5112

Huffman Codes

Data Compression

Homework

Due October 1 at midnight

3 notes:
1. Not due Sept ...
2. Correction coming soon
3. Supplementary files

Why are you taking this class?

Collaboration is allowed - up to 3 per group
- submit your own work

Encoding Messages

"I got this"

"Help! How did we do homework last year?"

How would I send this in binary?

Encoding Messages

First idea: 26 English letters + space + punctuation

= 32 characters $\quad 32 = 2^5$

a 00000
b 00001
c 00010
d 00011
e 00100
.
.
.

Encoding Messages

Encoding Messages

"Well, Prince, so Genoa and Lucca are now just family estates of the Buonapartes. But I warn you, if you don't tell me that this means war, if you still try to defend the infamies and horrors perpetrated by that Antichrist—I really believe he is Antichrist—I will have nothing more to do with you and you are no longer my friend, no longer my 'faithful slave,' as you call yourself! But how do you do? I see I have frightened you—sit down and tell me all the news."

A	01	N	10
B	1000	O	111
C	1010	P	0110
D	100	Q	1101
E	0	R	010
F	0010	S	000
G	110	T	1
H	0000	U	001
I	00	V	0001
J	0111	W	011
K	101	X	1001
L	0100	Y	1011
M	11	Z	1100

Morse Code

A	01	N	10
B	1000	O	111
C	1010	P	0110
D	100	Q	1101
E	0	R	010
F	0010	S	000
G	110	T	1
H	0000	U	001
I	00	V	0001
J	0111	W	011
K	101	X	1001
L	0100	Y	1011
M	11	Z	1100

Short encoding → more frequent

Long encoding → less frequent

Average bits per letter : ≈ 2.53

1010

Morse code :

•, —, pause

ternary encoding

C

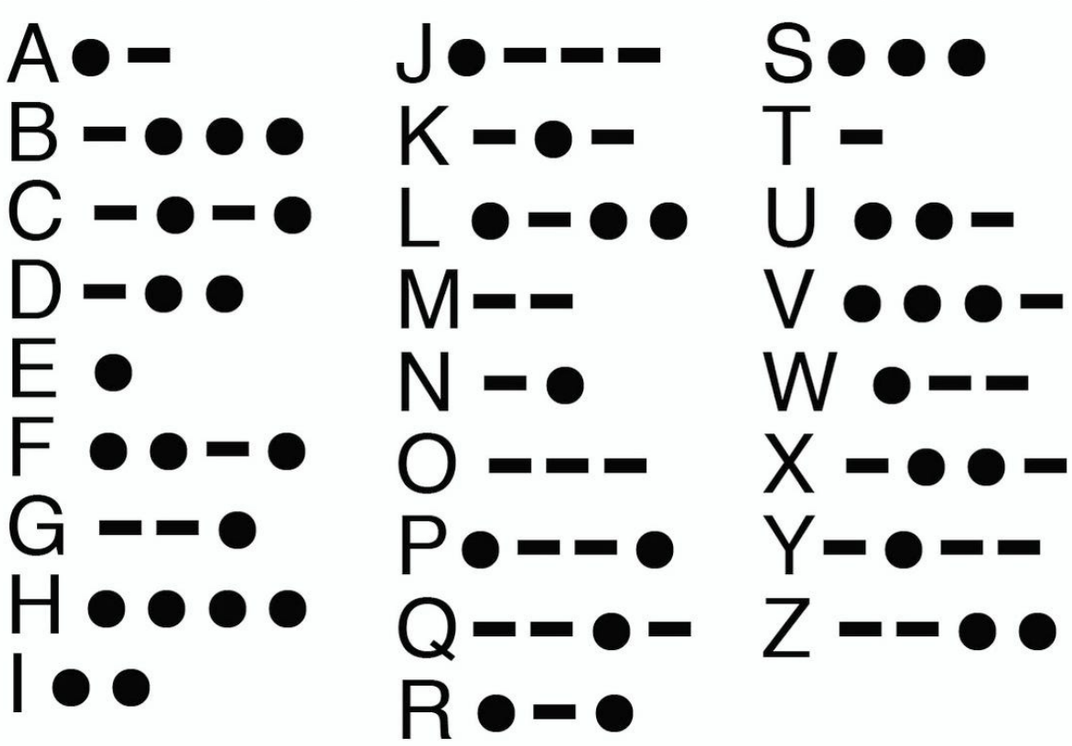
nn

etet

net

etn

?



Compression

Dates back to Shannon (1948)

Significant evolution

Many tools: gzip, bzip, deflate, brotli (2013)

Today: Huffman codes
- used in many compression algos, e.g. deflate

Prefix Codes

Set of characters $S \xrightarrow{\alpha}$ Sequences of $\{0,1\}$
such that if $x, y \in S$, $\alpha(x)$ is not a prefix of $\alpha(y)$

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

a	11
b	01
c	~~01~~1

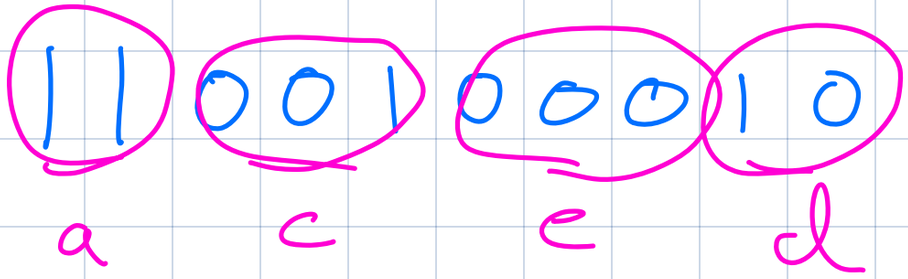
> b is a prefix of c!

Prefix Codes

Set of characters $S \xrightarrow{\alpha}$ Sequences of $\{0,1\}$
such that if $x, y \in S$, $\alpha(x)$ is not a prefix of $\alpha(y)$

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

a	11
b	01
c	001
d	10
e	000



I know this is "a" b/c of the prefix property

Prefix Codes

Frequency: Each letter $x \in S$ has some frequency f_x

fraction of characters in "the text" that are x

$$\sum_{x \in S} f_x = 1$$

Given a prefix code α and a message of length n,

$$\text{encoding length} = \sum_{x \in S} n f_x |\alpha(x)|$$

$$ABL(\alpha) = \text{average bits per letter} = \sum_{x \in S} f_x |\alpha(x)|$$

Prefix Codes

α_1

a ⟼ 11
b ⟼ 01
c ⟼ 001
d ⟼ 10
e ⟼ 000

$f_a = .32$
$f_b = .25$
$f_c = .20$
$f_d = .18$
$f_e = .05$

$$ABL(\alpha_1) = 2 \cdot .32 + 2 \cdot .25 + 3 \cdot .20 + 2 \cdot .18 + 3 \cdot .05 = 2.25 \text{ bits per letter}$$

α_2

a ⟼ 11
b ⟼ 10
c ⟼ 01
d ⟼ 001
e ⟼ 000

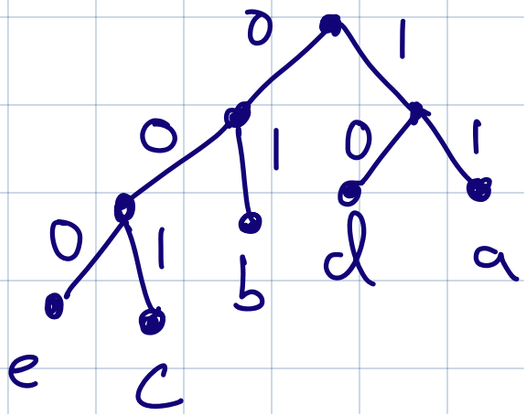
$$ABL(\alpha_2) = 2 \cdot .32 + 2 \cdot .25 + 2 \cdot .20 + 3 \cdot .18 + 3 \cdot .05 = 2.23$$

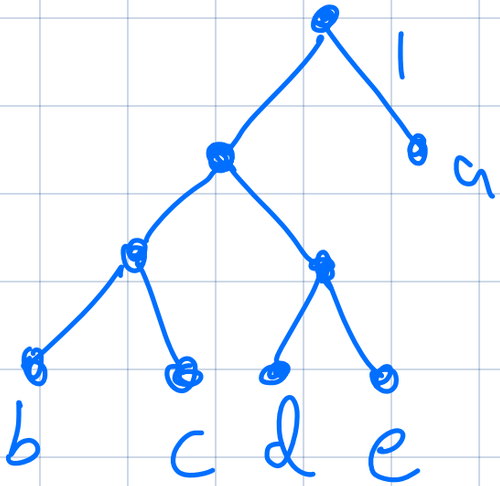
What is the optimal prefix code?

Prefix Codes and Binary Trees

α_1

$a \longmapsto 11$
$b \longmapsto 01$
$c \longmapsto 001$
$d \longmapsto 10$
$e \longmapsto 000$





Prefix code → binary tree w/o labels on interior nodes

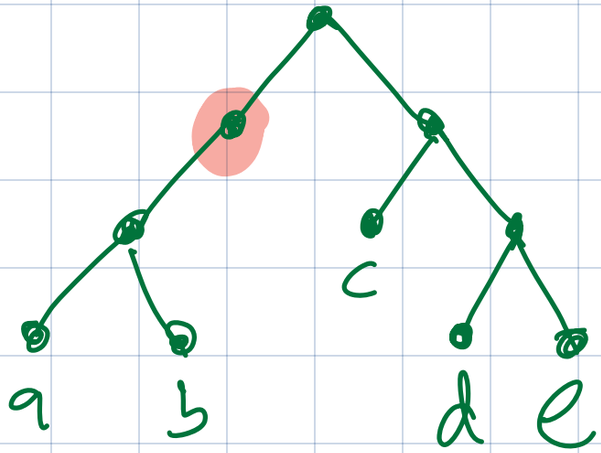
$a \longmapsto 1$
$b \longmapsto 000$
$c \longmapsto 001$
$d \longmapsto 010$
$e \longmapsto 011$

There is a 1-1 correspondence b/w prefix codes and binary trees with leaves labelled by S.

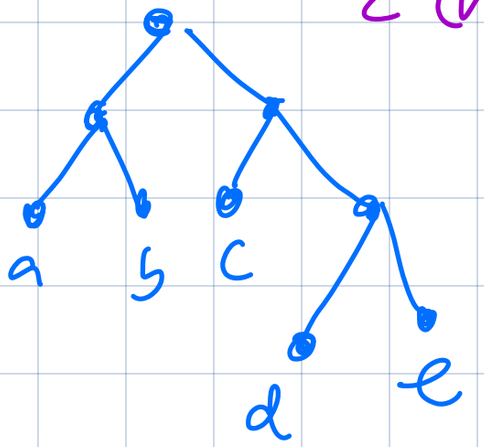
Optimal Prefix Codes

Observation: The binary tree corresponding to an optimal prefix code is full.

↑ every interior node has 2 children



remove the node

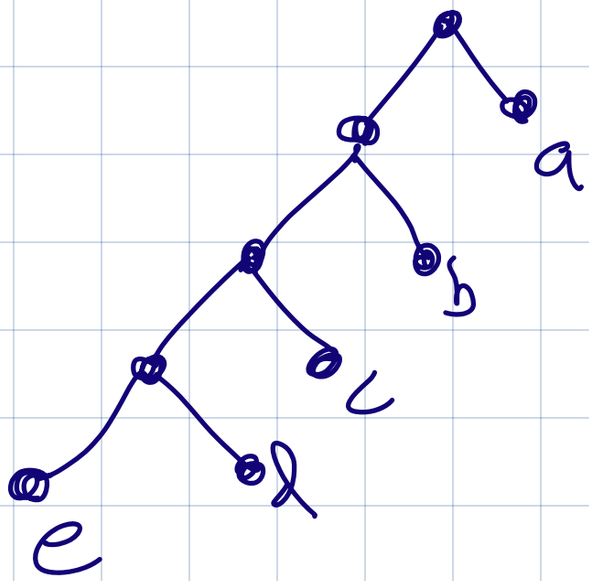


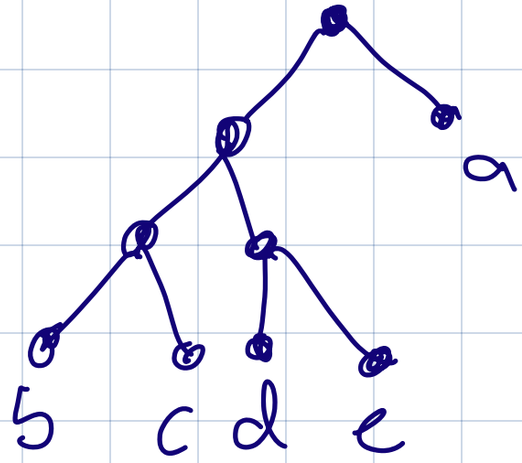
this is not optimal!

Optimal Prefix Codes

How would you label this tree?

$f_a = .32$
$f_b = .25$
$f_c = .20$
$f_d = .18$
$f_e = .05$





Huffman Codes

Start with the least frequent letter v.
If we had an optimal prefix code, v would have a sibling w.

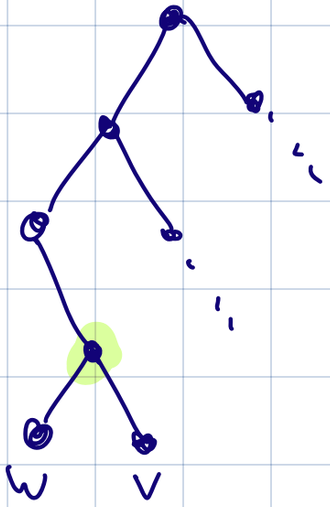
Consider the second least frequent letter y.

Know: $f_v < f_y < f_w$

$\Rightarrow$ w can't be lower than y.

$\Rightarrow$ w and y are on the same level

$\Rightarrow$ same # of bits in encoding



We could switch them w/o changing the ABL.

There exists an OPC where the two least frequent letters are siblings.

Huffman Codes

Start w/ two least frequent letters v, w.
Create a new letter $A = v \mid w$
$f_A = f_v + f_w$

Recurse on $S' = S \setminus \{v, w\} \cup \{A\}$

Repeat, repeat, repeat, ... until $|S| = 2$.

Huffman Codes

Huffman Codes