

Huffman Encoding

An Overview

The standard ASCII character encoding uses the same amount of space (one *byte* or eight *bits*, where each bit is either a 0 or a 1) to store each character. Common characters don't get any special treatment; they require the same 8 bits that are used for much rarer characters such as 'ü' or '¥'. A file of 1000 characters encoded using the ASCII scheme will take 1000 bytes (8000 bits); no more, no less, whether it be a file of 1000 spaces to a file containing 4 instances each of 250 different characters. A *fixed-length* encoding like ASCII is convenient because the boundaries between characters are easily determined and the pattern used for each character is completely fixed (i.e. 'a' is always exactly 97).

In practice, it is not the case that all 256 characters in the ASCII set occur with equal frequency. In an English text document, it might be the case that only 90 or so distinct characters are used at all (meaning 166 characters in the ASCII never even appear) and within those 90 there are likely to be significant differences in the character counts. The Huffman encoding scheme takes advantage of the disparity between frequencies and uses less storage for the frequently occurring characters at the expense of having to use more storage for each of the more rare characters. Huffman is an example of a *variable-length* encoding—some characters may only require 2 or 3 bits and other characters may require 7, 10, or 12 bits. The savings from not having to use a full 8 bits for the most common characters makes up for having to use more than 8 bits for the rare characters and the overall effect is that the file almost always requires less space.

ASCII Encoding

The example we're going to use throughout this handout is encoding the particular string **"happy hip hop"** (don't ask me what it means, I just made it up!). Using the standard ASCII encoding, this 13-character string requires $13 \times 8 = 104$ bits total. The table below shows the relevant subset of the standard ASCII table.

char	ASCII	bit pattern (binary)
h	104	01101000
a	97	01100001
p	112	01110000
y	121	01111001
i	105	01101001
o	111	01101111
space	32	00100000

The string **"happy hip hop"** would be encoded in ASCII as **104 97 112 112 121 32 104 105 112 32 104 111 112**. Although not easily readable by humans, it would be written as the following stream of bits (each byte is boxed to show the boundaries):

01101000	01100001	01110000	01110000	01111001	00100000	01101000
01101001	01110000	00100000	01101000	01101111	01110000	

To decode such a string (i.e. translate the binary encoding back to the original characters), we merely need to break the encoded stream of bits up into 8-bit bytes, and then convert each byte using the fixed ASCII encoding. The first 8 bits are **01101000**, which is the pattern for number 104, and position 104 in the ASCII set is assigned to lowercase **'h'**. A file encoded in ASCII does not require any additional information to be decoded since the mapping from binary to characters is the same for all files and computers.

A More Compact Encoding

The first thing you might notice about ASCII encoding is that using 8 bits per character can be excessively generous. Although it allows for the possibility of representing 256 different characters, we only have seven distinct characters in the phrase we're trying to encode, and thus could distinguish among these patterns with fewer bits. We could set up a special coding table just for this phrase using 3 bits for each character. Creating such an encoding is trivial: we create a list of the unique characters, and then go through and assign each a distinct encoded number from 0 to N-1. For example, here is one possible 3-bit encoding (of the 7! possible permutations):

char	number	bit pattern
h	0	000
a	1	001
p	2	010
y	3	011
i	4	100
o	5	101
space	6	110

Using this table, **"happy hip hop"** is encoded as **0 1 2 2 3 6 0 4 2 6 0 5 2**, or in binary:

000	001	010	010	011	110	000	100	010	110	000	101	010
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Using three bits per character, the encoded string requires 39 bits instead of the original 104 bits, compressing to 38% of its original size.

However, to decode this binary representation, one would need to know the special mapping used, since using **000** for **'h'** is not standard practice and in fact, in this scheme, each compressed string uses its own special-purpose mapping that is not necessarily like any other. Some sort of header or auxiliary file would have to be attached or included with the encoded representation that provides the mapping information. That header would take up some additional space that would cut into our compression savings. For a large enough file, though, the savings from trimming down the per-character cost would likely outweigh the expense of the additional table storage.

A Variable-Length Encoding

What if we drop the requirement that all characters take up the same number of bits? By using fewer bits to encode characters like **'p'**, **'h'**, and space that occur frequently and more to encode characters like **'y'** and **'o'** that occur less frequently, we may be able to compress even further. We'll later show how we generated the table below, but for now just take our word for it that it represents an optimal Huffman encoding for the string **"happy hip hop"**:

char	bit pattern
h	01
a	000
p	10
y	1111
i	001
o	1110
space	110

Each character has a unique bit pattern encoding, but not all characters use the same number of bits. The string **"happy hip hop"** encoded using the above variable-length code table is:

01	000	10	10	1111	110	01	001	10	110	01	1110	10
----	-----	----	----	------	-----	----	-----	----	-----	----	------	----

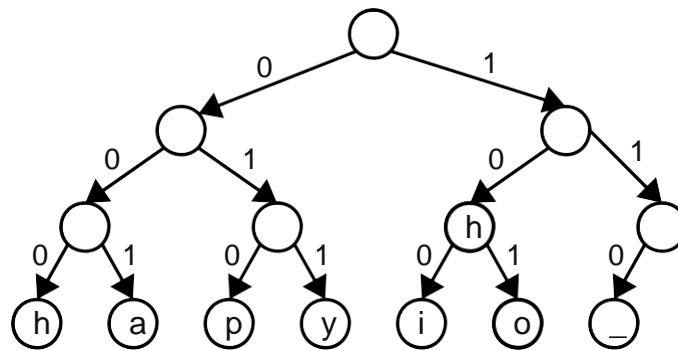
The encoded phrase requires a total of 34 bits, shaving a few more bits from the fixed-length version. What is tricky about a variable-length code is that we no longer can easily determine the boundaries between characters in the encoded stream of bits when decoding. I boxed every other character's bit pattern above to help you visualize the encoding, but without this aid, you might wonder how you will know whether the first character is encoded with the two bits **01** or the three bits **010** or perhaps just the first bit **0**? If you look at the encoding in the table above, you will see that only one of these options is possible. There is no character that encodes to the single bit **0** and no character that encodes to

the sequence **010** or **0100** or **01000** for that matter. There is, however, a character that encodes to **01** and that is '**h**'. One of the important features of the table produced by Huffman coding is the *prefix property*: no character's encoding is a prefix of any other (i.e. if '**h**' is encoded with **01** then no other character's encoding will start with **01** and no character is encoded to just **0**). With this guarantee, there is no ambiguity in determining where the character boundaries are. We start reading from the beginning, gathering bits in a sequence until we find a match. That indicates the end of a character and we move on to decoding the next character.

Like the special-purpose fixed-length encoding, a Huffman encoded file will need to provide a header with the information about the table used so we will be able to decode the file. Each file's table will be unique since it is explicitly constructed to be optimal for that file's contents.

Encoding Seen as a Tree

One way to visualize any particular encoding is to diagram it as a binary tree. Each character is stored at a leaf node. Any particular character encoding is obtained by tracing the path from the root to its node. Each left-going edge represents a 0, each right-going edge a 1. For example, this tree diagrams the compact fixed-length encoding we developed previously:



The path to **'h'** is just left right or **01**, the path to **'y'** is right-right-right-right or **1111**. Notice that the prefix property of the Huffman encoding is visually represented by the fact that characters only occupy leaf nodes (i.e. those nodes which are not a prefix of any further nodes).

The tree shown above for **"happy hip hop"** is, in fact, an optimal tree—there are no other tree encodings by character that use fewer than 34 bits to encode this string. There are other trees that use exactly 34 bits; for example you can simply swap any sibling nodes in the above tree and get a different but equally optimal encoding.

The Huffman tree doesn't appear as balanced as the fixed-length encoding tree. You've heard in our discussion on binary search trees that an unbalanced tree is bad thing. However, when a tree represents a character encoding, that lopsidedness is actually a good thing. The shorter paths represent those frequently occurring characters that are being encoded with fewer bits and the longer paths are used for more rare characters. Our plan is to shrink the total number of bits required by shortening the encoding for some characters at the expense of lengthening others. If all characters occurred with equal frequency, we would have a balanced tree where all paths were roughly equal. In such a situation we can't achieve much compression since there are no real repetitions or patterns to be exploited.

Decoding Using the Tree

A particularly compelling reason to diagram an encoding as a tree is the ease with which it supports decoding. Let's use the fixed-length tree to decode the stream of bits **011101010010011**. Start at the beginning of the bits and at the root of the tree. The first bit is **0**, so trace one step to the left, the next bit is **1**, so follow right from there, the following bit is **1**, so take another right. We have now landed at a leaf, which indicates that we have just completed reading the bit pattern for a single character. Looking at the leaf's label, we learn we just read a **'y'**. Now we pick up where we left off in the bits and start tracing again from the root. Tracing **101** leads us to **'i'**. Continuing through the remaining bits and we decode the string **"yippy"**.

The same path-following strategy works equally well on the Huffman tree. Decoding the stream of bits **111100110101111** will first trace four steps down to the left to hit the **'y'** leaf, then a left-left-right path to the **'i'** leaf and so on. Again the decoded string is **"yippy"**. Even though the encoded characters don't start and end at evenly spaced boundaries in the Huffman-encoded bits, we have no trouble determining where each character ends because we can easily detect when the path hits a leaf node in the encoding tree.

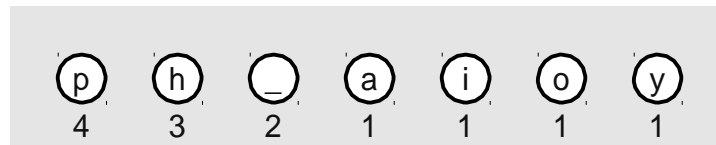
Generating an Optimal Tree

The pertinent question now: how is that special tree constructed? We need an algorithm for constructing the optimal tree giving a minimal per-character encoding for a particular file. The algorithm we will use here was invented by David Huffman in 1952.

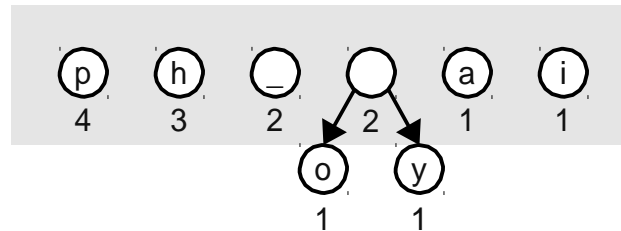
To begin generating the Huffman tree, each character gets a weight equal to the number of times it occurs in the file. For example, in the **"happy hip hop"** example, the character **'p'** has weight 4, **'h'** has weight 3, the space has weight 2, and the other characters have weight 1. Our first task is to calculate these weights, which we can do with a simple pass through the file to get the frequency counts. For each character, we create an unattached tree node containing the character value and its corresponding weight. You can think of each node as a tree with just one entry. The idea is to combine all these separate trees into an optimal tree by wiring them together from the bottom upwards. The general approach is as follows:

1. Create a collection of singleton trees, one for each character, with weight equal to the character frequency.
2. From the collection, pick out the two trees with the smallest weights and remove them. Combine them into a new tree whose root has a weight equal to the sum of the weights of the two trees and with the two trees as its left and right subtrees.
3. Add the new combined tree back into the collection.
4. Repeat steps 2 and 3 until there is only one tree left.
5. The remaining node is the root of the optimal encoding tree.

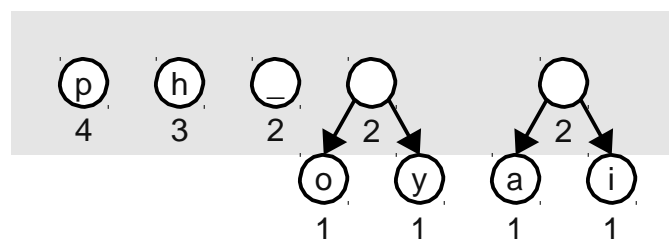
Sounds simple, doesn't it? Let's walk through building the optimal tree for our example string **"happy hip hop"**. We start with this collection of singletons, the weight of each node is labeled underneath:



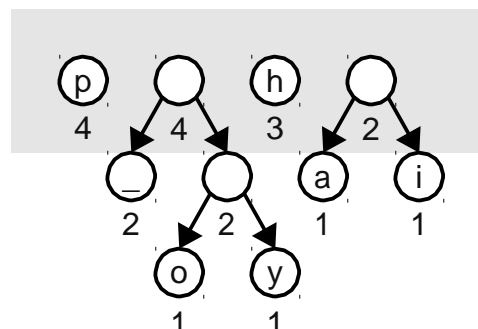
We start by choosing the two smallest nodes. There are four nodes with the minimal weight of one, it doesn't matter which two we pick. We choose 'o' and 'y' and combine them into a new tree whose root is the sum of the weights chosen. We replace those two nodes with the combined tree. The nodes remaining in the collection are shown in the light gray box at each stage.



Now we repeat that step, this time there is no choice for the minimal nodes, it must be 'a' and 'i'. We take those out and combine them into a weight 2 tree. Note how the collection of nodes shrinks by one each iteration (we remove two nodes and add a new one back in).

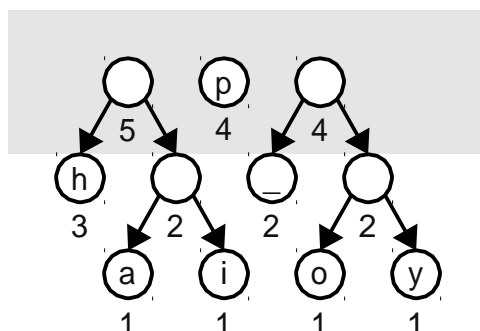


Again, we pull out the two smallest nodes and build a tree of weight 4:

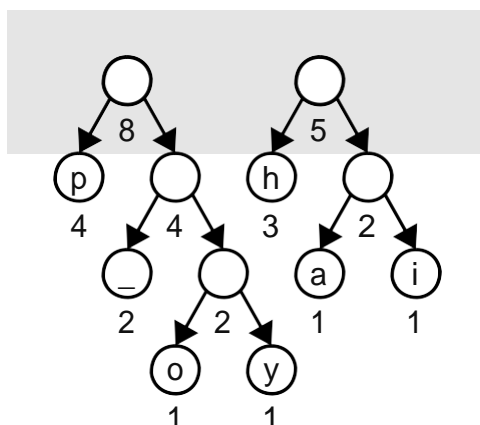


Note when we build a combined node, it doesn't represent a character like the leaf nodes do. These *interior nodes* are used along the paths that eventually lead to valid encodings, but the prefix itself does not encode a character.

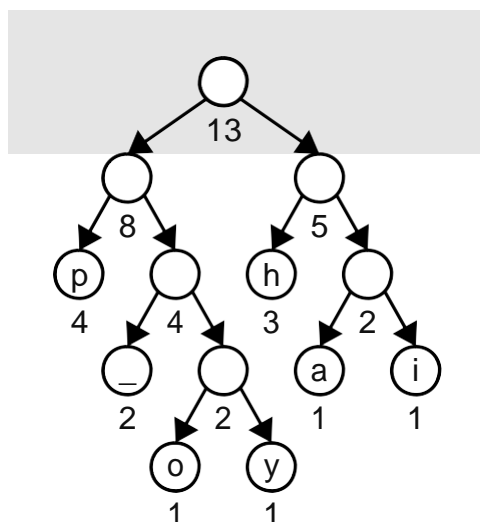
One more iteration combines the weight 3 and 2 trees into a combined tree of weight 5:



Combining the two 4s gets a tree of weight 8:



And finally, we combine the last two to get our final tree. The root node of the final tree will always have a weight equal to the number of characters in the input file.



Note that this tree is different from the tree on page 4, and has slightly different bit patterns, but both trees are optimal and the total number of bits required to encode **"happy hip hop"** is the same for either tree. When we have choices among equally weighted nodes (such as in the first step choosing among the four characters with weight 1) picking a different two will result in a different, but still optimal, encoding. Similarly when combining two subtrees, it is as equally valid to put one of the trees on the left and the other on the right as it is to reverse them.

Remember that it is essential that you use the same tree to do both encoding and decoding of your files. Since each Huffman tree creates a unique encoding of a particular file, you need to ensure that your decoding algorithm generates the exact same tree, so that you can get back the file.

Practical Considerations: The Encoding Table

There is one last issue we have not discussed yet. Suppose that I want to compress a message and send it to you. Using Huffman coding, I can convert the message (plus the pseudo-EOF) into a string of bits and send it to you. However, you cannot decompress the message, because you don't have the encoding tree that I used to send the message.

There are many ways to resolve this. We could agree on an encoding tree in advance, but this only works if we already know the distribution of the letters in advance. This might be true if we were always compressing normal English text, but in general is not possible to do.

A second option, and the option used in Assignment 6, is to prefix the bit sequence with a header containing enough information to reconstruct the Huffman encoding tree. There are many options you have for reading and writing the encoding table. You could store the table at the head of the file in a long, human-readable string format using the ASCII characters '0' and '1', one character entry per each line, like this:

```
h = 01
a = 000
p = 10
y = 1111
...
```

Reading this back in would allow you to recreate the tree path by path. You could have a line for every character in the ASCII set; characters that are unused would have an empty bit pattern. Or you could conserve space by only listing those characters that appear in the encoding. In such a case you must record a number that tells how many entries are in the table or put some sort of sentinel or marker at the end so you know when you have processed them all.

As an alternative to storing sequences of ASCII '0' and '1' characters for the bit patterns, you could store just the character frequency counts and rebuild the tree again from those counts in order to decompress. Again we might include the counts for all characters (including those that are zero) or optimize to only record those that are non-zero. Here is how we might encode the non-zero character counts for the "happy hip hop" string (the 7 at the front says there are 7 entries to follow—6 alphabetic characters, and the space character):

```
7 h3 a1 p4 y1 2 i1 o1
```

Greedy Algorithms

Huffman's algorithm is an example of a *greedy* algorithm. In general, greedy algorithms use small-grained, or local minimal/maximal choices in attempt to result in a global minimum/maximum. At each step, the algorithm makes the near choice that appears to lead toward the goal in the long-term. In Huffman's case, the insight for its strategy is that combining the two smallest nodes makes both of those character encodings one bit longer (because of the added parent node above them) but given these are the most rare characters, it is a better choice to assign them longer bit patterns than the more frequent characters. The Huffman strategy does, in fact, lead to an overall optimal character encoding.

Even when a greedy strategy may not result in the overall best result, it still can be used to approximate when the true optimal solution requires an exhaustive or expensive traversal. In a time or space-constrained situation, we might be willing to accept the quick and easy-to-determine greedy solution as an approximation.