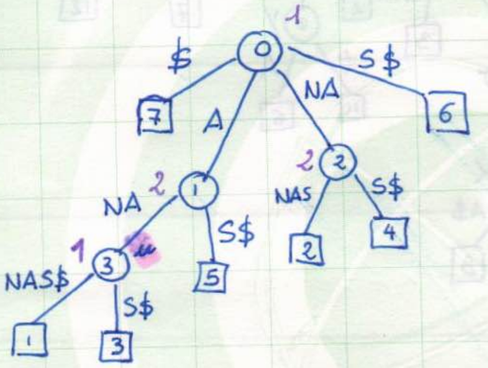
EXERCISES

6 dicembre 2018

EXCERCISE 1

Let T = ANANAS \$. Compute LZ - parsing using suffix tree.

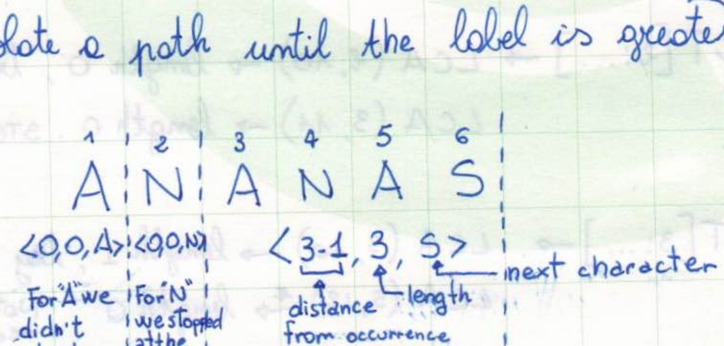
1) Build the suffix tree (notice that it is typical to add the "\$" at the end so no string is a prefix of all the others)



A	N	A	N	A	S	\$
1	2	3	4	5	6	7

2) Precompute minimum leaves for each node (purple color)

3) Percolate a path until the label is greater or equal than the position

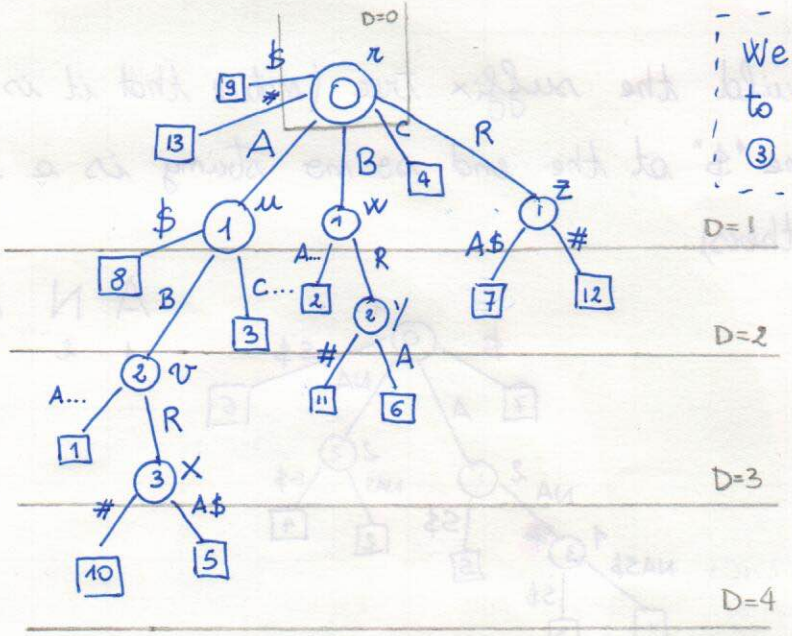


EXERCISE 2

Let T = ABACABRA and let P = ABR. How does the 1-mismatch algorithm works on T and P, given that it is provided for free a data structure that solves LCA in O(1) time. Only find a 1-mismatch.

① Concatenate T and P : ABACABRA$ABR#
1 2 3 4 5 6 7 8 9 10 11 12 13

② Build suffix tree



We label internal nodes to ease notation in step ③

These Ds indicate the depth of nodes and are needed for the second part of the exercise (see next page)

③ The algorithm compares P with the first suffix of T

- P ~ T[1: ...] → LCA (1, 10) has length 2, mismatch on (A, 3) and (R, 12) ✓
- P ~ T[2: ...] → LCA (2, 10) → length 0, try next character
  LCA (3, 11) → length 0 . STOP because 2 mismatches
- P ~ T[3: ...] → LCA (3, 10) → length 1, try next character
  LCA (5, 12) $\xrightarrow{x}$ length 0

  Notice that since the LCP is 1 it means that the second character is a mismatch, so we jump to the third character
- P ~ T[4: ...] → LCA (4, 10) $\xrightarrow{x}$ length 0, try next character
  LCA (5, 11) $\xrightarrow{x}$ length 0 STOP because 2 mismatches
- P ~ T[5: ...] → LCA (5, 10) $\xrightarrow{x}$ length 3, no mismatch ✓

With respect to this exercise provide a data structure that allows LCA in constant time and compute LCA(10, 6), block = 4

① Perform Euler tour (without characters "\$" and "#")

ET = r u 8 u v 1 v x 10 x 5 x v u 3 u r w 2 w y 11 y 6 y w r 4 r z 7 z 12 z r
D = 0 1 2 1 2 3 2 3 4 3 4 3 2 1 2 1 0 1 2 1 2 3 2 3 2 1 0 1 0 1 2 1 2 1 0

② Compute depth of each node

③ We recall that we have 3 pieces:
1) D' blocked D with sparsification (power of 2)
2) prefix-suffix minima of block of D
3) table of 0-1 s

In this kind of exercise we are asked to draw only the data structure(s) we need to complete the exercise

④ Partition D into blocks of size 4 and highlight 10 and 6 and compute minima

1	2	3	4	5	6	7	8	9
0121	2323	4343	2121	0121	2323	2101	0121	210
$\boxed{0}$	$\boxed{2}$	$\boxed{3}$	$\boxed{1}$	$\boxed{0}$	$\boxed{2}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$

⑤ We notice that we need only sparsification, since we need to compute LCA of blocks 3, 4, 5, 6.

3	1	0	2

Notice that we take the biggest power of 2 smaller than the length

$$\min \begin{cases} \min D'[3 ; 3+2-1] = 1 \\ \min D'[6-2^1+1 : 6] = 0 \end{cases} \rightarrow 0$$

⑥ Find the relative position of the minimum in the block, namely 1 and find the corresponding node in the Euler tour:

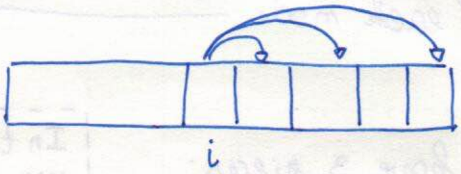
... ¦ v u 3 u ¦ r w 2 w ¦ ...

¦ 2 1 2 1 ¦ 0 1 2 1 ¦

position 1

Answer r.

Notice that the question was to provide a data structure that takes $n \log n$ the trivial answer would have been sparsification on the entire array.



EXERCISE 3

Provide the first 10 codewords of s.c code, where $s=1$, $c=3$.

Since $s+c=4=2^2$, the number of bits needed is 2

00	STOPPER
01	CONTINUERS
10	
11	

s	0	00
cs	1	01 00
	2	10 00
	3	11 00
ccs	4	01 01 00
	5	01 10 00
	6	01 11 00
	7	10 01 00
	8	10 10 00
	9	10 11 00
	10	11 01 00
		11 10 00
		11 11 00

we stop here since we are required 10 cws

A variation of this exercise is: provide the configuration of number 15.

\# of integers that can be written as $\boxed{S} \to 1$

\# " " " " " " " $\boxed{C}\ \boxed{S} \to 3 \cdot 1$

\# " " " " " " " $\boxed{C}\ \boxed{C}\ \boxed{S} \to 3 \cdot 3 \cdot 1$

\# " " " " " " " $\boxed{C}\ \boxed{C}\ \boxed{C}\ \boxed{S} \to 3 \cdot 3 \cdot 3 \cdot 1$

It comes without saying that since $15 > 1+3+9\underline{-1}$ the configuration is CCCS.

we start from 0

EXERCISE 4

Let $S = \underset{1}{2}\ \underset{2}{5}\ \underset{3}{8}\ \underset{4}{10}\ \underset{5}{11}\ \underset{6}{12}\ \underset{7}{13}$. Perform interpolative coding on one level of recursion.

1. $l = 1$, $r = 7$, $m = 4$, $low = 2$, $high = 13$

2. Start encoding 10, because $m = \frac{l+r}{2} = \frac{7+1}{2} = 4$ and $S_m = 10$

3. Compute range $= [5, 10]$

> Do you recall?
> range $= [low + m - l,\ high - r + m]$

4. Encode 5 (offset of S_m in the ~~gap~~ range) in $\lceil \log_2 \overset{\text{maximum gap}}{6} \rceil$ bits → 101

5. Recur on $\underset{1}{2}\ \underset{2}{5}\ \underset{3}{8}$ and on $\underset{5}{11}\ \underset{6}{12}\ \underset{7}{13}$

[1] $m = 3$ $l = 1$, $low = 2$, $r = 3$, $high = \underset{S_m - 1}{9}$	$l = 5$, $low = 11$, $r = 7$, $high = 13$, $m = 3$
[2] $m = 2 \Rightarrow S_m = 5$	Since $13 - 11 + 1 = 3 = m$ no bits are emitted for this part of S
[3] range $= [3, 8]$	
[4] Encode 5 - ~~[illegible]~~ = 2 in ~~[illegible]~~ 3 bits → 010	

EXERCISE 5

(a) Rice code, where $x = 5$, $K = 3$

① $q = \left\lfloor \frac{x-1}{2^k} \right\rfloor = 0$, $\quad r = 4$

② Concatenate q in unary with binary in K bits of r

$$\underbrace{1}_{\text{unary of } 0}\ \underbrace{100}_{\text{3 bits for 4}}$$

(b) Compute $x = 1$, $K = 3$ with Rice code

① $q = \left\lfloor \frac{x-1}{2^k} \right\rfloor = 0$, $\quad r = 0$

② Concatenate as before $\underbrace{1}_{\text{unary of } 0}\ \underbrace{000}_{\text{3 bits for } 0}$

(c) Decode 00100101000, given $K = 3$

① Split the sequence into block, knowing that we have a unary-binary sequence:

$$\underbrace{001}_{\text{unary}}\ \underbrace{001}_{\text{bin}} \;\Big|\; \underbrace{01}_{\text{unary}}\ \underbrace{000}_{\text{bin}}$$

$$q = 2,\ r = 1 \;\Big|\; q = 1,\ r = 0$$

$$\downarrow \qquad\qquad \downarrow$$

$$1 + 2^3 \cdot 2 + 1 \;\Big|\; 1 + 2^3 \cdot 1$$

$$\downarrow \qquad\qquad \downarrow$$

$$18 \qquad\qquad 9$$

Notice that the formula depends on the number $x-1$, which may be divisible by 2^k or not.

7th of December 2018

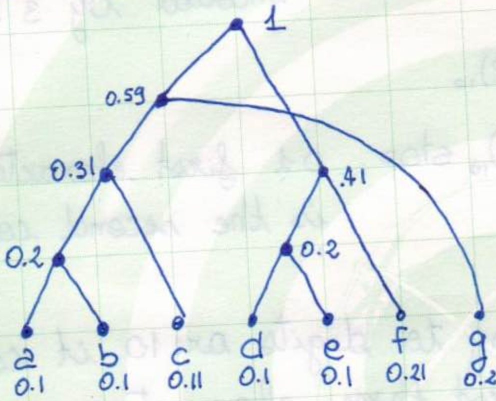
EXERCISES

EXERCISE 1

We are given a text, where the frequencies are the following

a	b	c	d	e	f	g
0.1	0.1	0.11	0.1	0.1	0.21	0.28

. Build a classical Huffman code



Then, the exercise asks to build the canonical Huffman.

1. Plot num array

ℓ	1	2	3	4
	0	2	3	2

2. Plot SYMB table

ℓ \ symb	0	1	2	3
1				
2	f	g		
3	c	d	e	
4	a	b		

③ Plot fc array

l	1	2	3	4
Fc	2	2	1	0

Then the exercise asks to decompress the first three symbols of following sequence

$$C = \underbrace{001}_{c}\underbrace{10}_{f}\underbrace{10}\,1011011100\ldots$$

$l = 1$, $(0)_2 < fc[1] = (2)_{10}$

$l = 2$, $(00)_2 < fc[2] = (2)_{10}$

$l = 3$ $(001)_2 = fc[3] = (1)_{10}$ stop. 001 first codeword, which corresponds to the first character encoded by 3 bits: c

$l = 1$, $(1)_2 < Fc[1] = (2)_{10}$

$l = 2$, $(10)_2 = Fc[2] = (2)_{10}$ stop. 01 first character encoded by 2 bits: f is the second codeword

Since the following to digits are 10 it comes without saying that the first three characters are c f f.

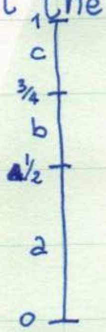
EXERCISE 2

Decode with arithmetic coding the first three characters of the following $c = 10011$, provided that $p(a) = \frac{1}{2}$ and $p(b) = p(c) = \frac{1}{4}$

① Diadic fraction : $1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} + 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{16} + 1 \cdot \frac{1}{32} = \frac{19}{32}$

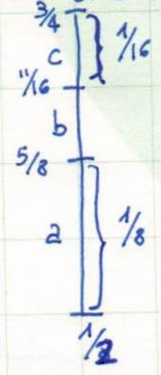
Always use fractions instead of floating points.
Always write letters sorted alphabetically

② Plot the intervals



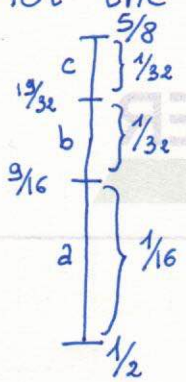
Since $\frac{1}{2} < \frac{19}{32} < \frac{3}{4}$ we split the second interval (b)

③ Plot the new intervals



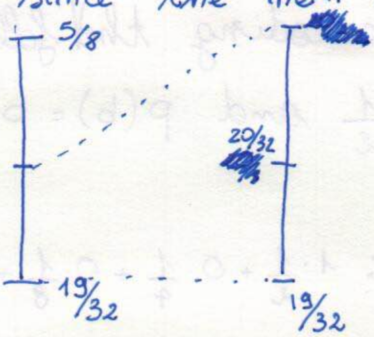
Since $\frac{19}{32} < \frac{5}{8}$ we split the first interval (a)

④ Plot the new intervals



Since $\frac{19}{32}$ is represented as a border of the interval we should continue splitting the interval $\left[\frac{9}{16}, \frac{19}{32}\right]$, but we are required 3 symbols and we obtained: bac

Observation: since the next intervals would have been



we can say that the sequence is bac$\bar{a}$

Exercise 3

Compute the perfect hash for $S = \{1, 5, 11, 3, 10, 8, 2, 4\}$ provided that the first level $m = 5$.

① Pick at random a function $h_{ab}(x) = ax + b \mod p$, say $h_{2,1}(x) = 2x + 1 \pmod 5$

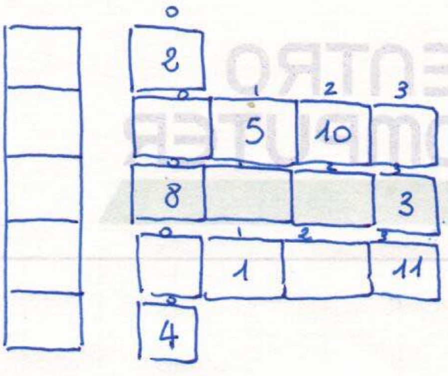
② Compute the number of collisions, suggestion: build graphically hashing with chaining

$m_0 = 1$	0	$\to$ 2
$m_1 = 2$	1	$\to$ 5, 10
$m_2 = 2$	2	$\to$ 3, 8
$m_3 = 2$	3	$\to$ 1, 11
$m_4 = 1$	4	$\to$ 4

Is this choice ok? $\sum_{i=0}^{4} (m_i)^2 \overset{?}{<} 2m \iff 1+4+4+4+1 \overset{?}{<} 16$

$\iff 14 < 16$ ok

③ Let us build $(m_i)^2$ buckets $\forall i = 0, \dots, 4$ and define 5 hash functions



$h_0 = x \pmod 1$

When $m_i = 1$ we may choose $a = 1$, $b = 1$ and $m = 1$

$h_1 = h_2 = h_3 = x \pmod 4$

$h_4 = x \pmod 1$

Perform search of key 7.

① $7 \cdot 2 + 1 \equiv 15 \equiv 0 \pmod 5 \rightarrow 1^\circ$ bucket

② $7 \equiv 0 \pmod 1 \rightarrow 1^\circ$ bucket

③ $\overset{0}{\boxed{2}}$ does not contain 7, hence 7 is not present.

EXERCISE 4

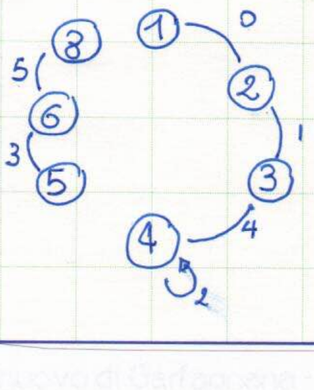
Given a set of strings $S = \{aa, ab, bb, bc, ca, db\}$, provided that the rank of the symbols is $a \to 1$, $b \to 2$, $c \to 3$, $d \to 4$ and $h_1(xy) = r(x) \cdot r(y) \pmod{m}$ and $h_2(xy) = r(x) + r(y) \pmod{m}$ build a MOPHF (Minimal Ordered Perfect Hash Function), on $m = 11$.

① Compute h_1, h_2 and h for all the Keys

	h_1	h_2	h
aa	1	2	0
ab	2	3	1
bb	4	4	2
bc	6	5	3
ca	3	4	4
db	8	6	5

② Build the graph

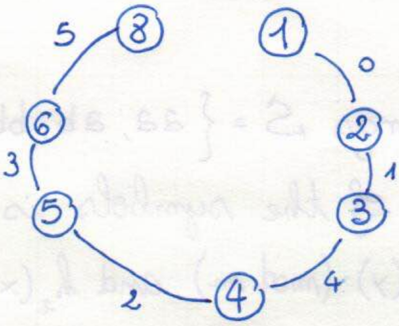
Only draw the nodes that are values of h_1 and h_2



Since the graph is not acyclic it is not possible to build a MOPHF.

The exercise asks to modify the graph to allow the creation of a MOPHF.

① Eliminate self loop, we choose



② Define by cases the function g

g	x
0	1
0	2
1	3
3	4
5	5
4	6
/	7
1	8

Notice that this method works only on keys in set S.
Let us take ac, $h_1(ac) = 3$, $h_2(ac) = 4 \Rightarrow g(h_1(ac)) = 1$, $g(h_2(ac)) = 3$, hence the algorithm would assign to "ac" rank 4, which is wrong.

Exercise 5

Assume we want to build a Bloom filter of a set of $n = 2^{16}$ keys building a binary array of $m = 2^{20}$ bits.

(a) Compute the optimal number of hash functions and the optimal error.

(b) Compute the size of the Bloom filter for $n = 2^{16}$ to guarantee $\varepsilon = 2^{-20}$ by using an optimal number of hash functions.

(a) $K_{opt} = \frac{m}{n} \cdot \ln 2 = 2^4 \cdot \ln 2$

$\varepsilon_{opt} = (0.6185)^{16}$

(b) Set $2^{-20} = (0.6185)^{\frac{m}{2^{16}}}$ and solve the equation