

CSCI 1312 (Introduction to Programming for Engineering), Fall 2016

Homework 8x

Credit: Up to 25 extra-credit points.

1 Reading

Be sure you have read (or at least skimmed) the assigned readings from chapters 7 and 11.

2 Honor Code Statement

Please include with each part of the assignment the Honor Code pledge or just the word “pledged”, plus one or more of the following about collaboration and help (as many as apply).¹ Text *in italics* is explanatory or something for you to fill in. For written assignments, it should go right after your name and the assignment number; for programming assignments, it should go in comments at the start of your program.

- This assignment is entirely my own work.
- This assignment is entirely my own work, except for portions I got from the assignment itself (*some programming assignments include “starter code”*) or sample programs for the course (*from which you can borrow freely — that’s what they’re for*).
- I worked with *names of other students* on this assignment.
- I got help with this assignment from *source of help — ACM tutoring, another student in the course, the instructor, etc.*
- I got significant help from *outside source — a book other than the textbook (give title and author), a Web site (give its URL), etc.. (“Significant” here means more than just a little assistance with tools — you don’t need to tell me that you looked up an error message on the Web, but if you found an algorithm or a code sketch, tell me about that.)*
- I provided significant help to *names of students* on this assignment. (*“Significant” here means more than just a little assistance with tools — you don’t need to tell me about helping other students decipher compiler error messages, but beyond that, do tell me.*)

3 Programming Problems

Do the following programming problems. You will end up with at least one code file per problem. Submit your program source (and any other needed files) by sending mail to `bmassing@cs.trinity.edu` with each file as an attachment. Please use a subject line that mentions the course and the assignment (e.g., “csci 1312 hw 8x” or “CS1 hw 8x”). You can develop your programs on any system that provides the needed functionality, but I will test them on one of the department’s

¹Credit where credit is due: I based the wording of this list on a posting to a SIGCSE mailing list. SIGCSE is the ACM’s Special Interest Group on CS Education.

Linux machines, so you should probably make sure they work in that environment before turning them in.

Yes, this writeup is long. But I think the code you write need not be, and it's an interesting problem!

You may have heard claims that E is the most frequently-used character in English text, followed by T, and so forth. Your mission for this assignment is to write two programs that together will allow you to find out how true this claim is for selected text (and, okay, to give you practice working with some course topics):

- The first program analyzes a single file of plain-text, counting occurrences of each alphabetic character and writing results (characters and counts, but only for characters that occur at least once) to an output file.
- The second program merges one or more files produced by the first program and writes results to an output file.

(Why two programs? Mostly pedagogical reasons.) Writing the programs from scratch is nontrivial (though you could probably do it), so to make it more doable I'm providing starter code that reduces what you need to do and also gives you some practice with UNIX `make`, discussed in class. Once you have an output file produced by the second program, you can use the Linux command

```
sort -n -r outfilename
```

to display the results in a way that shows the most-often-used letter first, etc.²

To make the programs as portable as possible, I've written a function that builds an "alphabet" (text string containing all the characters `ch` for which `islower(ch)` is true), and I want you to write two functions, described below, that use this string.⁴ You'll be using these functions in both programs, so it makes sense to put them in a separate file that both programs can make use of (rather than copying the code into both programs). There are several ways to combine this "library" code with the two programs, but what I want you to do is to use the Linux utility `make`, as discussed in class. Starter code, with `FIXME` comments showing where you need to add code:

- [alphabet.h](#)⁵, declarations of "library" functions.
- [alphabet.c](#)⁶, starter definitions of "library" functions.
- [countalpha.c](#)⁷, starter code for `countalpha` program.

²Where to find plain-text files to use as input is an interesting question; when I assigned this problem last year one could get such from [Project Gutenberg](#)³ but they don't seem to offer really-plain-ASCII-text versions any more. However, you can get `plain-text-in-UTF8` versions, and then you can convert to ASCII with the Linux command

```
iconv -f UTF-8 -t US-ASCII -c infile.txt >outfile.txt
```

Word-processing programs will also export to plain text, though once again you might be well-advised to check that what you get is really plain ASCII text and not something with other characters. The above command may help here too.

⁴Note, for what it's worth, that it would be easy to expand this "alphabet" to include digits or punctuation or any other characters you wanted to count — and you wouldn't need to change the rest of the program.

⁵http://www.cs.trinity.edu/~bmassing/Classes/CS1312_2016fall/Homeworks/HW08x/Problems/alphabet.h

⁶http://www.cs.trinity.edu/~bmassing/Classes/CS1312_2016fall/Homeworks/HW08x/Problems/alphabet.c

⁷http://www.cs.trinity.edu/~bmassing/Classes/CS1312_2016fall/Homeworks/HW08x/Problems/countalpha.c

- `mergecounts.c`⁸, starter code for `mergecounts` program.
- `Makefile`⁹, “makefile” to build the two programs.

The `Makefile` includes instructions for “building” the project. Note that just using `gcc` with a single program, as we’ve been doing, won’t work, but once you have all the above files downloaded, typing `make` will produce two executables, `countalpha` and `mergecounts` that you can run (although they won’t do anything very interesting). You might try that before starting to write code.

Instructions for specific files you need to change:

1. (Optional — up to 5 extra-credit points) The first file you need to change is `alphabet.c`. Fill in the bodies of the two functions for which code is not provided. You can check that your code at least compiles by typing `make` again.
2. (Optional — up to 10 extra-credit points) The next file you need to change is the code for the first program, the one that analyzes a single input file and produces an output file. The starter code checks that there are two command-line arguments (filenames for input and output) and opens the input file. Add code to do the following:
 - Declare an array of counters, one for each character in `alphabet`, and of course initialize it. (`strlen(alphabet)` tells you how many you need.)
 - Read the input file a character at a time and count, using the function `char_to_index`, how many times each alphabetic character occurs (but use `tolower()` first to turn any upper-case characters into lower-case). Note that this function also tells you whether the character is even alphabetic (so you don’t need `isalpha` — it returns -1 if not. Note also that to get full credit for this part you *must* use this function rather than trying to figure out another way which index to use.
 - Count the total number of characters and how many were alphabetic.
 - For every alphabetic character that occurs at least once, write to the output file a line with the character and the count. (Use the function `index_to_char` to get right character for each index.)
 - Print the total number of characters and the number of alphabetic characters.

This is probably easiest to understand with examples. If the input file looks like this:

```
testing 1 2 3 4?
```

```
TESTING 4 3 2 1!
```

the output file should look like this:

```
e 2
g 2
i 2
n 2
s 2
t 4
```

⁸http://www.cs.trinity.edu/~bmassing/Classes/CS1312_2016fall/Homeworks/HW08x/Problems/mergecounts.c

⁹http://www.cs.trinity.edu/~bmassing/Classes/CS1312_2016fall/Homeworks/HW08x/Problems/Makefile

and the program should print this:

```
alphabet 'abcdefghijklmnopqrstuvwxy'  
14 alphabetic characters, 36 total characters
```

And if the input file looks like this:

```
Now is the time for all good persons  
to come to the aid of their party!
```

the output file should look like this:

```
a 3  
c 1  
d 2  
e 6  
f 2  
g 1  
h 3  
i 4  
l 2  
m 2  
n 2  
o 9  
p 2  
r 4  
s 3  
t 7  
w 1  
y 1
```

and the program should print this:

```
alphabet 'abcdefghijklmnopqrstuvwxy'  
55 alphabetic characters, 72 total characters
```

3. (Optional — up to 10 extra-credit points) The last file you need to change is the code for the second program, the one that merges output from repeated executions of the first program. The starter code checks that there is at least one command-line argument, builds the “alphabet”, and calls a function `process_file` for each input filename to process that single file. Add code to do the following:
- Actually do something in `process_file`, in addition to printing the filename — read the file a line at a time (see below) and use this information to update the array of counters. See below for a discussion of how to do this. Print an error message if the file cannot be opened or has errors (see below), The function should return `true` if everything was okay, `false` if there was an error.
 - After all input files have been processed, write to the output file a line for each nonzero alphabetic character, with the count of that character and the character.

About reading lines from the input file, to get some practice with an additional way of reading text input, I want you to use `fgets` to get a line at a time and `sscanf` to then pick out the

character and the count. Your program should do something sensible if an input line is too long to fit into the array you declare to hold it (such as printing an error message and throwing away the rest of the line). It should also print an error message for any input line that isn't in the right form (character, space, integer, end-of-line). The starter code has some additional hints.

Here too this is probably easiest to understand with an example. Given the two output files shown earlier, the program should combine them to produce an output file containing

```
3 a
1 c
2 d
8 e
2 f
3 g
3 h
6 i
2 l
2 m
4 n
9 o
2 p
4 r
5 s
11 t
1 w
1 y
```

and print this:

```
alphabet 'abcdefghijklmnopqrstuvwxy'
processing input file sample1-out.txt
processing input file sample2-out.txt
```

Finally, the program should give an error message for every line of this input file:

```
hello
x
100
x 1000x
```