第四次作业

多媒体数据压缩

无损数据压缩

- 名词解释：UTF-8、UTF-16、UTF-32。
  - 三者均是计算机中对字符的编码方案
  - UTF-8 使用变长字节表示字符的编号。
    - 对于单字节的符号，字节的第一位设为 0，后面的 7 位为这个符号的 Unicode 码，因此对于英文字母，UTF-8 编码和 ASCII 码是相同的。
    - 对于 n 字节的符号 (n>1),第一个字节的前 n 位都设为 1，第 n+1 位设为 0，后面字节的前两位一律设为 10，剩下的位用来存储字符的 Unicode 码。
  - UTF-16 使用变长字节表示字符的编号。
    - 对于编号在 U+0000 到 U+FFFF 的字符，直接用两个字节表示。
    - 编号在 U+10000 到 U+10FFFF 之间的字符，用四个字节表示。
  - UTF-32 用 4 个字节存储字符的编号。
    - 还需要考虑计算机的端模式 (大端、小端)
- 某符号的 Unicode 数字编号为 0x4E2D，写出 UTF-8 编号后的 16 进制结果。
  - ‘1110 0100 1000 1000 1010 1101 ’
- 已知信源 X: {x1, x2, x3, x4, x5, x6, x7}，各信源符号的概率依次为 P(X): {0.2, 0.19, 0.18, 0.17, 0.15, 0.1, 0.01}。求霍夫曼编码，并计算编码效率。
  - 编码：

X	x1	x2	x3	x4	x5	x6	x7
P(x)	0.2	0.19	0.18	0.17	0.15	0.1	0.01
Code	10	11	000	001	010	0110	0111
Length	2	2	3	3	3	4	4

  - 码长：

$$2\times(0.2+0.19)+3\times(0.18+0.17+0.15)+4\times(0.1+0.01)=2.72$$

  - 最低码长：

$$-log_2^{0.2}-log_2^{0.19}-log_2^{0.18}-log_2^{0.17}-log_2^{0.15}-log_2^{0.1}-log_2^{0.01}=2.6087$$

  - 编码效率：

$$\frac{2.6087}{2.72}=95.91\%$$

- 对一个具有符号集 B= {b1, b2} = {0, 1}，设信源产生 2 个符号的概率分别为 P(b1)=0.2, P(b2)=0.8。对二进制数 1001 进行算术编码（结果用十进制数表示）。
  - 1001：（0.2048, 0.224）0.2048=0.001101000111
  - 0b001101000111=7256

- 对信息 00002033001110000600110111 1 进行行程（游程）编码。
  - 401212232031401620211041