Question 1

a) The maximum possible entropy for 4 symbols is $\log_2 4 = 2$ bits. This occurs if each symbol is equally likely, i.e., the file is not predictable.

b)

File 1

x_i	a	b	c	d
$p(x_i)$	0.5	0.25	0.2	0.05
$h(x_i) = -\log_2 p(x_i)$	1.000	2.000	2.322	4.322

$$\begin{aligned} H(X) = E[h(X)] = \sum h(x_i)\, p(x_i) &= 1.000(0.5) + 2.000(0.25) + 2.322(0.2) + 4.322(0.05) \\ &= 0.500 + 0.500 + 0.464 + 0.216 \\ &= 1.68 \text{ bits} \end{aligned}$$

c)

File 2

x_i	a	b	c	d
$p(x_i)$	0.1	0.8	0.05	0.05
$h(x_i) = -\log_2 p(x_i)$	3.322	0.322	4.322	4.322

$$\begin{aligned} H(X) = E[h(X)] = \sum h(x_i)\, p(x_i) &= 3.322(0.1) + 0.322(0.8) + 4.322(0.05) + 4.322(0.05) \\ &= 0.332 + 0.258 + 0.216 + 0.216 \\ &= 1.02 \text{ bits} \end{aligned}$$

d) The entropy for File 2 is lower since the symbols are more predictable, i.e., there is an 80% chance that the symbol will be "a".

The entropy for File 1 is higher since it is less predictable. However, there is still predictability since we know "d" is unlikely.

Both entropies are below the maximum of 2 bits due to the inherent predictability in both files.

Question 2

a) $C_2 = \{0, 1, 01, 11\}$ is *not* a prefix code: 0 is a prefix of 01 and, also, 1 is a prefix of 11.

$C_3 = \{0, 10, 11, 001\}$ is *not* a prefix code: 0 is a prefix of 001.

$C_5 = \{0, 10, 010, 111\}$ is *not* a prefix code: 0 is a prefix of 010.

The remaining codes C_1, C_4 and C_6 are prefix codes and, hence, they are uniquely decodable.

b) $\ell(C_2) = (1, 1, 2, 2) \Rightarrow K_2 = \dfrac{1}{2^1} + \dfrac{1}{2^1} + \dfrac{1}{2^2} + \dfrac{1}{2^2}$

$$= \frac{2}{2} + \frac{2}{4} = 1.5 > 1$$

$\ell(C_3) = (1, 2, 2, 3) \Rightarrow K_3 = \dfrac{1}{2^1} + \dfrac{1}{2^2} + \dfrac{1}{2^2} + \dfrac{1}{2^3}$

$$= \frac{1}{2} + \frac{2}{4} + \frac{1}{8} = 1.125 > 1$$

$\ell(C_5) = (1, 2, 3, 3) \Rightarrow K_5 = \dfrac{1}{2^1} + \dfrac{1}{2^2} + \dfrac{1}{2^3} + \dfrac{1}{2^3}$

$$= \frac{1}{2} + \frac{1}{4} + \frac{2}{8} = 1 \leq 1$$

$\Rightarrow C_2$ and C_3 are not uniquely decodable.

This is easy to show by a manual approach:

- C_2: ab = 01 = c
- C_3: da = 0010 = aab

c) C_5 is not uniquely decodable: ab = 010 = c.

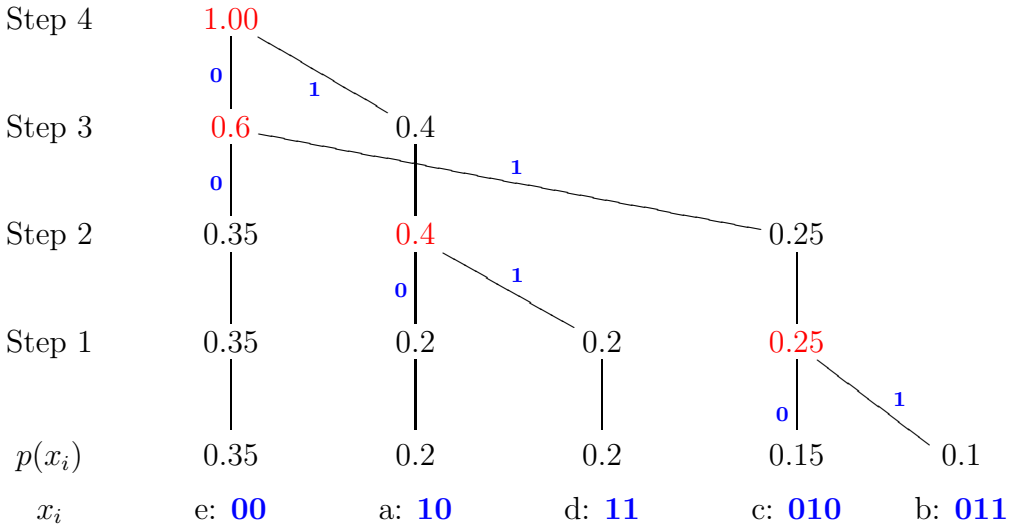
Question 3

a) First we must calculate the information contents: $h(x_i) = -\log_2 p(x_i)$. Note: the table has been reordered for the purpose of constructing the Huffman code.

x_i	e	a	d	c	b
$p(x_i)$	0.35	0.2	0.2	0.15	0.1
$h(x_i)$	1.515	2.322	2.322	2.737	3.322

$$H(X) = E[h(X)] = \sum h(x_i)\,p(x_i) = 1.515(0.35) + 2.322(0.2) + 2.322(0.2) + 2.737(0.15) + 3.322(0.1)$$
$$= 0.530 + 0.464 + 0.464 + 0.411 + 0.332 = 2.201 \text{ bits.}$$

b)



c)

x_i	e	a	d	c	b
$p(x_i)$	0.35	0.2	0.2	0.15	0.1
$h(x_i)$	1.515	2.322	2.322	2.737	3.322
$c(x_i)$	00	10	11	010	011
$\ell(x_i)$	2	2	2	3	3

$$E(L) = E[\ell(X)] = \sum \ell(x_i)\,p(x_i) = 2(0.35) + 2(0.2) + 2(0.2) + 3(0.15) + 3(0.1)$$
$$= 0.70 + 0.40 + 0.40 + 0.45 + 0.30 = 2.25 \text{ bits.}$$

d)
$$e = \frac{H(X)}{E(L)} = \frac{2.201}{2.25} = 0.978.$$

$\Rightarrow$ This Huffman code is 97.8% efficient.