CMPS 2200 Recitation 7

In this recitation, we'll look at huffman coding.

To make grading easier, please place all written solutions directly in `answers.md`, rather than scanning in handwritten work or editing this file.

All coding portions should go in `main.py` as usual.

Fixed-Length vs. Variable-Length Codes

In class we looked at the Huffman coding algorithm for data compresssion. Let's implement the algorithm and look at its empirical performance on a dataset of 5 text files, which are `alice29.txt`, `asyoulik.txt`, `f1.txt`, `fields.c`, and `grammar.lsp`.

a) We have implemented a means to compute character frequencies in a text file with the function `get_frequencies` in `main.py`. Compute cost for a fixed length encoding for each text file in function `fixed_length_cost(f)` by calling function `get_frequencies`.

b) Complete the implementation of Huffman coding in `make_huffman_tree`. Note that we manipulate binary trees in the priority queue using the object `TreeNode`. Moreover, once the tree is constructed, we must compute the actual encodings by traversing the Huffman tree that has been constructed. To do this, complete the implementation of `get_code`, which is a typical recursive binary tree traversal. That is, given a tree node, we recursively visit the left and right subtrees, appending a `0` or `1` to the encoding in each direction as appropriate. If we visit a leaf of the tree (which represents a character in the alphabet) we store the collected encoding for that character in `code`.

c) Now implement `huffman_cost` to compute the cost of a Huffman encoding for a character set with given frequencies.

d) Test your implementation of Huffman coding on the 5 given text files, and fill out a table of the encoding cost of each file for fixed-length and Huffman. Fill out a final column which gives the ratio of Huffman coding cost to fixed-length coding cost. Do you see a consistent trend? If so, what is it?

enter answer in `answers.md`

e) Suppose that we used Huffman coding on a document with alphabet Σ in which every character had the same frequency. What is the expected cost of a Huffman encoding for the document? Is it consistent across documents?

enter answer in `answers.md`