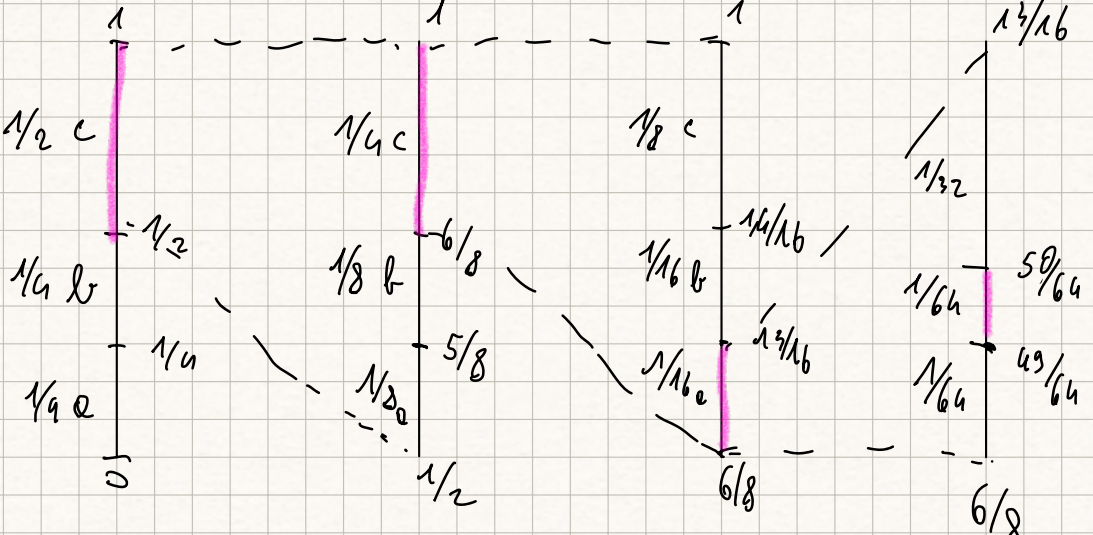
FC[3] = $\frac{0+2}{2} = 1$

FC[2] = $\frac{1+3}{2} = 2$

FC[1] = $\frac{2+0}{2} = 1$

Given the text T = ccab, compress it via Arithmetic coding, using as probabilities of the symbols their frequency of occurrence in T.

P(a) = 1/4 P(b) = 1/4 P(c) = 1/2



$\delta_0 = 1$ $\delta_1 = 1/2$ $\delta_2 = 1/4$ $\delta_3 = 1/16$

$\ell_0 = 0$ $\ell_1 = 1/2$ $\ell_2 = 6/8$ $\ell_3 = 6/8$

$\lceil \log \frac{2}{\delta_n} \rceil = \lceil \log \frac{2}{1/64} \rceil = 7$

codifico: $\ell_n + \frac{\Delta_n}{2}$

$= 49/64 + 1/128 = \frac{99}{128}$

$2 \cdot \frac{99}{128} = \frac{99}{64} \geq 1$ esci $1, \frac{99}{64} - 1 = \frac{35}{64}$

$2 \cdot \frac{35}{64} = \frac{70}{64} \geq 1$ esci $1, \frac{70}{64} - 1 = \frac{6}{64} = \frac{3}{32}$

$2 \cdot \frac{3}{32} = \frac{6}{32} < 0$ esce 0,

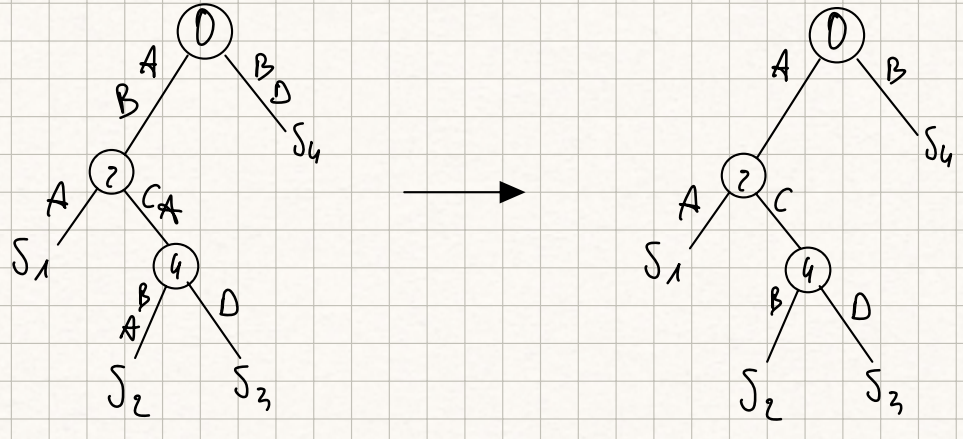
$2 \cdot \frac{6}{32} = \frac{12}{32} < 0$ esce 0,

$2 \cdot \frac{12}{32} = \frac{24}{32} < 0$ esce 0,

$2 \cdot \frac{24}{32} = \frac{48}{32} \geq 1$ esce 1 $\quad \frac{48}{32} - 1 = \frac{16}{32} = \frac{1}{2}$

$2 \cdot \frac{1}{2} = 1 \geq 1$ esce 1

Build a Patricia trie on the set of strings S={aba, abcaba, abcad, bd} and show how it is executed a lexicographic search for the string P1 = acdac, and P2=aabBuild a Patricia trie on the set of strings S={aba, abcaba, abcad, bd} and show how it is executed a lexicographic search for the string P1 = acdac, and P2=aab



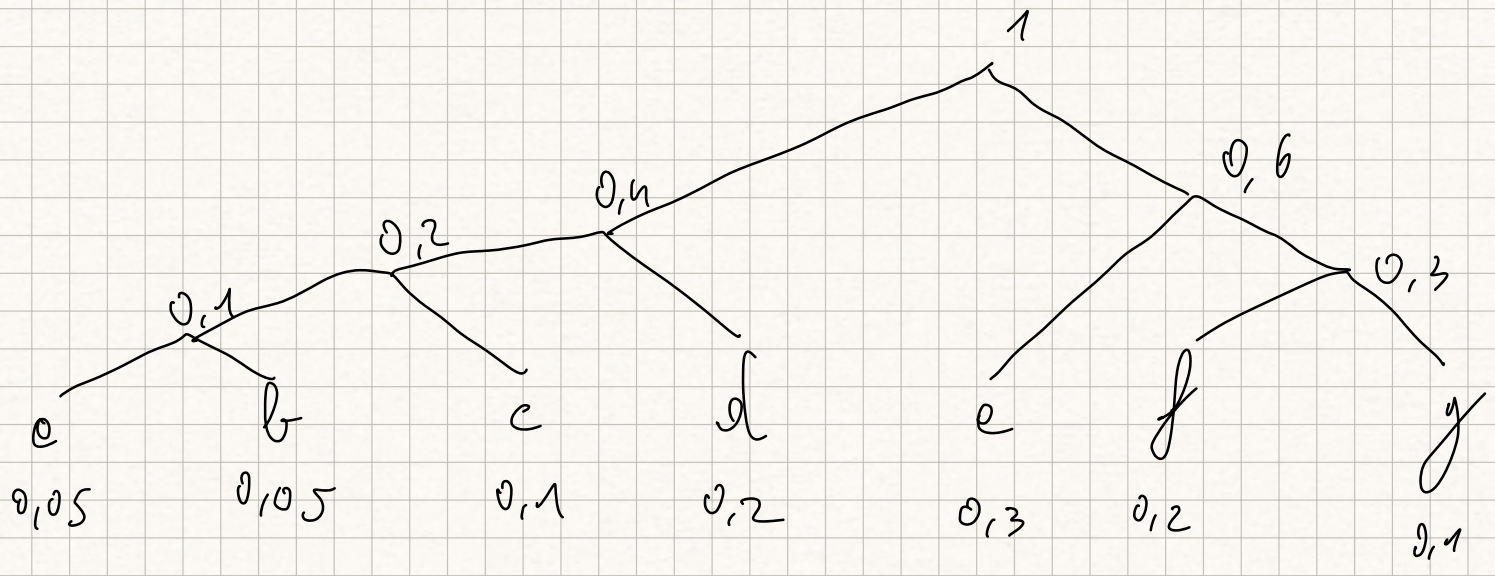
$P_1 = ACDAC$ (c0 the $(S_1, P_1) = A$ cerco tra S_2 e S_4

$P_2 = AAB$ (c0 the $(S_1, P_1) = A$ cerco prima di S_1

Let us given the symbols and their probabilities: p(a) = p(b) = 0.05, p(c) = p(g) = 0.1, p(d) = p(f) = 0.2, p(e) = 0.3.

a. Compute the Canonical Huffman code for this distribution

b. Decode the first 3 symbols of the coded sequence 100010110111100…



num (1, 4) = [0, 2, 3, 2]

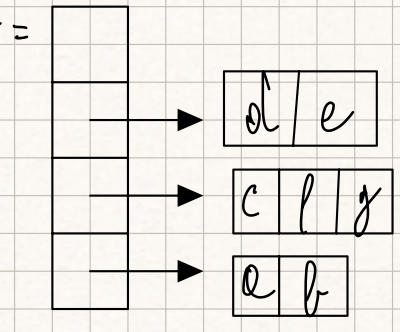
FC [4] = 0

$FC[3] = \frac{0+2}{2} = 1$

$FC[2] = \frac{1+3}{2} = 2$

$FC[1] = \frac{2+2}{2} = 2$

symb =



1 0 0 0 1 0 1 1 0 1 1 1 0 0

v = 1 l = 1 1 < 2

v = 2+0 l = 2 2 ≥ 2

symb [2, 2 - FC[2]] = d

v = 0 l = 1 0 < 2

v = 0 l = 2 0 < 2

$v=1 \quad l=3 \quad 1 \geq 1$

$\text{mult}[3, 1-1] = C$

$v=0 \quad l=1 \quad 0 < 2$

$v=1 \quad l=2 \quad 1 < 2$

$v=2 \quad l=3 \quad 3 \geq 1$

$\text{mult}[3, 3-1] = Y$

Given the string S=baobab, compute the Suffix Array, and then derive the LCP array by using the linear-time algorithm seen in class, showing the amount of characters really compared when determining the various LCPs.

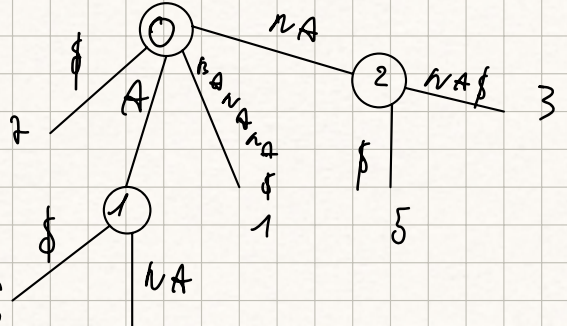
BAOBAB$ → BAOBAB$, AOBAB$, OBAB$, BAB$, AB$, B$, $

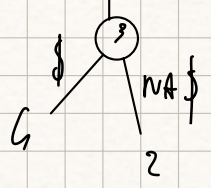
1 2 3 4 5 6 7

	SA	SUFF
	7	$
0	5	AB$
1	2	AOBAB$
0	6	B$
1	4	BAB$
2	1	BAOBAB$
0	3	OBAB$

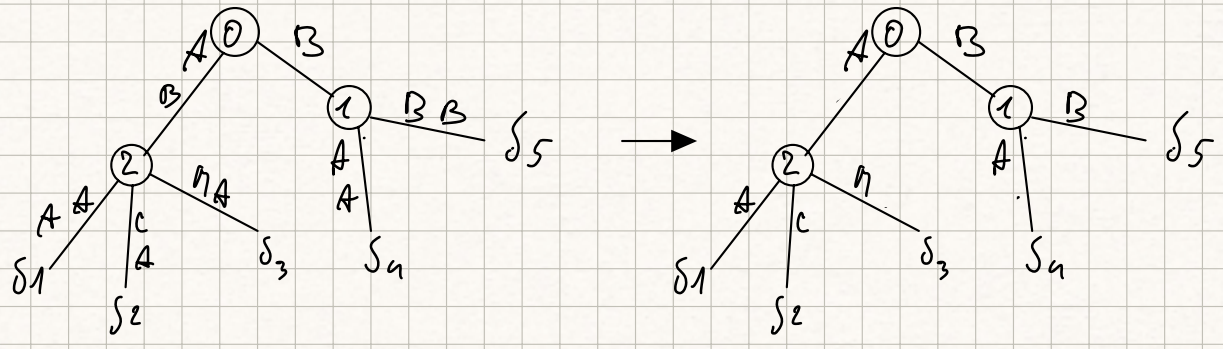
1. Build the suffix tree of the string T = BANANA$.

BANANA$, ANANA$, NANA$, ANA$, NA$, A$, $





Given the set of strings S = {abaa, abca, abma, baa, bbb}, build a Patricia trie and show the steps for the lexicographic search of the strings P1 = aaa, P2 = abb



P1 = AAA prendo S_1 $LCP(S_1, P_1) = A$ quindi cerco prima di S_1

P2 = ABB ho mismatch in posizione 3, prendo qualsiasi stringa sotto il nodo ②

$LCP(S_1, P_2) = 2 = AB$ quindi cerco tra S_1 S_2

Given the sequence of integers (11, 14, 16, 17, 19, 20, 21, 31), show how to encode them based on

- Elias-Fano Code
- PForDelta with base = 11, and b=3

$n = 32$ $b = \lceil \log_2 n \rceil = 5$ $l = \lceil \log_2 \frac{u}{n} \rceil = 2$ $h = b - l = 5 - 2 = 3$

01011	11
01110	14
10000	16
10001	17

L = 11 10 00 01 11 00 01 11

H = 0 0 10 10 11 10 11 10 0 10

10011	19
10100	20
10101	21
11111	31
h	l

base = 1 b = 3

. : 0, 3, 5, 6, 8, 9, 10, 20

gap : 0, 3, 2, 1, 2, 1, 1, 10

000 011 010 001 010 001 001 111

000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	exep.

Given the sequence of integers (11, 14, 16, 19, 20, 21, 22), show how to encode them based on

- Elias-Fano Code
- Interpolative Code (just one level of recursion, hence 3 numbers)

$u = 23 \quad b = \lceil \log_2 u \rceil = 5 \quad l = \lceil \log_2 \frac{u}{m} \rceil = 4 \quad h = 5 - 4 = 1$

01011	11
01110	14
10000	16
10011	19
10100	20
10101	21
10110	22

L = 1101 0011 00 01 10

H = 0 01010 10 11100 00

$\langle l = 1, 2 = 8, low = 11, high = 22 \rangle$

$m = \left\lfloor \frac{1+7}{2} \right\rfloor = 4 \quad \delta_m = 19$

$a = low + (m - l) = 11 + (4 - 1) = 14$

$b = high - (r - m) = 22 - (8 - 4) = 19$

$\delta_m - a = 5 \quad m \lceil \log_2 (19 - 14 + 1) \rceil = 3$

$(101)_2$

$\langle l = 1, 2 = 3, low = 11, high = 18 \rangle$

$m = \left\lfloor \frac{1+3}{2} \right\rfloor = 2 \quad \delta_m = 14$

$a = 14 + (2 - 1) = 12$

$b = 18 - (3 - 2) = 17$

$\delta_m - a = 2 \quad m \lceil \log_2 (17 - 12 + 1) \rceil = 3$

$(010)_2$

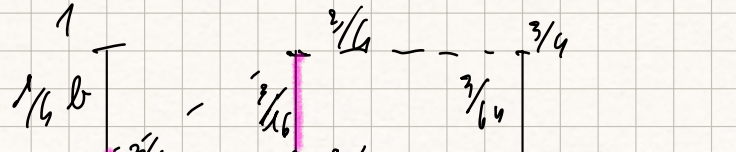
$\langle l = 5, 2 = 7, low = 20, high = 22 \rangle$

$m = \frac{5+7}{2} = 6 \quad \delta_m = 21$

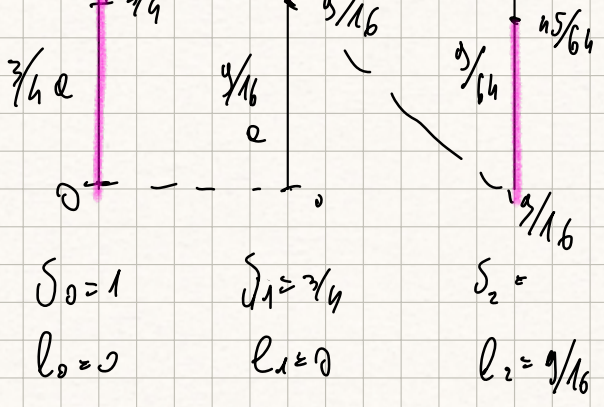
$a = 20 + (6 - 5) = 21$ no bit emitted

$b = 22 - (7 - 6) = 21$

Given the string T = aba, and the probabilities P(a) = ¾ and P(b) = ¼, show the result of the Arithmetic Coding applied on T. (hint: Please work with the dyadic fractions, not change them into reals.)



$d = \lceil \log_2 \frac{2}{9/64} \rceil = 4$



$S_0 = 1$ $S_1 = 3/4$ $S_2 =$

$l_0 = 0$ $l_1 = 0$ $l_2 = 9/16$

$9/16 + \frac{9}{64} \cdot 2 = \frac{81}{128}$

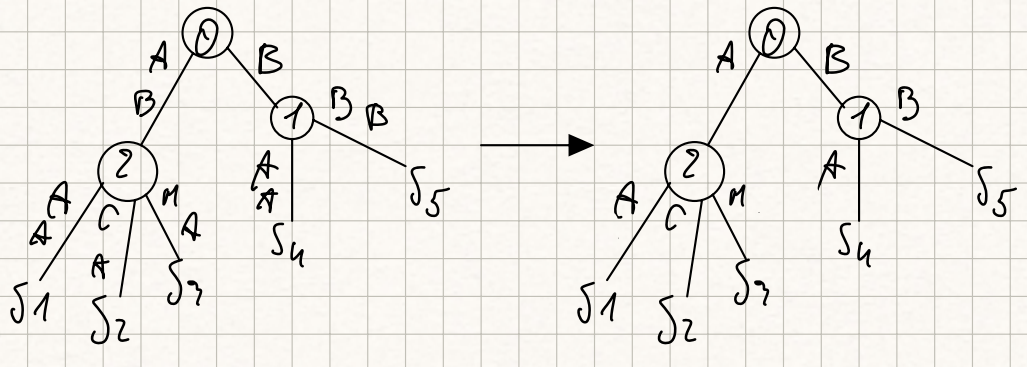
$\frac{81}{128} \cdot 2 = \frac{81}{64} > 1$ entra 1 $\frac{81}{64} - 1 = \frac{17}{64}$

$\frac{17}{64} \cdot 2 = \frac{17}{32} < 1$ entra 0

$\frac{17}{32} \cdot 2 = \frac{17}{16} > 1$ entra 1 $\frac{17}{16} - 1 = \frac{1}{16}$

$\frac{1}{16} \cdot 2 = \frac{1}{8} < 1$ entra 0

Given the set of strings S = {abaa, abca, abma, baa, bbb}, build a Patricia trie and show the steps for the lexicographic search of the strings P1 = aaa, P2 = abb.



$P_1 = AAA$ prende S_1

0 2

$LCP(AAA, ABAA) = A$

mismatch in posizione 2 vado a sinistra di S_1

$P_2 = ABB$

0 2 mismatch, prendo qualsiasi foglia del sottoalbero ② $LCP(S_1, P_2) = AB$

cerco tra S_1 e S_2

Given the string S=annapanna, compute its parsing LZ77 and

LZ78

ANNAPANNA

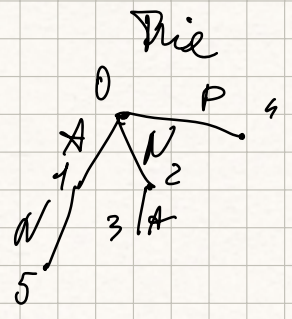
<0,0,A>, <0,0,N> <1,1,A> <0,0,P> <5,4,EOF>

LZ77

<1,N> <3,EOF>

ANNAPANNA

<0,A> <0,N> <2,A> <0,P>



Question #4 [scores 3+3+2]. Given the string T formed by your Matricola, and hence consisting of 6 digits:

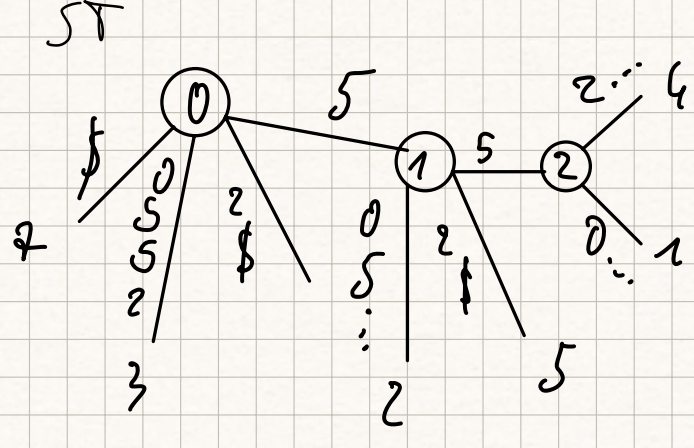
• Show the suffix array of T, in which every digit is interpreted as a symbol;

Form the string P as given by the two middle digits of T (i.e. if the Matricola is 123456, then P=34). Then describe the algorithm that efficiently counts the occurrences of P in T.

• Comment on the time complexity of the counting algorithm as a function of n (= T's length), p (= P's length) and the number occ of P's occurrences.

		LCP
1	550552$	0
2	50552$	0
3	0552$	0
4	552$	1
5	52$	1
6	2$	2
7	$	

ST



Question #2 [scores 5]. Take your Matricola (of 6 digits), change every occurrence of 0 with 1 (if any), and then interpret each digit as an integer gap, and finally derive an increasing integer sequence by summing those gaps: namely, if the Matricola is 120304, then you transform it into 121314, and then you get the corresponding integer sequence as 1, 3 (=1+2), 4 (=1+2+1), 7 (=1+2+1+3), 8 (=1+2+1+3+1), 12(=1+2+1+3+1+4).Compress the resulting increasing integer sequence with Elias-Fano.

$551552 \Rightarrow 5, 10, 11, 16, 21, 23$ $\quad u=24 \quad b=\lceil \log_2 24 \rceil = 5$

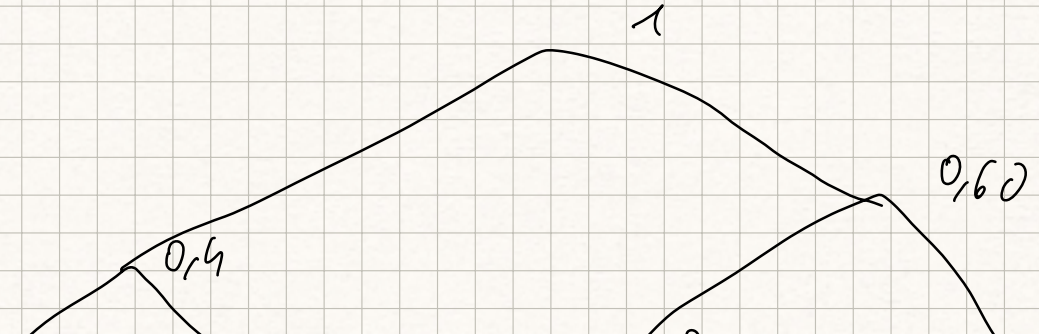
00101	5
01010	10
01011	11
10000	16
10101	21
10111	23

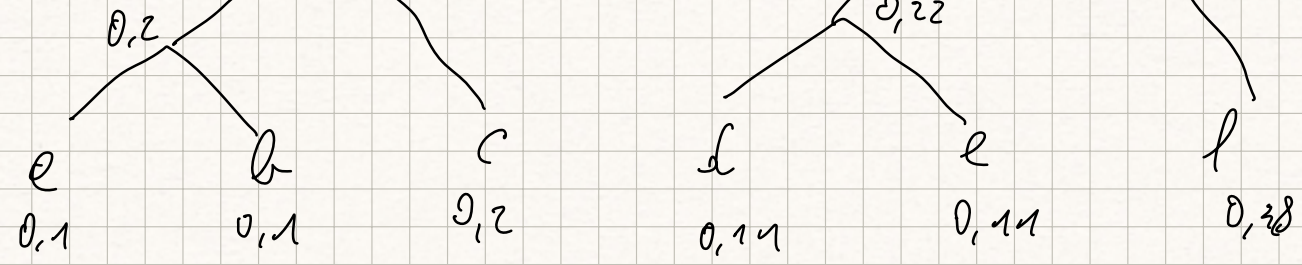
$\ell = \lceil \log_2 \frac{u}{n} \rceil = 2 \quad h = 3$

$L = 01\,10\,11\,00\,01\,11$

$U = 0\,10\,110\,0\,10\,110\,00$

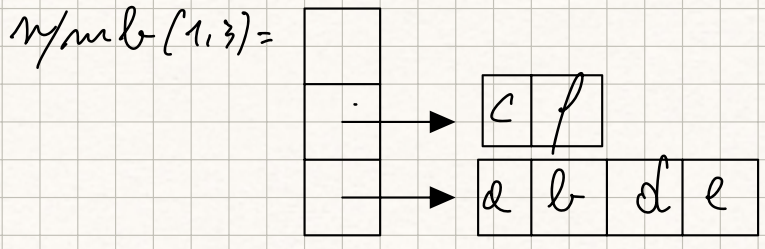
Question #1 [scores 2+3+4]. Given the symbols and their probabilities: p(a) : p(b) = 0.1, p(c)= 0.2, p(d)=p(e)=0.11, p(f)=0.38.
Compute the Huffman code for this distribution.
Compute the Canonical variant of the Huffman code (by sorting alphabetically the letters in every SYMB's list).
Decode the first 2 symbols of the coded sequence 11010.





min (1,3) = [0, 2, 4]

FC[3] = 0



FC[2] = $\frac{0+4}{2} = 2$

FC[1] = $\frac{2+2}{2} = 2$

v = 1 l = 1 1 < 2

v = 2·1 + 1 l = 2 3 > 2

min [2, 3-2] = f

v = 0 l = 1 0 < 2

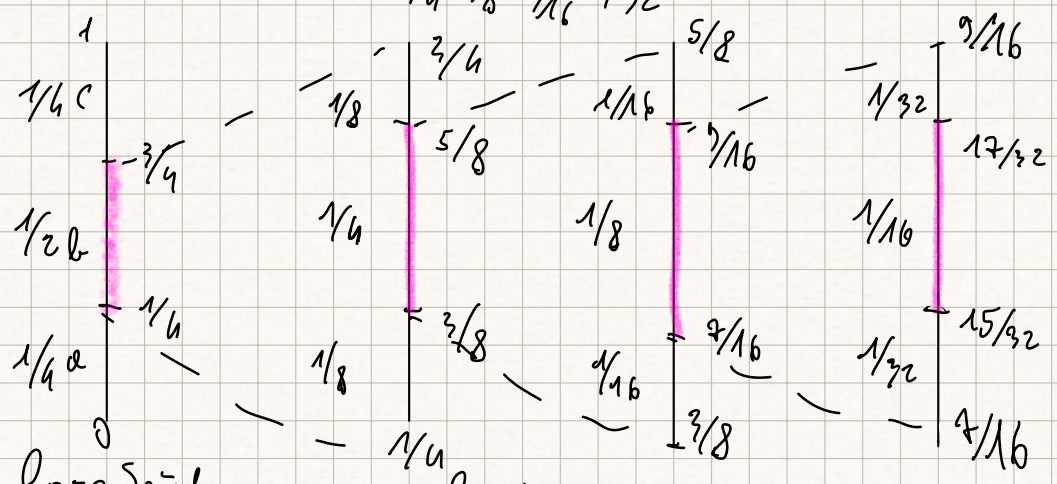
v = 2·0 + 1 l = 2 1 < 2

v = 2·1 + 0 l = 3 2 > 0

min [3, 2] = d

Decode the compressed sequence <4, 011110> produced by arithmetic code, by assuming probabilities P[a]=P[c]=1/4 and P[b]=1/2.

0 1 1 1 1 0 = $\frac{1+2+4+8}{32} = \frac{15}{32}$



Burrows-Wheeler

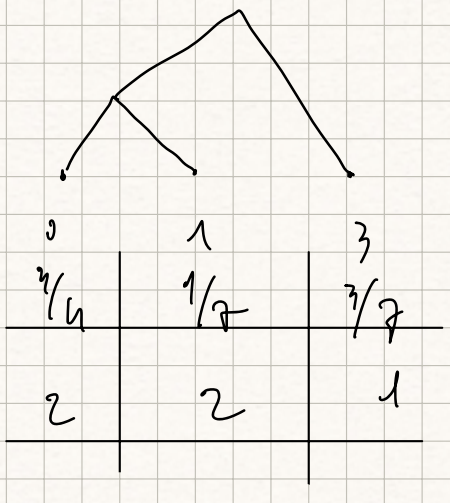
$l_1 = 1/4$ $S_1 = 1/2$ $l_2 = 3/8$ $S_2 = 1/4$ $l_3 = 7/16$ $S_3 = 1/16$

text = ANANANANAN#

1 2 3 4 5 6 7 8 9 10

		L
10	#A	A
9	A#	N
1	AN	#
7	ANA#	N
5	ANANA#	N
3	ANA	N
2	N	A
8	NA#	A
6	NANA#	A
4	NA	A

P(0) = 3/7 P(1) = 1/7 P(3) = 3/7



Body: 00 101 1 0000
0 3 1 3 3 0 0

① BWT

② MTF

③ RLE Ø

④ Canonical Huffman

Preamble

-) portion # = 3
-) L = {A, N, T}

1 3 1 1 3 3 1 1 1

A T T T N A A A A → 0 3 1 3 3 0 0

$L = \{A, N, T\}$ (1 2 3)

$L = \{T, A, N\}$ (1 2 3)

$L = \{N, T, A\}$ (1 2 3)

num (1,2) = 1, 2

-) symb (1,2) : → 3, → 0,1

-) FC[2] = 0

-) $FC[1] = \frac{0+2}{2} = 1$

Decompression of B-W

B = 00 | 1 | 01 | 11 0000

0 1 3

v = 0 l = 1 0 < 1

v = 0 l = 2 0 ≥ 0

num [2, 0-0] = 0

v = 1 l = 1 1 ≥ 1

num [1, 0] = 3

v = 0 l = 1 0 < 1

v = 1 l = 2 1 ≥ 1

num [2, 1] = 1

num (1,2) = → 3, → 0,1

FC [2] = 0

FC [1] = 1

0 3 1 3 3 0 0

00 11 100

2 3 4

1 2 3

-1 -1 -1

1	3	1	1	3	3	1	1	1
A	T	T	T	M	A	A	A	A

L = (A, M, T)

L = (T, A, M)

L = (M, T, A)

L = (A, M, T)

F L

A
A T
A #
A T
A T
A M
M A
T A
T A
A

T = A M A T A T A T A #

A = 000 0000

B = 000 0010

C = 000 1100

D = 000 1111

E = 100

F = 1010

2 level - index, in which B=2 strings

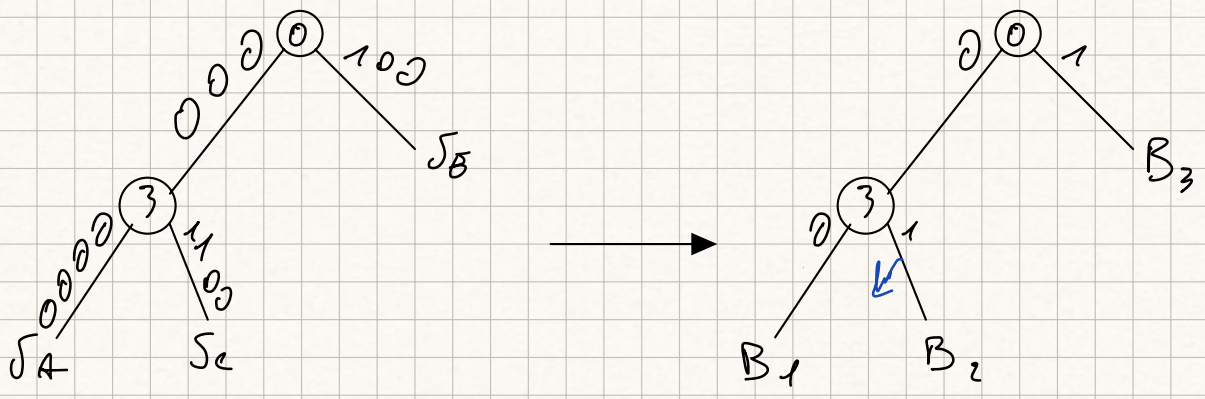
Patricia Trie in int. memory and

FC on disk

$B_1 = \boxed{\langle 0, 0000000 \rangle \langle 5, 10 \rangle}$

$B_2 = \boxed{\langle 0, 000110 0 \rangle \langle 5, 10 \rangle}$

$B_3 = \boxed{\langle 0, 100 \rangle \langle 2, 10 \rangle}$



lexicographic search P = 000 1 0 1 $\quad lcp(P,C) = \frac{0001}{4}$

↑ 0 ↑ 3 mismatch

We search in B_1

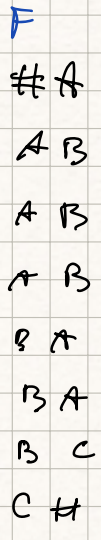
$P_1 = 10\$ \quad lcp(P_1, E) = 2$ Jump in B_3 and scan

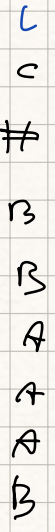
$P_2 = 10\# \quad lcp(P_2, E) = 2$ Jump in B_3 and scan

If you are to the left you have to scan the previous block, if you are to the right you have to scan the right block

Given the string S = abababc show the result of the algorithmic pipeline BWT + MTF + RLE0 + Huffman, where RLE0 is the application of the special RunLengthEncoding algorithm over 0-runs and the Wheeler code.

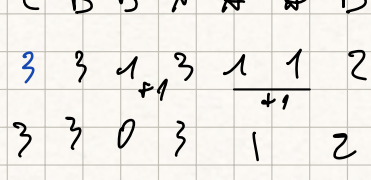
sorting →

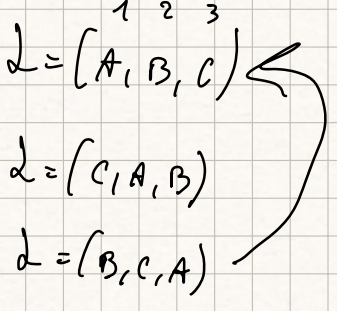




Preamble

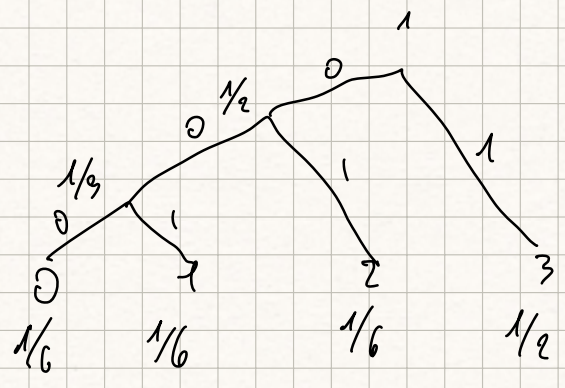
in pos. 2



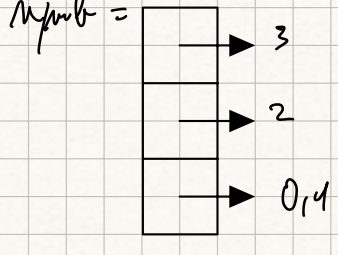


$P(0) = 1/6 \quad P(1) = 1/6$

$P(2) = 1/6 \quad P(3) = 1/2$



num (1, 3) = (1, 1, 2)



FC[3] = 0

$FC[2] = \frac{0+2}{2} = 1$

$FC[1] = \frac{1+1}{2} = 1$

BODY = 3 3 0 3 1 2
1 1 000 1 001 01

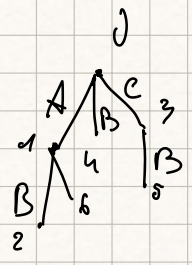
Question #1 [scores 4+6+6] Given the string S = aabcbcbada, show the:

- LZ77 parsing of S.
- LZ78 parsing of S, along with the auxiliary data structure used to compute it.
- Canonical Huffman encoding of S

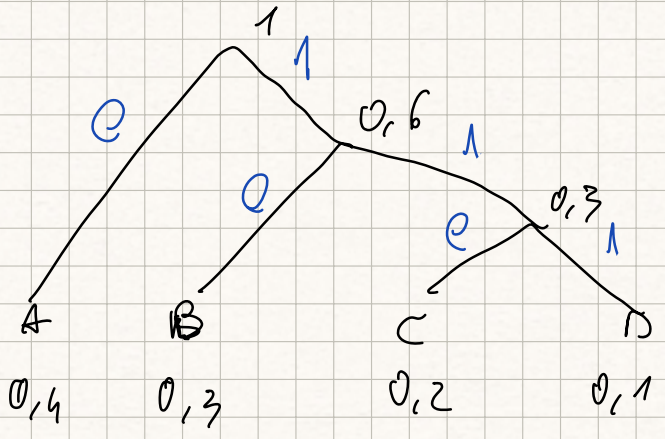
A | A | B | C | B | C | B | A | D | A

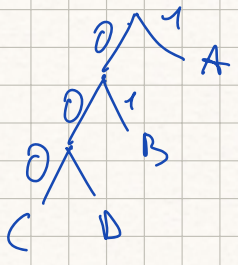
1) ⟨0,0,A⟩ ⟨1,1,B⟩ ⟨0,0,C⟩ ⟨2,3,A⟩ ⟨0,0,D⟩ ⟨2,1,EOF⟩

2) ⟨0,A⟩ ⟨1,B⟩ ⟨0,C⟩ ⟨0,B⟩ ⟨3,B⟩ ⟨1,D⟩ ⟨1,EOF⟩



P(A) = 4/10 P(B) = 3/10 P(C) = 2/10 P(D) = 1/10





nuova (1,3) = (1,1,2)

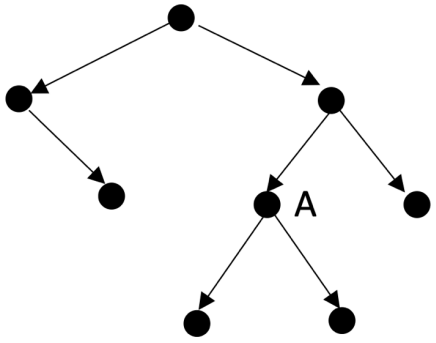
lista (1,3) = A, B, C, D

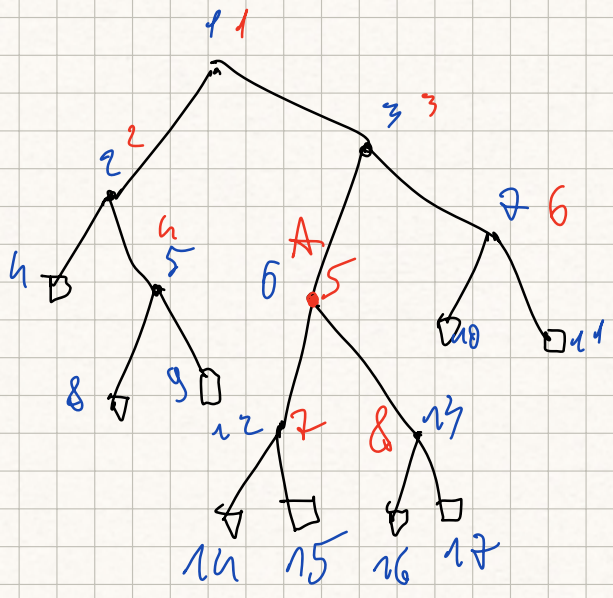
FC[3] = 0

FC[2] = (0+2)/2 = 1

FC[1] = (1+1)/2 = 1

Question #4 [scores 4] Show the succinct representation of the following binary tree. Then show how to use this representation to navigate from the root to the node labelled A, and then back to the parent of A.





B = 1 1 1 0 1 1 0 0 0 0 1 1 0 0 0 0

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Rank$_1$[2x] → go left if B[2x] == 1

Rank$_1$[2x+1] → go right if B[2x+1] == 1

Rank$_1$[2·1+1] = 3 B[3] = 1 ok

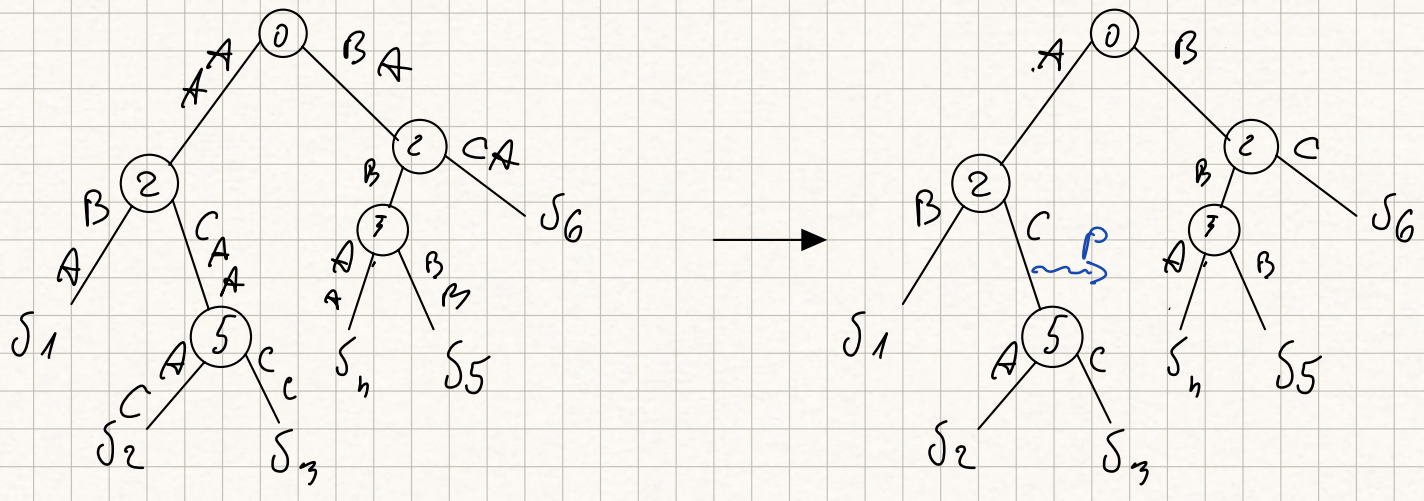
Rank$_1$[2·3] = 5 B[6] = 1 ok

going to parent $\left\lfloor \frac{\text{Select}_1(5)}{2} \right\rfloor = 3$

Given the ordered set of strings:

S = { AABA, AACAAAC, AACAACC, BABAA, BABBB, BACA }

- Build the Patricia trie PT for S.
- Show the steps executed to lexicographically search for the pattern P = AACBACD in the set S by means of PT.



P = A A C B A C D primo S_3 $LCP(P, S_3) = \frac{AAC}{3}$ Cerco tra S_3 e S_4

Given the string “ABABABC” compress it by using the pipeline BWT + MTF + RLE0 + Huffman, where MTF counts letter’s positions from 0, and RLE0 uses the Wheeler’s code.

T = A B A B A B C $

$ A B A B A B C

C $ A B A B A B

B C $ A B A B A

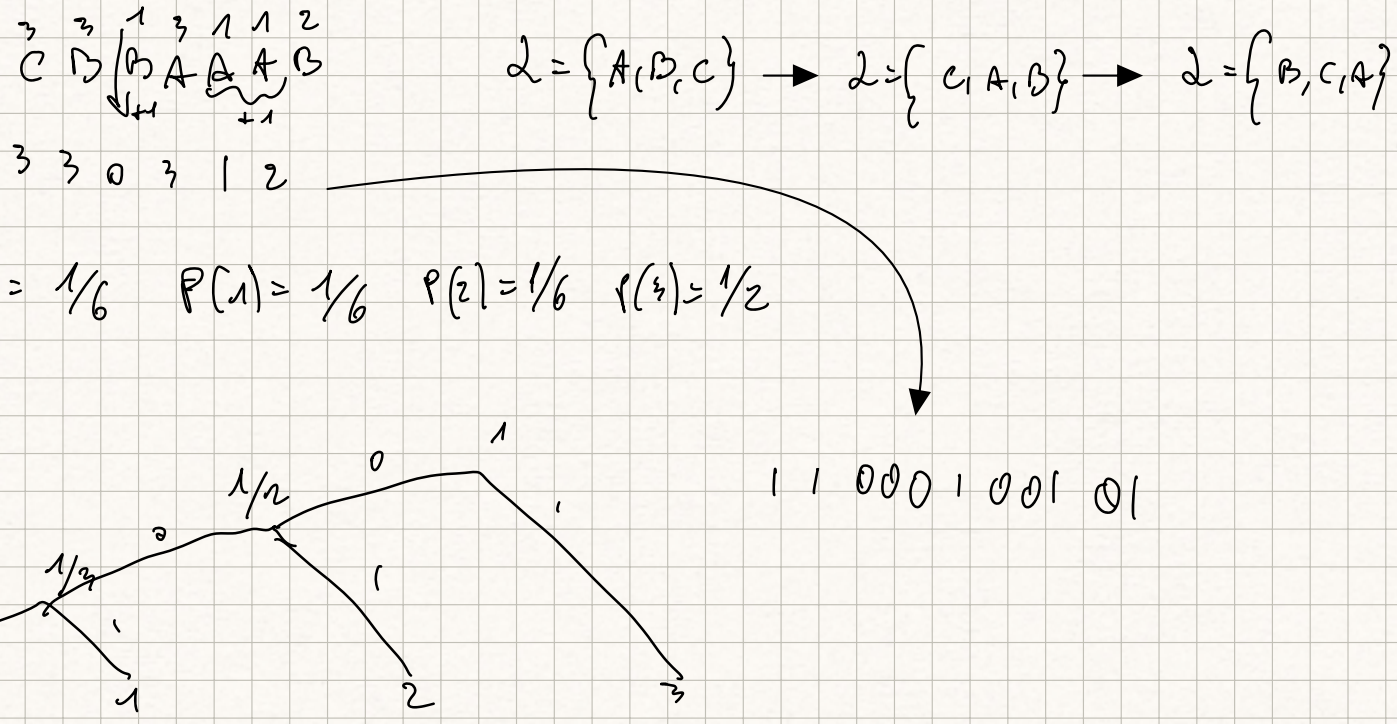
A B C $ A B A B

B A B C $ A B A

A B A B C $ A B

B A B A B C $ A

P	L
$ A	C
A B	$ # pos 2
A B	B
A B	B
B A	A
B A	A
B C	A
C $	B



Given the following SYMB and FC arrays for a Canonical Huffman code: SYMB[1] = [], SYMB[2] = [A,B], SYMB[3] = [C,D,E], SYMB[4] = [F,G]

FC=[2,2,1,0]

Decompress the first 2 letters of the following compressed bit sequence: 000111...

$v = 0 \quad l = 1 \quad 0 < 2$

$v = 0 \quad l = 2 \quad 0 < 2$ → symb[0, 1-0] = A

$v = 0 \quad l = 3 \quad 0 < 1$

$v = 1 \quad l = 4 \quad 1 \geq 0$

$v = 1 \quad l = 1 \quad 1 < 2$

$v = 2 \cdot 1 + 1 = 3 \quad l = 2 \quad 3 \geq 2$ → symb[2, 3-2] = B

Given the string “CABABCA” compress it by using the pipeline BWT + MTF + RLE0 + Huffman, where MTF counts letter’s positions from 0, and RLE0 uses the Wheeler’s code.

CABABCA#

#CABABCA

A#CABABC

CA#CABAB

BCA#CABA

ABCA#CAB

BABCA#CA

ABABCA#C

F	L	
#	C	A
A	#	C
A	B	C
A	B	B
B	A	A
B	C	A
C	A	B
C	A	#

in pos 8

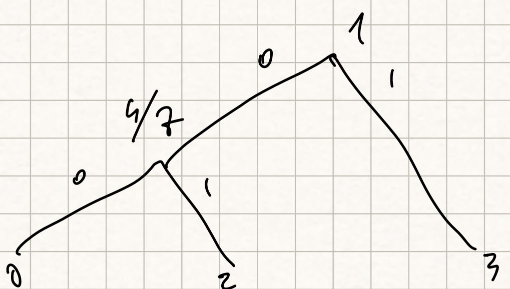
<8, A C C B A A B>

1 3 1 3 3 1 2

0 3 0 3 3 0 2

$\alpha = \{A, B, C\} \rightarrow \alpha = \{C, A, B\} \rightarrow \alpha = \{B, C, A\}$

$P(0) = 3/7 \quad P(2) = 1/7 \quad P(3) = 3/7$

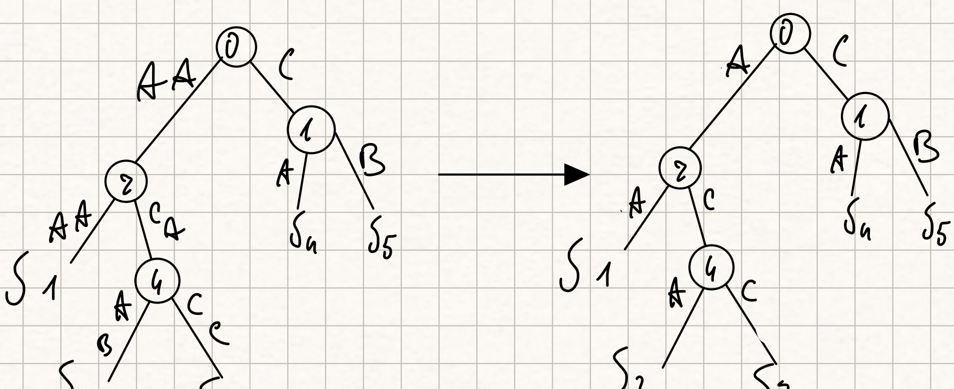


00 1 00 11 00 01

Given the set of strings

S = {AAAA, AACAAB, AACACC, CA, CB}

- Built the Patricia trie PT for S
- Show the steps executed to lexicographically search for P1 = ABC in the PT



$P_1 = ABC$ Prendo S_2

0 2

$LCP(S_2, P_1) = 1$ Cerco

tra S_3, S_4

Given the list of integers A = {2, 3, 5, 6, 8, 10, 11}, use Interpolative Code and emit the encoding of the “first” three integers 6, 3 and 10.

$l = 1, r = 7, low = 2, high = 11$

$m = \left\lfloor \frac{l+2}{2} \right\rfloor = 4 \quad S_m = 6$

$a = 2 + (4 - 1) = 5$

$b = 11 - (7 - 4) = 8$

$S_m - a = 1 \quad in \lceil \log_2(b - a + 1) \rceil = 2$

$(01)_2$

$l = 1, r = 3, low = 2, high = 5$

$m = 2 \quad S_m = 3$

$a = 2 + (2 - 1) = 3$

$b = 5 - (3 - 2) = 4$

$3 - 3 = 0 \quad m \lceil \log_2(4 - 3 + 1) \rceil = 1$

0

$l = 5, r = 7, low = 7, high = 11$

$m = 6 \quad S_m = 10$

$a = 7 + (6 - 5) = 8$

$b = 11 - (7 - 6) = 10$

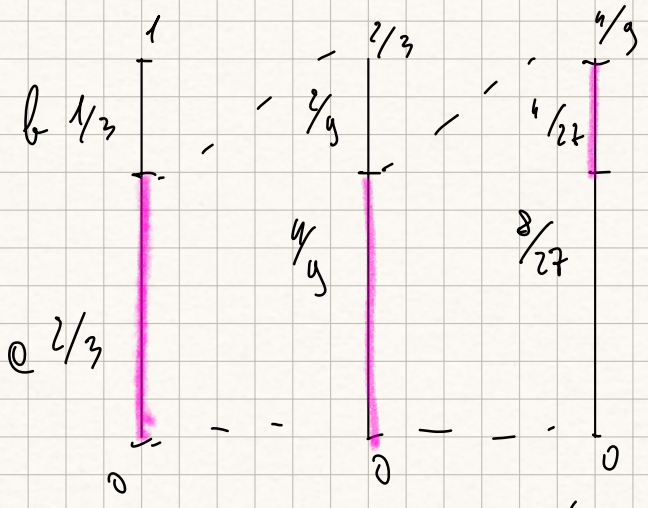
$S_m - a = 2 \quad m \lceil \log_2(10 - 8 + 1) \rceil = 2$

$(10)_2$

Given the text T=aab, encode it using Arithmetic Coding and the

semi- static model. (Associate the intervals from the bottom to the top as “a”, “b”; and use the fractions directly.)

$P(a) = 2/3 \qquad P(b) = 1/3$



$\frac{S_3}{2} + l_2 = \frac{10}{27}$

$\left\lceil \log_2 \frac{2}{S_3} \right\rceil = \left\lceil \log_2 \frac{27}{2} \right\rceil = 4$

$l_0 = 0 \quad S_0 = 1$

$l_1 = 0 \quad S_1 = 2/3$

$S_2 = 4/9$

$l_2 = 0$

$S_3 = 4/27$

$l_3 = 8/27$

$\frac{20}{27} < 1 \rightarrow 0$

$\frac{40}{27} \geq 1 \rightarrow 1 \qquad x = \frac{40}{27} - 1 = \frac{13}{27}$

$\frac{26}{27} < 1 \rightarrow 0$

$\frac{S^2}{27} > 1 \rightarrow 1$

$S = (1, 2, 3, 4, 6, 8, 9)$ interpolative code

$l = 1$, $r = 7$, $low = 1$, $high = 9$

$m = \left\lfloor \frac{1+7}{2} \right\rfloor = 4$ $\quad S_m = 4$

$a = 1 + (4 - 1) = 4$ $\quad S_m - a = 0$ $\quad m \lceil \log_2(6 - 4 + 1) \rceil = 2$

$b = 5 - (7 - 4) = 6$ $\quad (00)_2$

$l = 1$, $r = 3$, $low = 1$, $high = 3$

$m = 2$ $\quad S_m = 2$

$a = 1 + (2 - 1) = 2$ $\quad S_m - a = 0$ $\quad m \lceil \log_2(2 - 2 + 1) \rceil = 0$ no bit emitted

$b = 3 - (3 - 2) = 2$

$l = 5$, $r = 7$, $low = 5$, $high = 9$

$m = 6$ $\quad S_m = 8$

$a = 5 + (6 - 5) = 6$ $\quad S_m - a = 2$ $\quad m \lceil \log_2(8 - 6 + 1) \rceil = 2$ $\quad 10$

$b = 9 - (7 - 6) = 8$

$L = 11\,10\,00\,01\,11\,00\,11$

$H = \overset{0}{0}\,\overset{1}{0}\,\overset{2}{1}\,\overset{3}{0}\,\overset{4}{1}\,\overset{5}{1}\,\overset{6}{0}\,\overset{7}{1}$

$n = \#1 \text{ in } H = 8$

$l = \frac{L}{n} = \frac{16}{8} = 2$

h	l
0 0 0	1 1
0 0 1	1 0
0 1 0	0 0
0 1 1	0 1
1 0 0	1 1
1 0 1	0 0
1 1 0	0 1
1 1 1	1 1

$h = \log_2 \# 0 = 3$

a a b a b c

LET $\langle 0,0,a \rangle \langle 1,1,b \rangle \langle 2,2,c \rangle$

LESS $\langle 0,a \rangle \langle 1,1 \rangle \langle 0,b \rangle \langle 2,2 \rangle \langle 0,c \rangle$

$\langle 0, \text{dest} \rangle$ new dest

$\langle d, l \rangle$ copy