G51PRG Exercise Four-C: Character Frequency

Steven R. Bagley

Introduction

This exercise involves finding the frequency of individual letters (i.e. '`a`' to '`z`') in a plain text file. This will require you to process the file character-by-character, and once it has processed the whole file it should print out a histogram of the number of times each character appears.

Letter frequency

For this exercise, you are to develop a C program that can open a text file (using `fopen()` and with the name of the file passed in on the command line) and to find the number of times of each letter appears in the file (you may want to look at using the `isalpha()` function here).

You'll need an array of 26 `int`s to store the count of each letter — you can treat upper and lower-case as identical. Initialise all its elements to zero, and then increment the appropriate counter as each letter is encountered.

Output

Once you've successfully processed the entire file, you should output the count of each character *that appears in the file* (i.e. those with a count greater than zero) and also as a vertical histogram. So, for the input:

```
The cat sat on the green mat
```

your program should output:

```
The letter 'a' occurs 3 times.
The letter 'c' occurs 1 times.
The letter 'e' occurs 4 times.
The letter 'g' occurs 1 times.

The letter 'h' occurs 2 times.
The letter 'm' occurs 1 times.
The letter 'n' occurs 2 times.

The letter 'o' occurs 1 times.
The letter 'r' occurs 1 times.
The letter 's' occurs 1 times.
The letter 't' occurs 5 times.

  5                     *
  4                     *
  4     *               *
  3     *               *
  3 *   *               *
  2 *   *               *
  2 *   *  *     *      *
  1 *   *  *     *      *
  1 * * * **    ***   ***
  0 * * * **    ***   ***
  0 *************************
... abcdefghijklmnopqrstuvwxyz
```

Each bar in the histogram represents the number of times a character appeared in the file and that each step up the histogram should be represented by two lines (as represented in the above example). You will need to draw this downwards from the top printing a '`*`' or a space appropriately for each letter depending on how far down the graph you have got. You should not print more than 20 lines (i.e. from zero to ten) in your histogram, if the count is above ten then you should print a '`+`' at the top of the histogram to show it goes higher.

A selection of text files for you to test your program on can be found at:

`http://g51prg.cs.nott.ac.uk/Distribution/Coursework/cswk4c.zip`