UMass Boston CS 240
Homework 2
Due 10/02/2019 18:45

1 Histogram of Word Lengths

Write a C program that reads from `stdin` till EOF and analyzes the lengths of the words in the input. We constrain the input characters are all printable characters, which can be show in a text file.
For all the printable characters, there are alphanumeric characters which are digits, and letters of both uppercase and lower case, and punctuation characters or non alphanumeric characters like, ',",- _, etc.

Now, Let's define that a word is only made of alphanumeric characters, and delimiters between words are all the non-alphanumeric printable characters.

For example, each of the following is a word.

- Homework
- CS240
- 2

The following strings should be broken into multiple words.

- "you'll" has two words: "you" and "ll".
- "ALL'S" has two words: "ALL" and "S".
- "UTF-8" has two words: "UTF" and "8".
- "2/19/2019 17:00" has five words: "2", "19", "2019", "17", and "00".
- "www.gutenberg.org" has three words: "www", "gutenberg", and "org".

You can build upon the word counting code on page 20 of K&R. As you read a word one character at a time, keep track the number of characters you have read. When you reach the end of a word, you have its length. Then you increment a counter that keeps track the number of words of this particular length. Use an array of these counters. We test your code with `CompleteShakespeare.txt`. The longest word you will encounter has 27 characters and there is only one of it.

For output, you should print 29 lines:
1, For the first line, should print out the only longest word.
2, The second line is a new line.
3, In each line of the rest, you print the word length (in width 2, right side aligned), a space, the number of words of that length (width 6, right side aligned), a space, and then asterisks. Use one asterisk for each 4,000 words. If there are fewer than 4,000, you still print one asterisk for them, because we cannot print a fractional asterisk. For example, print none for 0 length , print one asterisk for 1 to 4,000 words, and two asterisks for 4,001 to 8,000 words, and so on. The asterisks constitute the histogram of word lengths. For the input `CompleteShakespeare.txt`, your code should print exactly like Figure 1.

```
honorificabilitudinitatibus

 1  63691 ***************
 2 166375 ****************************************
 3 204211 **************************************************
 4 223161 *******************************************************
 5 121472 ******************************
 6  80386 ********************
 7  59379 ***************
 8  35083 *********
 9  20351 ******
10  10067 ***
11   3771 *
12   1353 *
13    454 *
14    247 *
15     77 *
16      3 *
17      4 *
18      0
19      0
20      0
21      0
22      0
23      0
24      0
25      0
26      0
27      1 *
```

Figure 1: Output for `CompleteShakespeare.txt`

2 Directory for This Assignment

Make a directory `/home/user??/cs240/hw2`. Write your program in the name of `histo.c`. Compile and test it as follows.

```
.../hw2$ gcc histo.c -o histo.out
.../hw2$ ./histo.out < CompleteShakespeare.txt > output.txt
```

Write comments to explain your code when necessary – how you determine a character is alphanumeric, how you extract one word, and how you convert a count to the number of asterisks to print.

There should be three files in the directory: `histo.c`, `histo.out`, and `output.txt`. Other files are allowed.

3 Requirements

Do not modify the files after the due date until you receive your score.

I will examine your codes one by one and test them with another text file after the deadline.

Write your code with comments and proper spacing, especially when you don't think you will get the full score. Partial scores will be given when you clearly express what you have done for different snippets of your code.

If your program does not compile or run and your code is hard to read, then a lower score is guaranteed.

No late homework will be accepted.