IN ENGLISH, THE FREQUENCIES OF THE LETTERS FOLLOW THIS "LAW":

"E":	12.70%	"E":	11.39%
"T":	9.10%	"A":	11.76%
"A":	8.20%	"I":	11.28%
"O":	7.50%	"O":	9.83%
⋮		⋮	
"Q":	0.10%		
"Z":	0.07%	"Z":	0.49%

GIVEN THE NON-UNIFORMITY OF THE FREQUENCIES, ONE WOULD GUESS THAT USING THE SAME NUMBER OF BITS FOR EACH LETTER IS A BIT OF A WASTE.

PREFIX CODING

DEF: A PREFIX CODE FOR A SET OF LETTERS/SYMBOLS S IS A FUNCTION γ FROM S TO THE STRINGS OF BITS, SUCH THAT

$$\forall x, y \in S,\ x \neq y,\ \gamma(x) \text{ IS NOT A PREFIX OF } \gamma(y).$$

$S = \{'A', 'B', 'C'\}$

NOT A PREFIX CODE ('x='A' AND y='B')

'A' → 0
'B' → 01
'C' → 1

IF I GIVE YOU THE CODE '01', YOU DO NOT KNOW IF THE TEXT IS 'AC' OR 'B'.

A PREFIX CODE

'A' → 0
'B' → 10
'C' → 11

DECODING IS UNIQUE
'01110' → ACB
A C B

TO REPRESENT A SEQUENCE OF LETTERS $X = x_1 x_2 \cdots x_m$, WE USE THE STRING OF BITS

$$\Gamma = \gamma(x_1)\gamma(x_2)\cdots\gamma(x_m). \qquad ('ACB' \to 01110)$$

HOW TO RECONSTRUCT X FROM Γ?

- WE SCAN Γ L-TO-R;
- AS SOON AS WE REACH A PREFIX OF Γ EQUAL TO SOME $\gamma(x)$, WE OUTPUT x AND WE REMOVE THE PREFIX $\gamma(x)$ FROM Γ.

THIS DECODING WILL SUCCEED GIVEN THAT γ IS A PREFIX-CODE.

OPTIMAL PREFIX CODE

SUPPOSE THAT WE ONLY HAVE LETTERS A, B, C, D, E, AND THAT THEIR FREQUENCIES ARE

$$f_A = 0.32 \qquad f_B = 0.25 \qquad f_C = 0.20 \qquad f_D = 0.18 \qquad f_E = 0.05$$

IF WE USE THE PREFIX CODE γ_1:

$$\gamma_1(A) = 11,\ \gamma_1(B) = 01,\ \gamma_1(C) = 001,\ \gamma_1(D) = 10,\ \gamma_1(E) = 000$$

THE AVERAGE BIT-COST PER LETTER OF γ_1 (AND f) IS THEN:

$$2 \cdot 0.32 + 2 \cdot 0.25 + 3 \cdot 0.20 + 2 \cdot 0.18 + 3 \cdot 0.05 = 2.25.$$

IF WE SWAP THE ENCODING FOR C AND D, WE GET

$$\gamma_2(A) = 11 \qquad \gamma_2(B) = 01 \qquad \gamma_2(C) = 10 \qquad \gamma_2(D) = 001 \qquad \gamma_2(E) = 000$$

THE AVG BIT COST IS NOW

$$2 \cdot 0.32 + 2 \cdot 0.25 + 2 \cdot 0.20 + 3 \cdot 0.18 + 3 \cdot 0.05 = 2.23 < 2.25$$

ALGORITHMIC PROBLEM:
GIVEN AN ALPHABET S, AND A FREQUENCY TABLE f FOR THE LETTERS IN S, FIND A PREFIX-CODE γ THAT MINIMIZES

$$\text{AVERAGE BIT-LENGTH}(\gamma) = \sum_{x \in S} \left(f_x \cdot |\gamma(x)| \right).$$

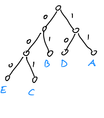
THE SOLUTION SPACE IS SUPER-EXPONENTIALLY LARGE AND "HARD" TO NAVIGATE.

NEXT WEEK THERE WILL BE 4 ALGORITHMS LECTURES:

TUE	12:00–16:00
WED	11:00–14:00
THU	11:00–14:00
FRI	11:00–13:00

PREFIX CODES AND BINARY TREES

LET T BE A BINARY TREE WHERE EACH LEAF HAS A UNIQUE LETTER ASSOCIATED TO.



IF WE TAKE A GENERIC LEAF, WITH LABEL $x \in S$, WE DEFINE AS THE ENCODING OF x THE SEQUENCE OF 0'S AND 1'S THAT WE ENCOUNTER AS WE TRAVERSE THE TREE FROM ITS ROOT TO THAT LEAF (WHEN WE GO LEFT, WE ADD A "0", WHEN WE GO RIGHT WE ADD "1").

$$\gamma(A) = 11,\ \gamma(B) = 01,\ \gamma(C) = 001,\ \gamma(D) = 10,\ \gamma(E) = 000$$

L: THE ENCODING OF S GENERATED BY TRAVERSING THE BINARY TREE T IS A PREFIX-CODE.

P: IF THE ENCODING OF $x \in S$ IS A PREFIX OF THE ENCODING OF $y \in S$, $y \neq x$, THEN THE PATH FROM THE ROOT TO x IS A PREFIX OF THE PATH FROM THE ROOT TO y. BUT x IS A LEAF, AND $y \neq x$. THUS, WE HAVE A CONTRADICTION. □

WE CAN, ALSO, CREATE BINARY TREES STARTING FROM PREFIX CODES.

$$\gamma_0(A) = 1$$
$$\gamma_0(B) = 011$$
$$\gamma_0(C) = 010$$
$$\gamma_0(D) = 001$$
$$\gamma_0(E) = 000$$

$\text{DEPTH}_T(A) = 1$
$\text{DEPTH}_T(B) = 3$

THM: THERE IS A BIJECTION FROM BINARY TREES WITH DISTINCT LABELS ON THE LEAVES, AND PREFIX CODES.

GIVEN A BINARY TREE T WITH LABELS ON THE LEAVES, WE WRITE $\text{DEPTH}_T(x)$ TO DENOTE THE DEPTH OF THE LEAF LABELED BY x IN T.

OUR ALGORITHMIC PROBLEM IS THEN: FIND A LABELED BINARY TREE T THAT MINIMIZES

$$\text{ABL}(T) = \sum_{x \in S} \left(f_x \cdot \text{DEPTH}_T(x) \right).$$