**Data structures and Algorithms Assignment 1 – Text Compression: Huffman Coding**

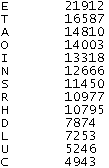
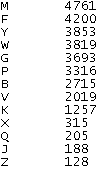
**Weighting 25%**

**Introduction**

Huffman coding is a scheme that assigns variable-length bit-codes (binary strings) to characters, such that the lengths of the codes depend on the frequencies of the characters in a typical message. As a result, encoded messages take less space (as compared to fixed-length encoding such as ASCII or UNICODE) since the letters that appear more frequently are assigned shorter codes. This is performed by first building a Huffman coding **Tree** based on a given set of frequencies. From the tree, bit-codes for each character are determined and then used to encode a message. The tree is also used to decode an encoded message as it provides a way to determine which bit sequences translate back to a character. (See the text passage at the end of the assignment for more details).

You are to implement a Huffman coding system using the frequency table given below (this is also in a text file on moodle, LetterCount.txt). Your final solution should be able to encode any message that uses the characters of the english alphabet or decode any bit sequence to the corresponding characters. There are three main parts to the system, *generating the huffman tree, encoding and decoding*.

**Frequency Table**

**Part 1: Generating the Huffman Tree (40 marks)**

Here you are required to generate **a Reference Based Binary tree** (Huffman tree) representing the bit-codes of each character in the alphabet. The characters that occur more frequently will have shorter bit codes than those occurring less frequently. Traversing the huffman tree from root to leaf should reveal the bit-code of the characters, where a left branch encodes a 0 (zero) and a right branch a 1. In this way, the leaves of the tree will represent the characters in our alphabet and the paths from the root to the leaves will reveal the binary codes for the characters. See the tutorial at the end of the document for an example on how to create a Huffman tree from a frequency table. You can use Java classes you have seen as part of your studies. For example you can use the TreeNode class to represent a node in your HuffmanTree but you will have to extend it to capture the symbol and its frequency. You could also use an ADT List of TreeNodes to represent the frequency table and grow the tree from here.

**Part 2: Encode (20 marks)**

This part of the assignment will involve generating the binary code for a message from the alphabet. This will be obtained by traversing the Huffman tree from the root to the leaf for each character in the message. There are a number of approaches that could be taken to this part. One might be to find the leaf node in the tree for a particular character and then back track from there to the root recording zero’s and one’s depending on whether you encounter a left or right branch. You would also need to reverse the code at the end as you went backwards up the tree. Display also the compression ratio for the encoded message by assuming that the fixed length representation for each character is 7 bits (the ASCII standard).

**Part 3: Decode (20 marks)**

This part of the assignment involves taking a binary string (the bit code) and regenerating the character message from the Huffman tree. To do this you will need to use the binary code to guide you down the right branches (remember, 0 for a left branch and 1 for a right branch) until you encounter a leaf and therefor e a character in the message.

**Part 4: Program Interface (20 marks)**

The interface for this program should be quite straightforward and resemble something like the following (with a bit more colour and dazzle ☺):

Text Area for output message after decoding or bitcode after encoding

Text Area for input message for encoding or bitcode for decoding

Decode button

Encode button

**You will receive marks for:**

1. All work attempted and submitted.
2. The quality of your solution design and your report describing this in detail.
3. The quality of your code and in-program documentation (comments).
4. You will receive marks for a bug-free, fully tested, application.
5. The quality of the test data.
6. The quality and professionalism of the design of the user graphical user interface. Ideally, this should be user friendly, realistic and have a well-designed Java look and feel while not permitting the entry of erroneous data.
7. Innovation and creativity in your deign and solution.

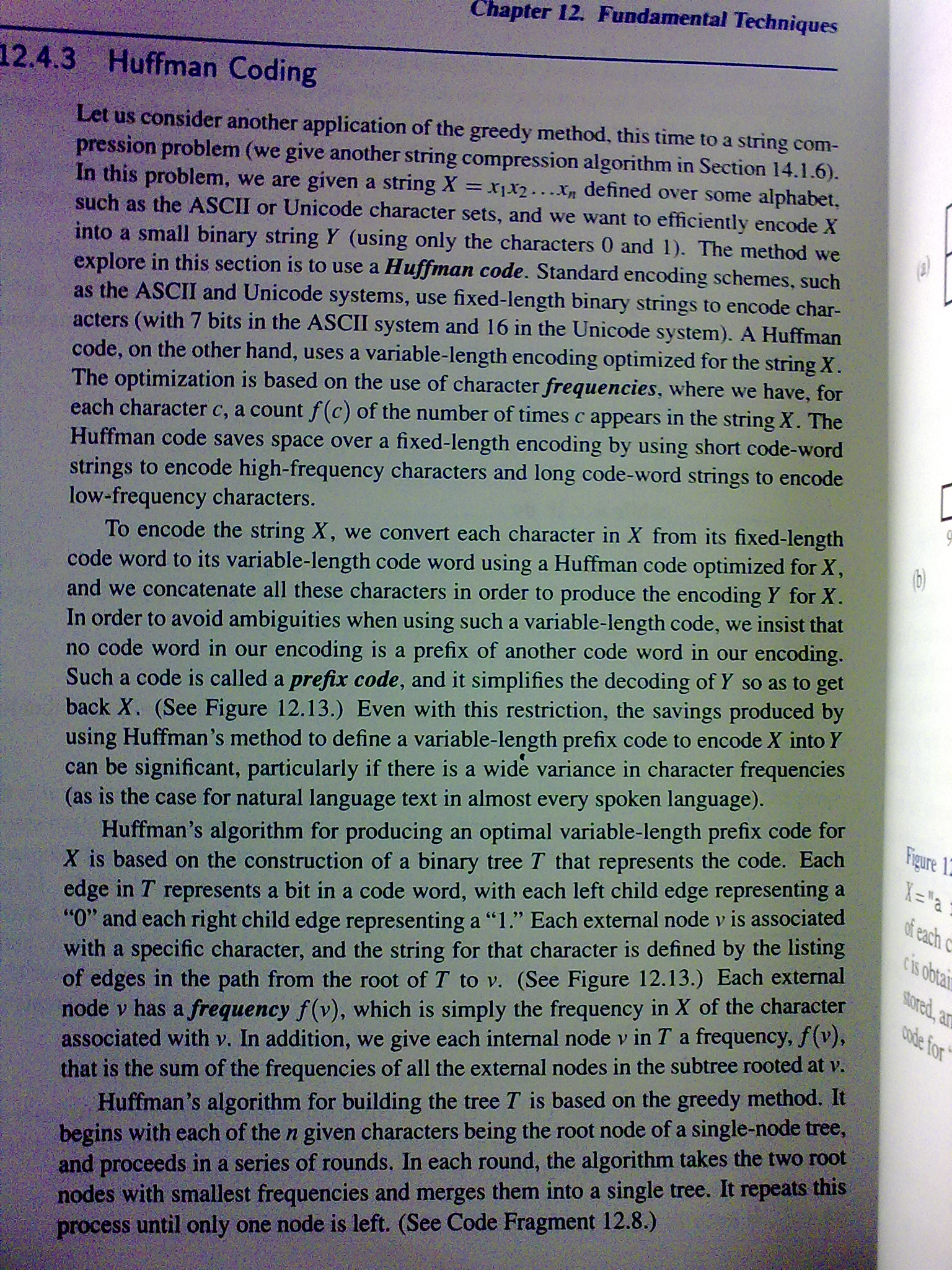
**You are required to deliver:**

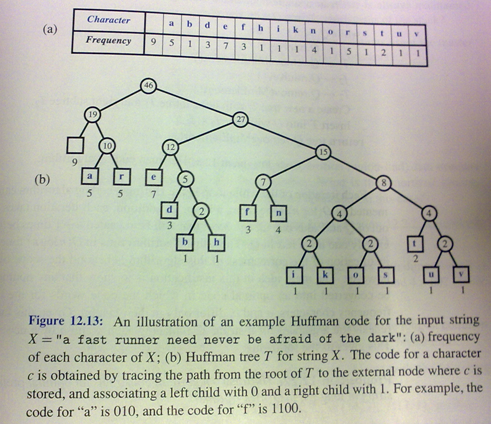
1. A Java application including **source code and .class file**
2. A **description of your program design** stating the design problems that you needed to address and how you surmounted them. For each of these, state the design problem and them your solution.
3. **Screen shots** illustrating test cases
4. A professionally written and word-processed **report** that contains all of the above.
5. Signed copy of the plagiarism form

### Academic Honesty

*Any work that you submit for continuous assessment or assignments must be done by you. Failure to acknowledge the source of a significant idea or approach is considered plagiarism and not allowed. Academic dishonesty will be dealt with severely. At a minimum, you will receive a mark of zero for the assignment.*

Extract from Data Structures and Algorithms in Java, Goodrich and Tamassia.





**A quick tutorial on generating a huffman tree – siggraph.org**

Let’s say you have a set of numbers and their frequency of use and want to create a huffman encoding for them:

**FREQUENCY** VALUES of **ALPHABET**

--------- -----

5 1

7 2

10 3

15 4

20 5

45 6

Creating a huffman tree is simple. Sort this list by frequency and make the two-lowest elements into leaves, creating a parent node with a frequency that is the sum of the two lower element's frequencies:

12:\*

/ \

5:1 7:2

The two elements are removed from the list and the new parent node, with frequency 12, is inserted into the list by frequency. So now the list, sorted by frequency, is:

10:3

12:\*

15:4

20:5

45:6

You then repeat the loop, combining the two lowest elements. This results in:

22:\*

/ \

10:3 12:\*

/ \

5:1 7:2

and the list is now:

15:4

20:5

22:\*

45:6

You repeat until there is only one element left in the list.

35:\*

/ \

15:4 20:5

22:\*

35:\*

45:6

57:\*

\_\_\_/ \\_\_\_

/ \

22:\* 35:\*

/ \ / \

10:3 12:\* 15:4 20:5

/ \

5:1 7:2

45:6

57:\*

102:\*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ \\_\_

/ \

57:\* 45:6

\_\_\_/ \\_\_\_

/ \

22:\* 35:\*

/ \ / \

10:3 12:\* 15:4 20:5

/ \

5:1 7:2

Now the list is just one element containing 102:\*, you are done.

This element becomes the root of your binary huffman tree. To generate a huffman code you traverse the tree to the value you want, outputing a **0** every time you take a left hand branch, and a **1** every time you take a right hand branch. (normally you traverse the tree backwards from the code you want and build the binary huffman encoding string backwards as well, since the *first* bit must start from the top).

**Example**: The encoding for the value **4** (15:4) is **010**. The encoding for the value **6** (45:6) is **1**

Decoding a huffman encoding is just as easy : as you read bits in from your input stream you traverse the tree beginning at the root, taking the left hand path if you read a **0** and the right hand path if you read a **1**. When you hit a leaf, you have found the code.