Greedy algorithms (contd.)

"What kind of question will be asked in the quiz".

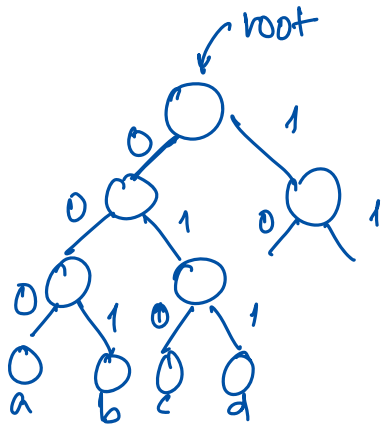
↓

"will_be_asked".

For each letter, we could use 6 bits to represent.

Qn: Can we have a better representation s.t average bit length all letters is smaller than 6.

Prefix tree:



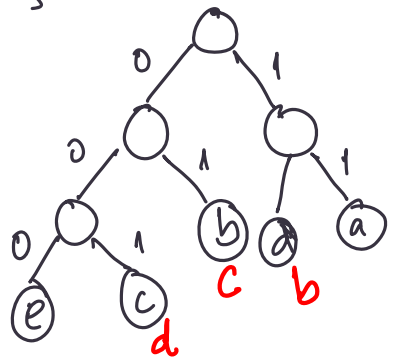
Enc(a) = 000
Enc(b) = 001
Enc(c) = 010
Enc(d) = 011.

← Leaves are labelled by letters

Encoding of a letter is given by the path from root to the leaf labelled by that letter.

$S = \{a, b, c, d, e\}$

11	$\gamma_1(a) = 11$	.32
10	$\gamma_1(b) = 01$	.25
01	$\gamma_1(c) = 001$	.20
001	$\gamma_1(d) = 10$	.18
000	$\gamma_1(e) = 000$	.05



avg depth = $\sum_{u \in S} \text{prob}(u) \cdot \text{bit length}(u)$

~~Exp~~

$= 0.32 \times 2 + 0.25 \times 2 + 0.20 \times 3 + 0.18 \times 2 + 0.05 \times 3$

$= 2.25.$ 2.23.

abcde 11100100 1000

Prefix property: For any two letters u and v, $Enc(u)$ is not a prefix of $Enc(v)$, and vice versa.

For any node, encoding is described by the path from root to that node. Any letter only appears at a leaf.

u' and v' are s.t $Enc(u')$ is a prefix of $Enc(v')$ $\quad u' \neq v'$

$\Rightarrow u'$ is an internal node on path from root to v'.

$\Rightarrow u'$ cannot be a letter as letters are on leaves.

Task: Construct optimal prefix trees given a document/text.

↑ wrt avg depth.

Shannon-Fano schemes:

$S \rightarrow$ Split S into 2 parts S_1 and S_2 s.t $prob(S_1) \approx prob(S_2)$

Bad ↳ This scheme won't always work.

Equi Partition can be hard to come up with.

Huffman Coding:

S

P: List of prob of appearances of letters in S in incr. order.

Idea: Smallest depth for most freq. characters

|||

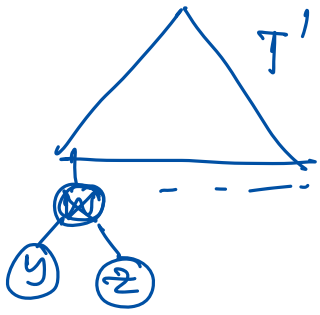
Largest depth for least freq. characters.

Algo (S, P)

- Pick two least freq. characters, say y and z.
- Remove y and z from S and P.
- Add a new letter w to S s.t

$$prob(w) = prob(y) + prob(z).$$

T =



.32 a
.25 b
.20 c
.18 d
.05 e

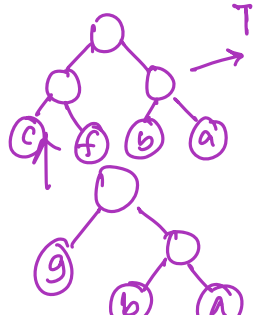
- Call updated lists S' and P'.
- Tree T' ← Algo (S', P').
- Obtain tree T by replacing leaf w by a node with children y and z.
- Return T.

$S = \{a, b, c, d, e\}$ $P = \{.32, .25, .20, .18, .05\}$

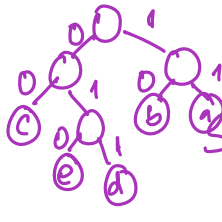
$S' = \{a, b, c, f\}$ $P' = \{0.32, .25, 0.20, 0.23\}$

$S'' = \{a, b, g\}$ $P'' = \{0.32, .25, 0.43\}$

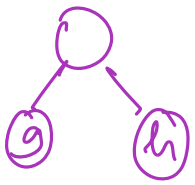
T' =



T =



T'' =

T''' = 

$S''' = \{h, g\}$ $\quad \downarrow P''' = \{0.57, 0.43\}$

$a = 11$, $b = 10$, $c = 00$, $d = 011$, $e = 010$.

Avg. depth $= 2 \times 0.32 + 2 \times 0.25 + 2 \times 0.20 + 3 \times 0.18 + 3 \times 0.05$

$= 0.64 + 0.5 + 0.4 + 0.54 + 0.15 = 2.23$.