HW5

1. **假定一个数据文件由 8 位字符组成，其中所有 256 个字符出现的频率大致相同：最高的频率也低于最低频率的 2 倍。证明：在此情况下，赫夫曼编码并不比 8 位固定长度编码更有效。**

   在Huffman编码中，最开始将两个频率最小的节点合并，由于最高的频率低于最低频率的2倍，故合并后的节点大于最高频率的节点，故只要还有字符存在，就会合并字符中频率最下的两个。因此一开始会形成128个子树，每个子树的孩子都是字符。由于$2f_{max} \leq 4f_{min}$，设128个子树的频率为f'，则有$f'_{max} \leq 2f'_{min}$，和之前一样，会形成64个子树，每个子树的孩子节点都为上一轮合并的节点。以此类推，最终得到高度为log(256)=8的Huffman树，且所有的256个字符的深度均为8，故Huffman编码并不比 8 位固定长度编码更有效。

2. **令 S 是一个有限集，S1,S2,...,Sk 是 S 的一个划分，这些集合都是非空且不相交的。定义结构 (S,I) 满足条件 I = {A : |A∩Si | ≤ 1,i = 1,..., k}。证明：(S,I) 是一个拟阵。也就是说，与划分中所有子集都最多有一个共同元素的集合 A 组成的集合构成了拟阵的独立集。**

   (1) S是一个有限集

   (2) 若集合$A \subseteq I$，且$B \subseteq A$，则有 $|A \bigcap S\ i\ | \leq 1, i = 1, \dots, k$，由于B是A的子集，故 $|B \bigcap S\ i\ | \leq 1, i = 1, \dots, k$，进而$B \subseteq I$，故是遗传的

   (3) 假设不满足交换性质，也就是说，$A \subseteq I$、$B \subseteq I$且$|A| < |B|$，有 $\forall\ x \in B - A, A \bigcup \{x\} \notin I$，即$|\{\{x\} \bigcup A\} \bigcap S\ i\ | > 2$，而$|A \bigcap S\ i\ | \leq 1$，故有$x \in S_i$。即当(S,I)不满足交换性质时，$\forall\ x \in B - A$，有$x \in S_i$，故为使$|B \bigcap S\ i\ | \leq 1$，$|B - A|$应小于等于1。

   当$|B| > |A| + 1$时，当A是B的真子集时，$|B - A|$最小，而此时$|B - A| > 1$，此时不满足$|B - A|$小于等于1（由上述结论）。

   当$|B| = |A| + 1$时，为使$|B - A|$小于等于1，A应为B的真子集，设$b \in B - A$，则$b \in S_i$。因为$A \subseteq I$、$B \subseteq I$，B最多和S_i有一个重复元素，而$b \in S_i$，故B中除去b以外的元素组成的集合与S_i没有交集，而$A = B - \{x\}$，即A与S_i没有交集，此时将b加入集合A中，新的集合A' 与S_i只有一个重复元素，即$|\{\{b\} \bigcup A\} \subseteq I$，这与假设不满足交换性质矛盾。

   由上可得，(S , I) 满足交换性质。

   综上，(S , I)是一个拟阵。

3. **A = a1,...,an 表示一个正整数集合。A 中的元素之和为 N。设计一个 O(n ·N) 的算法来确定是否存在一个 A 的子集 B，使得 $\sum_{a_i \in B} a_i = \sum_{a_i \in A-B} a_i$**

即找到子集B，使其和为N/2

定义$S[i,k]$表示在$\{a_1, \dots, a_i\}$中若能找到子集，使得其和为k，则为true，否则为false

$$S[i,k] = \begin{cases} false & \text{其他} \\ true & S[i-1, k-a_i] \text{为} true \end{cases}$$

采用自底向上计算， 先将S[1,a_i]、S[1,0]设为true，其余的S[1,k]设为false，再依次计算接下来的行数的S的值。

若存在一个 i，使得$S[i, N/2] = true$，则存在一个 A 的子集 B，使得 $\sum_{a_i \in B} a_i = \sum_{a_i \in A-B} a_i$

此算法对矩阵S中的每个元素遍历了一遍，故算法的时间复杂度为$O(n \cdot N/2)=O(n \cdot N)$