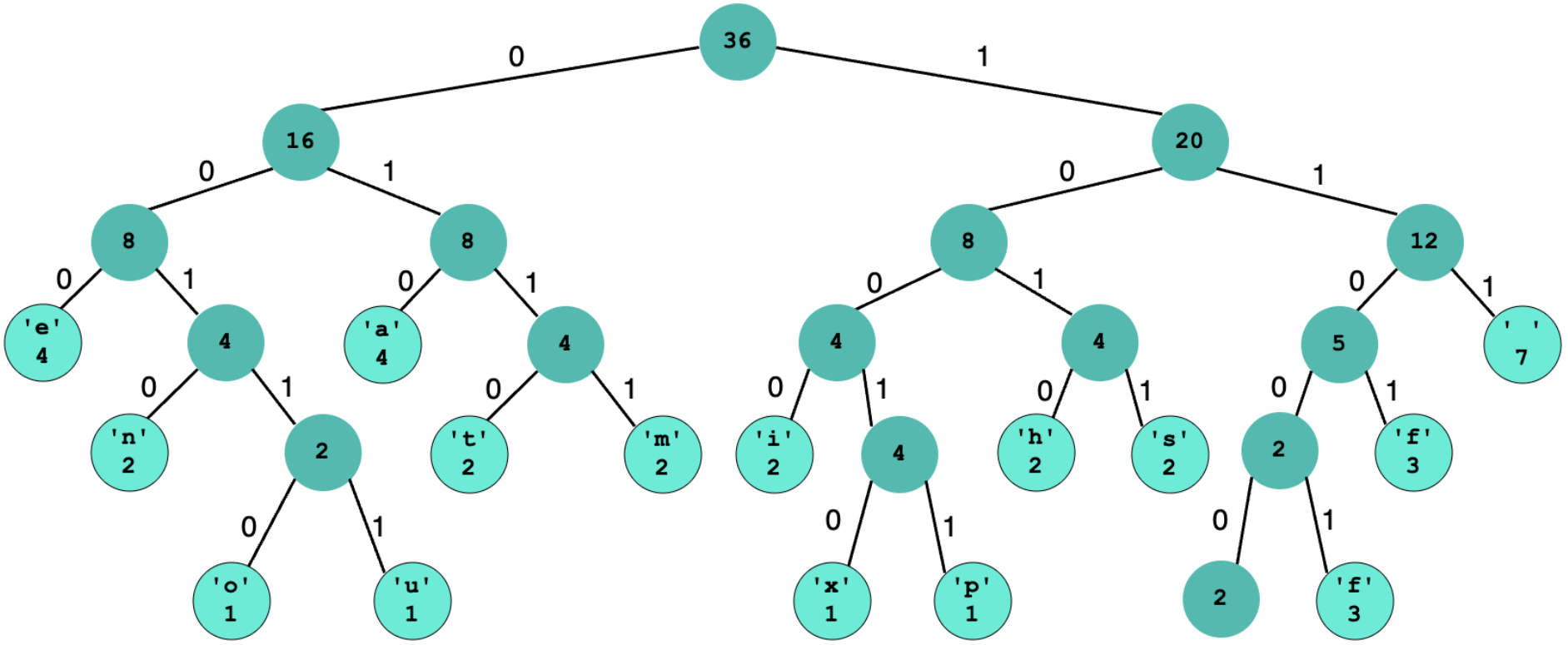
Week 25 Extension Question: Huffman Trees

A Huffman code is a type of lossless compression. It encodes characters as binary sequences. Each character sequence represents a path on the coded binary tree which leads to a leaf node character: 0 represents following the left child and 1 following the right child.
E.g. the character "u" is encoded as `00111` in the Huffman Tree below:



(a) Why do you think characters like ‘e’, ‘a’ and ‘[space]’ are at a lesser depth than other characters?
Answer: they appear more frequently in bodies of text
(b) Why do we not have to put any special character or break between the sequences for each character?
Answer: prefix-free property - no sequence can be a prefix of another so every time we reach a leaf, we simply record the character and start again from the root
(c) Decode the following sequence of bits: `011010101000101111101101100000000011110001011111001000001001 10`
Answer: ‘this tree is neat’