CS 223 Lec 16

Question: how do we find an optimal prefix code?

Answer: Huffman trees

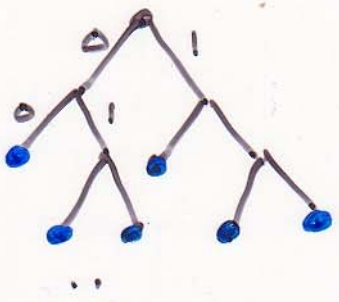
Let F be a text file, over an alphabet C.

It is easy to compute the frequency of each character $c \in C$.

$f[c]$ = # of occurrences of c in F

Let T be a tree representation of a prefix code.

T



If we use T to encode F

how long is the bit string produced?

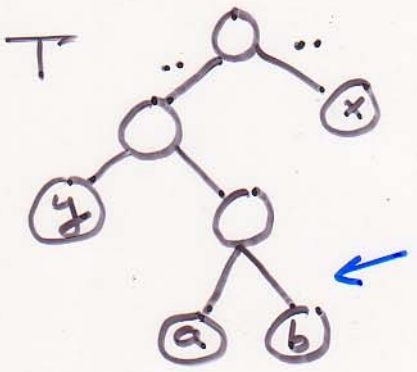
$B(T)$ = # of bits used

$$= \sum_{c \in C} f(c) \cdot \text{depth}_T(c)$$

goal: find a tree T that minimizes $B(T)$

Lemma Let $x, y \in C$ that have lowest frequencies. $\exists$ an optimal tree T where x and y are siblings.

Proof Let T be some optimal tree. Suppose x and y are not currently siblings.



← a and b are two nodes that are siblings and at maximum depth.

Observe we can assume

$f(a) \leq f(b)$ and $f(x) \leq f(y)$

Let T' be the tree obtained by exchanging a and x

T'

$$B(T) - B(T') = \sum_c f(c)\, d_T(c) - \sum_c f(c)\, d_{T'}(c)$$

$$= f(x)\, d_T(x) + f(a)\, d_T(a) - f(x)\, d_{T'}(x) - f(a)\, d_{T'}(a)$$

$$= f(x)\left(d_T(x) - d_{T'}(x)\right) + f(a)\left(d_T(a) - d_{T'}(a)\right)$$

~~$= (f(a) - f(x))(\quad)$~~

$$= f(x)\left(d_T(x) - d_T(a)\right) + f(a)\left(d_T(a) - d_T(x)\right)$$

$$= \underbrace{(f(x) - f(a))}_{\leq 0} \underbrace{(d_T(x) - d_T(a))}_{\leq 0}$$

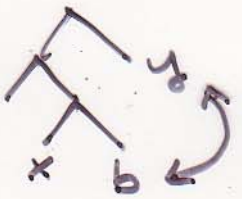
$$\geq 0$$

$$\Rightarrow \quad B(T) \geq B(T')$$

$B(T)$: optimal

So, T' must be optimal.

T'':



It is easy to show

$$B(T'') \leq B(T')$$

So T'' is also optimal.

Idea: merge the two least frequently occurring characters (nodes) to form a new node whose freq. is the sum of two children.

repeat

Example

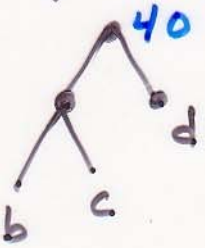
$C = \{a, b, c, d, e, f\}$

	f
a	100
b	10
c	5
d	25
e	30
f	60

{ a, b, c, d, e, f }

{ a, 

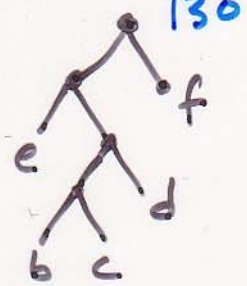
, d, e, f }

{ a, 

, e, f }

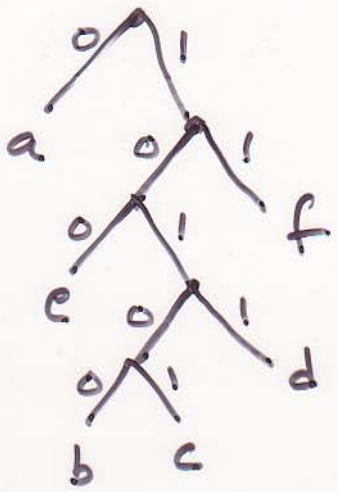
{ a, 

, f }

{ a, 

}

optimal ⇒ tree



freq	code	
100	a	0
10	b	1 0 1 0 0
5	c	1 0 1 0 1
25	d	1 0 1 1
30	e	1 0 0
60	f	1 1

$B(T)$ = # of bits used

$= 100 \cdot 1 + 10 \cdot 5 + 5 \cdot 5 + 25 \cdot 4 + 30 \cdot 3 + 60 \cdot 2$

$= 485$ bits.

Lemma Huffman's procedure leads to an optimal code.

Pf Form a tree by recursively merging the two least freq. nodes ... results in optimal tree

T be the optimal tree for original problem

T' the optimal tree for alphabet $C - \{x, y\} + \{z\}$

$\Rightarrow$ T is optimal provided T' is $\Rightarrow$ establish that Huffman's procedure is correct.