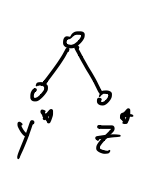
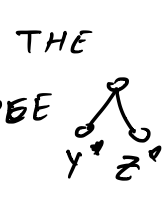
HUFFMAN CODING

T4: $\exists$ OPTIMAL PREFIX CODE, WITH TREE T^*, THAT ASSIGNS TWO LETTERS OF MINIMUM FREQUENCY TO TWO SIBLING LEAVES OF T^*.

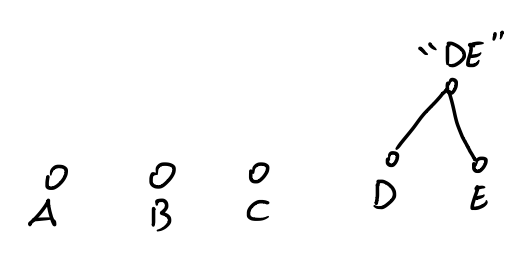
HUFFMAN(S, f): $(|S| \geq 2)$

IF $|S| == 2$:
- LET $S = \{y^*, z^*\}$
- ENCODE y^* WITH 0 AND z^* WITH 1 (OR, VICEVERSA).
- SET $T =$ 

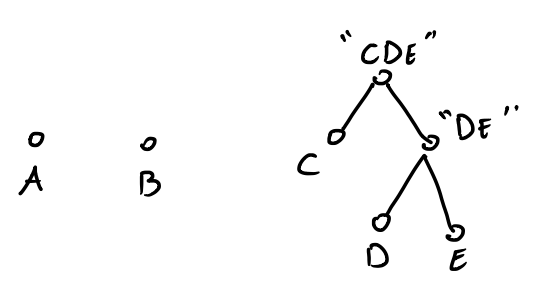
ELSE:
- LET y^* AND z^* BE TWO LETTERS OF SMALLEST FREQUENCIES
- LET $S' = S - \{y^*, z^*\} \cup \{\omega_{y^*, z^*}\}$
- LET f' BE SUCH THAT
  (I) $f'_{\omega_{y^*,z^*}} = f_{y^*} + f_{z^*}$, AND
  (II) $f'_x = f_x \quad \forall x \in S - \{y^*, z^*\}$
- RECURSIVELY BUILD AN OPTIMAL PREFIX-CODE FOR S', f'. LET T' BE TREE ASSOCIATED TO THIS CODE.
- CREATE A TREE T BY SUBSTITUTING THE LEAF LABELED ω_{y^*,z^*} IN T', WITH THE TREE  .

RETURN T.

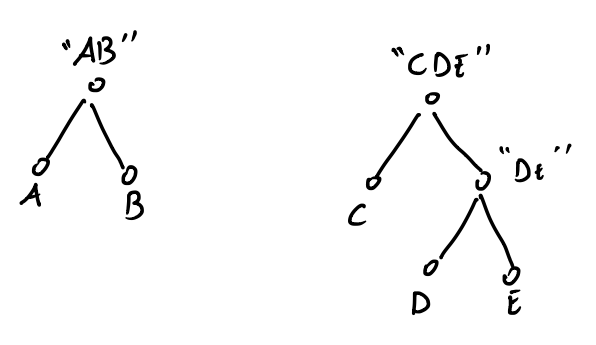
$f_A = 0.32 \quad f_B = 0.25 \quad f_C = 0.20 \quad f_D = 0.18 \quad f_E = 0.05$



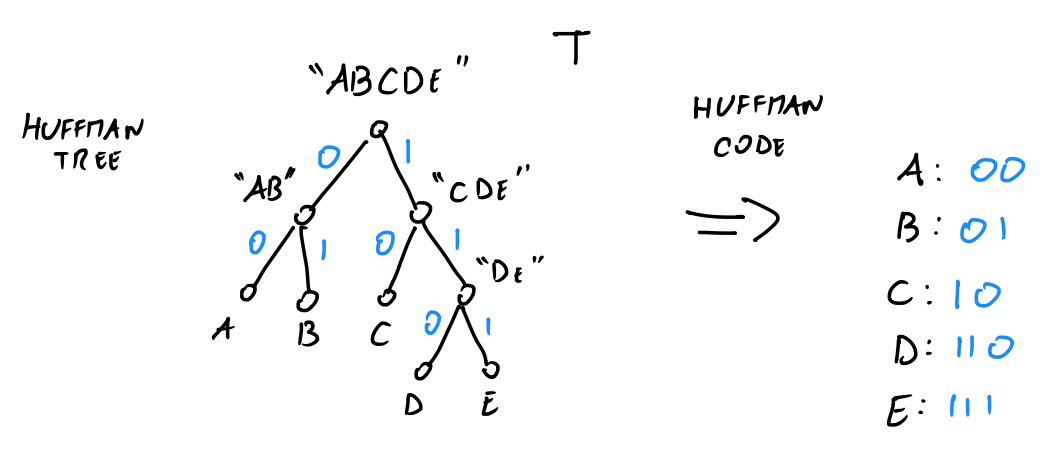
$f_A = 0.32 \quad f_B = 0.25 \quad f_C = 0.20 \quad f_{DE} = 0.23$



$f_A = 0.32 \quad f_B = 0.25 \quad f_{CDE} = 0.43$

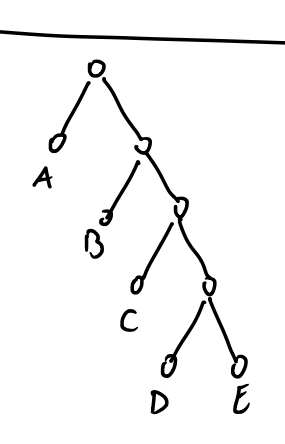


$f_{AB} = 0.57 \quad f_{CDE} = 0.43$



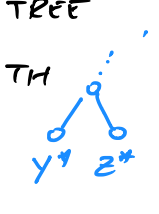
$ABL(T) = f_A \cdot 2 + f_B \cdot 2 + f_C \cdot 2 + f_D \cdot 3 + f_E \cdot 3$

$f_A = \frac{1}{2}$
$f_B = \frac{1}{4}$
$f_C = \frac{1}{8}$
$f_D = \frac{1}{16}$
$f_E = \frac{1}{16}$



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

IF $y^*, z^* \in S$ ARE TWO LETTERS OF SMALLEST FREQUENCIES, THE ALGORITHM (I) SUBSTITUTES THEM WITH A NEW LETTER ω (WITH $f_\omega = f_{y^*} + f_{z^*}$), (II) OBTAINS A NEW ALPHABET $S' = S - \{y^*, z^*\} \cup \{\omega\}$, AND (III) BUILDS THE OPTIMAL TREE T' FOR S'.

FINALLY, (IV) THE ALGORITHM TRANSFORMS T' IN A TREE T FOR S, SUBSTITUTING THE LEAF  OF T', WITH  .

L5: $ABL(T') = ABL(T) - f_\omega$

P: THE DEPTH OF EACH LETTER $x \in S - \{y^*, z^*\}$ IS SAME IN T AND T'.

ALSO THE DEPTH OF $x \in \{y^*, z^*\}$ IN T IS EQUAL TO THE DEPTH OF ω IN T' PLUS 1.

$$ABL(T) = \sum_{x \in S} \left(f_x \ \text{DEPTH}_T(x)\right)$$
$$= f_{y^*} \ \text{DEPTH}_T(y^*) + f_{z^*} \ \text{DEPTH}_T(z^*) + \sum_{x \in S - \{y^*, z^*\}} \left(f_x \ \text{DEPTH}_T(x)\right)$$
$$= f_{y^*} \left(1 + \text{DEPTH}_{T'}(\omega)\right) + f_{z^*} \left(1 + \text{DEPTH}_{T'}(\omega)\right) + \sum_{x \in S - \{y^*, z^*\}} \left(f_x \ \text{DEPTH}_{T'}(x)\right)$$
$$= (f_{y^*} + f_{z^*}) \left(1 + \text{DEPTH}_{T'}(\omega)\right) + \sum_{x \in S' - \{\omega\}} \left(f_x \ \text{DEPTH}_{T'}(x)\right)$$
$$= f_\omega \left(1 + \text{DEPTH}_{T'}(\omega)\right) + \sum_{x \in S' - \{\omega\}} \left(f_x \ \text{DEPTH}_{T'}(x)\right)$$
$$= f_\omega + f_\omega \ \text{DEPTH}_{T'}(\omega) + \sum_{x \in S' - \{\omega\}} \left(f_x \ \text{DEPTH}_{T'}(x)\right)$$
$$= f_\omega + \sum_{x \in S'} \left(f_x \ \text{DEPTH}_{T'}(x)\right) = f_\omega + ABL(T'). \ \square$$

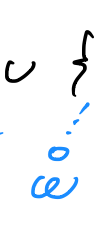
T: HUFFMAN'S ALGORITHM RETURNS A PREFIX CODE OF MINIMUM ABL (THAT IS, AN OPTIMAL PREFIX CODE).

P: WE PROVE THE CLAIM BY INDUCTION ON $|S|$.
IF $|S| = 2$, THE PREFIX CODE IS CLEARLY OPTIMAL.

OTHERWISE, LET T BE THE TREE RETURNED BY THE ALGORITHM.

BY CONTRADICTION, ASSUME THAT $\exists$ AN OPTIMAL TREE Z SUCH THAT $ABL(Z) < ABL(T)$.

WE CAN ALSO ASSUME (BY T4) THAT, IF y^*, z^* ARE TWO LETTERS OF SMALLEST FREQUENCIES, THE TWO NODES OF y^* AND z^* IN THE TREE Z ARE SIBLINGS.

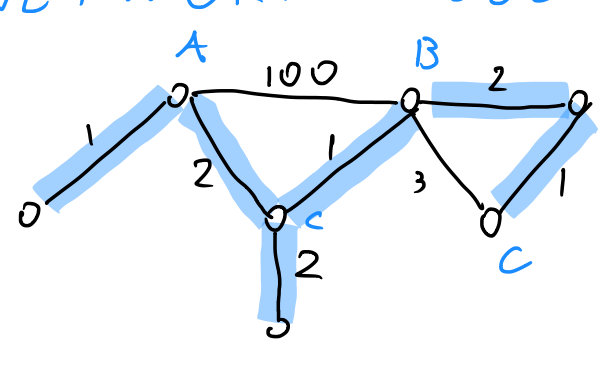
NOW, IF WE DELETE y^* AND z^* (THE NODES LABELLED BY y^* AND z^*) FROM Z AND WE LABEL ITS PARENT WITH ω, WE OBTAIN A TREE Z' THAT CORRESPONDS TO A PREFIX CODE FOR $S - \{y^*, z^*\} \cup \{\omega\}$. ($T$ IS OBTAINED BY SUBSTITUTING  WITH  IN T', AND Z CAN BE OBTAINED FROM Z' WITH THE SAME SUBSTITUTION).

BY L5, $ABL(T') = ABL(T) - f_\omega$; LIKEWISE, $ABL(Z') = ABL(Z) - f_\omega$.

BY CONTRADICTION WE ASSUMED THAT $ABL(Z) < ABL(T)$.
THEN,
$$ABL(Z') = ABL(Z) - f_\omega < ABL(T) - f_\omega = ABL(T').$$

RECALL THAT T' IS THE TREE RETURNED BY THE ALGORITHM ON AN ALPHABET OF SIZE $|S| - 1$. BY INDUCTION, THEN, T' HAS TO BE OPTIMAL — THEREFORE THERE CANNOT EXIST A TREE Z' S.T. $ABL(Z') < ABL(T')$. CONTRADICTION. $\square$

"NETWORK DESIGN PROBLEM"



$G(V, E)$ IS A WEIGHTED, CONNECTED, GRAPH WITH WEIGHTS $c: E \to \mathbb{R}_{\geq 0}$. (NON-NEGATIVE WEIGHTS)

WE HAVE A SET $V = \{v_1, \dots, v_n\}$ OF LOCATIONS. SOME PAIRS OF LOCATIONS (THOSE IN E) CAN BE DIRECTLY LINKED AT A COST. IN PARTICULAR, FOR EACH $\{v_i, v_j\} \in E$, DIRECTLY CONNECTING v_i TO v_j COSTS $c(v_i, v_j) > 0$ EUROS.

ASSUMING $G(V, E)$ IS CONNECTED, WHAT IS THE MINIMUM PRICE TO PAY TO (INDIRECTLY) CONNECT EACH PAIR OF LOCATIONS IN V?

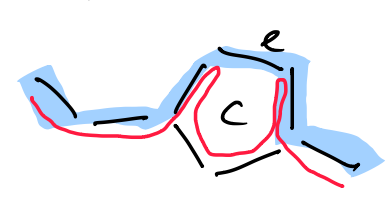
WE AIM TO FIND A SUBSET $T \subseteq E$ SO THAT:
- $G(V, T)$ IS CONNECTED, AND
- $COST(T) = \sum_{e \in T} c(e)$ IS MINIMUM.

L: LET T BE AN OPTIMAL SOLUTION TO THE NETWORK DESIGN PROBLEM. THEN, $G(V, T)$ IS A TREE.

P: BY DEF., $G(V, T)$ HAS TO BE CONNECTED.
WE SHOW THAT IT CANNOT CONTAIN CYCLES — THEN, IT MUST BE A TREE.

BY CONTR., SUPPOSE THAT $G(V, T)$ CONTAINS A CYCLE C. LET e BE ANY EDGE OF C.

$G(V, T - \{e\})$ IS A CONNECTED GRAPH (INDEED ANY PATH THAT TRAVERSES e CAN BE REROUTED THROUGH $C - \{e\}$)



THUS, $T - \{e\}$ IS A VALID (FEASIBLE) SOLUTION TO THE NETWORK DESIGN PROBLEM.
ITS COST IS

$$COST(T - \{e\}) = \sum_{e' \in T - \{e\}} c(e') = \left(\sum_{e' \in T} c(e')\right) - c(e) =$$
$$= COST(T) - c(e)$$
$$< COST(T),$$

SINCE $c(e') > 0 \quad \forall e' \in E$.

THUS, $T - \{e\}$ IS CHEAPER THAN T, AND T IS THEN NOT AN OPTIMAL SOLUTION. CONTRADICTION. $\square$

THE NETWORK DESIGN PROBLEM IS THEN ACTUALLY ASKING TO FIND A SUB TREE OF $G(V, E)$ OF MINIMUM COST THAT SPANS THE WHOLE GRAPH (THAT CONNECTS ANY TWO NODES OF V).

THIS PROBLEM IS FAMOUS UNDER THE NAME OF MINIMUM SPANNING TREE (MST).