8. Trees :: Applications :: Huffman Encoding

Suppose we wish to transmit, across a communication medium, an English text file (a book) which contains only lowercase letters a–z. In the ASCII character set, characters are **encoded** using 8-bits:

a = 01100001 (97_{10})
b = 01100010 (98_{10})
c = 01100011 (99_{10})
...
z = 01111010 (122_{10})

However, in English certain letters occur in words and sentences with greater frequency than other letters. For example, in the last sentence:

> however in english certain letters occur in words and sentences with greater frequency than other letters

the count for each letter is:

a = 4	f = 1	k = 0	p = 0	u = 2	z = 0
b = 0	g = 2	l = 3	q = 1	v = 1	
c = 5	h = 5	m = 0	r = 10	w = 3	
d = 2	i = 5	n = 9	s = 6	x = 0	
e = 16	j = 0	o = 4	t = 10	y = 1	

If we encoded and transmitted this 90-letter sentence using 8-bits per letter it would require us to transfer 720 bits.

8. Trees :: Applications :: Huffman Encoding (continued)

But, suppose we used a different encoding scheme for each letter, i.e., one where the number of bits per letter varies (a **variable-length encoding**). To encode our sentence using fewest number of bits as possible, it would be wise to encode the most frequently-occurring letters using a fewer number of bits and the less frequently-occurring letters using more bits.

I analyzed several classic English books (Dorian Gray by Wilde, Fall of the House of Usher by Poe, The House of Seven Gables by Hawthorne, The Jungle Book by Kipling, The Jungle by Sinclair, Metamorphosis by Kafka, Moby Dick by Melville, Portrait of the Artist as a Young Man by Joyce, Pygmalion by Shaw, The Adventures of Sherlock Holmes by Conan Doyle, A Tale of Two Cities by Dickens, The Time Machine by Wells, Tom Sawyer by Twain, and Walden by Thoreau) and determined the following letter frequencies (expressed as percentages):

a = 8.12	f = 2.23	k = 0.86	p = 1.66	u = 2.82	z = 0.07
b = 1.57	g = 2.12	l = 4.19	q = 0.10	v = 0.90	
c = 2.31	h = 6.68	m = 2.48	r = 5.64	w = 2.46	
d = 4.44	i = 6.81	n = 6.86	s = 6.29	x = 0.12	
e = 12.36	j = 0.14	o = 7.60	t = 9.18	y = 1.99	

As you were probably aware, the most frequently used letter in English is e followed by t and a, and the least frequently used letters are x, q and z.

Suppose then that we encode the letter e as 0, the letter t as 1, the letter a as 01, and so on, using longer bit patterns such as 01011101 for q and 01011110 for z.

We will continue Huffman Encoding in Trees : Section 9.