Project 3b CPE202

Huffman Decoding

For this project, you will implement a program to decode text that has been encoded using Huffman coding. This project assumes that the input file is in the format that was output from Assignment 3a.

For decoding, you will need to recreate the Huffman tree that was used to determine the Huffman codes used for encoding. The header in the input file contains the character frequency information needed for this. After reading in the frequency information, you will be able to use the same create_huff_tree() function that you wrote for the encoding portion of the assignment.

Once the Huffman tree is recreated, you will be able to traverse the tree using the encoded string of 1's and 0's from the input file. The decoding process starts by beginning at the root node of the Huffman tree. A '0' will direct the navigation to the left, while a '1' will direct the navigation to the right. When a leaf node is reached, the character stored in that node is the decoded character that is added to the decoded output. Navigation of the Huffman tree is then restarted at the root node, until all of the 1's and 0's have been consumed.

Note:

- Every function must come with a signature and a purpose statement.
- You must provide test cases for all functions.
- Use descriptive names for data structures and helper functions.

2 Recreating Huffman Tree, Decoding

Header Information

In the encoding portion of this assignment, we included a header on the first line of the output file that contains the character frequency information.

Implementation

- You should start with your huffman.py file from the first portion of the assignment, and add the functions necessary for the decode portion of the assignment.
- Write a function called **huffman_decode(encoded_file, decode_file)** (use that exact name) that reads an encoded text file, **encoded_file,** and writes the decoded text into an output text file, **decode_file,** using the Huffman Tree produced by using the header information. If the encoded_file does not exist, your program should raise the FileNotFoundError exception. If the specified output file already exists, its old contents will be overwritten.
- Before recreating the Huffman tree you will need to create a **freq_list** from the information stored in the header. For example, a header of “97 3 98 4 99 2” is associated with an original file of "aaabbbbcc" (3 a's, 4 b's, 2 c's).
- Write a function, **parse_header(header_string),** that takes a string input parameter (the first line of the input file) and returns a list of frequencies. The list of frequencies should be in the same format that cnt_freq() returned in the first part of this assignment (a list/arrary with 256 entries, indexed by the ASCII value of the characters). For the example above, **freq_list**[97:101] would be [3, 4, 2, 0].

- One you have re-created the list of freqs, pass it to your function **create_huff_tree(freq_list)** to recreate the Huffman Tree used for encoding**.**
- Once you have recreated the Huffman tree, you should have all the information you need to decode the encoded text and write it back out to the **decode_file**.

3 Tests

- Write sufficient tests using unittest to ensure full functionality and correctness of your program.
- When testing, always consider *edge conditions* like the following cases:
  - Single character files
  - Empty files
  - In your code, you will likely need to "special case" these types of input files

4 Some Notes

- The file huffman_tests_partb.py is in PolyLearn, under the Project 3b assignment. You can copy that file into your project directory, and use those as a starting point for your part b tests.
- When writing your own test files or using copy and paste from an editor, take into account that most text editors will add a newline character to the end of the file. If you end up having an additional newline character '\n' in your text file, that wasn't there before, then this '\n' character will also be added to the Huffman tree and will therefore appear in your generated string from the Huffman tree. In addition, an actual newline character is treated different, depending on your computer platform. Different operating systems have different codes for "newline". Windows uses '\r\n' = 0x0d 0x0a, Unix and Mac use '\n' = 0x0a. It is always useful to use a hex editor to verify why certain files that appear identical in a regular text editor are actually not identical.

  This can be mitigated when opening files for write by including the newline input parameter in the function call:

  ```
  open(out_file, 'w', newline='')
  ```

  This will result in only the 0x0a character being written for a newline, regardless of platform.

5 Submission

You must commit and push the following files, using the same Project 3 repository as we did for part a:

- **huffman.py**: The functions from the first portion of the assignment, and the newly specified and functions
  - **parse_header(header_string)** takes in input header string and returns Python list (array) of frequencies (256 entry list, indexed by ASCII value of character)
  - **huffman_decode(encoded_file, decode_file)** decodes encoded file, writes it to decode file
- **huffman_tests.py**, containing test cases for the functions specified by the assignment
- Any text files needed for your test cases