**COMP10002 Assignment 1, 2017s2**

**Last updated: September 6, 2017**

[**Submission Instructions**](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/submission.html)**A complete submission is shown in the lecture recording starting at around the 10 minute mark of the class on 5 September.**

**Information and Resources**

1. *Blank lines*: Sir, what about blank lines? What about if the whole input file is empty? *Answer*: If you look at the outputs linked below you'll see that blank lines are presemed to have line numbers, but don't generate any output. A completely empty input file will generate the Stage 1 output, but no Stage 2 or Stage 3 or Stage 4 output.
2. *Apostophes and etc*: Dear Sir, your output has the lines
3. many miles I've fallen by this time?' she said aloud. 'I must be getting
4. S2: line = 2, bytes = 72, words = 15

but it seems you are counting "I've" as two words rather than one. Shouldn't that be just one word? *Answer*: The definition of "word" we are working to makes it two words. (And, in any case, isn't "I've" short for "I have"?)

But Sir, if we count "I've" as two words, and one of the query terms is "v", then won't that mean there is a match that has to be counted? *Answer*: Correct, and that is what the specification intends and that is what my program does.

1. *Value of argc*: Dear Sir, you say in the handout that argc of zero represents no query. But don't you mean argc of 1? *Answer*: Yes, I guess I do...
2. *Access Denied on dimefox*: Note that you won't have an account yet on dimefox if you have never logged in to a lab machine so far this semester, for example, if you have been using your own computer for all the workshops, or if you simply haven't attended any workshops. The symptoms of this will be an "access denied" message when you try and connect with scp/puttyscp or ssh/putty when you want to copy/submit your program. You will need to login to a lab machine, get you account initialized, and then wait a few hours for everything to percolate through the various processes involved with transferring those account details on to dimefox. Then you'll be in a position to ssh/scp and eventually submit.

If you still have problems after you have taken this step, and are sure you are using the right password (and can log in to other University services using it), send me an email confirming that you have taken all of these steps and **still** can't get access to dimefox. *Don't leave this until the last minute. It is a problem that can't be fixed in a minute!*

1. *What Can Be Stored?*: Dear Sir, you say "You can only retain five lines and their scores at any given time, plus the current line that is being processed", do you mean that *exactly*, or do you really mean "You can only retain five data structures representing five lines (possibly, for example, in original as well as processed/parsed format) and their scores at any given time, plus the current line that is being processed?". *Answer:*Yes, good question, and I mean: you can retain at most five instances of the data structure(s) that represent a single line.
2. *Error in Printed Handout*: There was a small typo in the printed version of the handout that was ciculated in class on Friday 1 September that has now been corrected in the online version[here](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/ass1.pdf): the very first example in the handout shows
3. mac: ./ass1 < alice-eg.txt
4. S1: No query specified, must provide at least one word
5. mac: ./ass1 Lat 66 loNg 32 words < alice-eg.txt
6. S1: query = lat 66 loNg 32 words
7. S1: loNg: invalid character(s) in query
8. mac:

and it should be showing

mac: ./ass1 < alice-eg.txt

S1: No query specified, must provide at least one word

mac: ./ass1 lat 66 loNg 32 words < alice-eg.txt

S1: query = lat 66 loNg 32 words <--------

S1: loNg: invalid character(s) in query

mac:

with a lowercase "ell" in the second example commandline marked by the arrow.

1. *Marking Rubric*: The marking rubric is linked [here](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/ass1-rubric-master.html). Lines that do not apply to your program will be removed during the marking process; your mark will then be the sum of the lines that remain, positives and negatives. Marks won't go below zero in each section, and won't go below zero overall either.
2. *Attribution for Re-Used Code*: It is ok to make use of code (for example, insertionsort, and/or getword(), and etc) from the book or from the lecture slides or from other published/public sources, but you should remember to add an attribution as a comment to each relevant function, saying where you got it from, what modifications you added to make it suit your purpose, and so on -- exactly as you would when quoting some other author when writing an essay.

Of course, the expectation is that the assembly and "glueing together" of these bits to make a final program will all be your own work, and that the "quoted" bits will be a relatively small fraction of the "new" output you are being asked to generate. So it is **not** ok to take a whole solution from somewhere else, even if it appears on the web; and it is **not** ok to solicit or commission a solution by posting the specification to a forum or web site and asking for "assistance" or "guidance" or "suggestions". Just as it wouldn't be ok to submit something you found online in response to an assignment that involved writing an essay.

And a reminder of what it says in the specification: we **will** be using similarity-checking software across all the submissions, and we **will** be referring cases of suspected academic misconduct for disciplinary hearings run by the School of Engineering, and in the past those hearings **have** resulted (including multiple times in **this** subject) in students being awarded penalties including final marks of **zero** for the subject, regardless of their other components of assessment.

1. *Debugging (more)*: If a C program encounters a run-time error and exits, there might still be pending output that has not been written. This is a particular problem in the submit environment, because it can look like the program is failing before it generates any output at all. If in doubt, add fflush(stdout) function calls after each of your your debugging printf()'s (or even, add it to the macro), to force all pending output to be written immediately. You'll then be able to get a much clearer idea of how far the program is getting before it fails.
2. *Debugging*: Try putting this at the top of your program:
3. #define DEBUG 1
4. #if DEBUG
5. #define DUMP\_DBL(x) printf("line %d: %s = %.5f\n", \_\_LINE\_\_, #x, x)
6. #else
7. #define DUMP\_DBL(x)
8. #endif

and then, later in your code, where you have a double variable (say) score, try

DUMP\_DBL(score);

Then change DEBUG to 0 at the top of the program, compile it again, and then run it again. Get it?

Can then add DUMP\_INT and DUMP\_STR, and get the extra output turned on whenever you need it to understand what your program is doing. Then turn it all off again with one simple edit.

1. *Trouble with newline characters*: Text files that are created on a PC, or copied to a PC, edited and then saved again on the PC, may end up with PC-format two-character (CR+LF) newline sequence, see the [Wiki page](https://en.wikipedia.org/wiki/Newline) for details.

If you have compiled your program on a PC, and it receives a CR+LF sequence, then getchar() will consume them both, and hand a single-character '\n' newline to your program. So in that sense, everything works as expected. Likewise, on a PC when you write a '\n' to stdout, a CR+LF pair will be placed in to the output file (or passed through the pipe to the next program in the chain).

The problems arise when you copy your program and a PC-format test file to a Unix system and then try compiling and executing your program there. Now the CR characters get in the way and arrive via getchar() into your program as stand-alone '\r' characters.

The easiest way to defend against these confusions is to write your program so that it looks at every character that it reads, and if it ever sees a CR come through, it throws it away. That way, if you do accidentally get CR characters in your test files on the Unix server (or on your Mac) your program won't be disrupted by them. Here is a function that you should use to do this:

int

mygetchar() {

int c;

while ((c=getchar())=='\r') {

}

return c;

}

Then just call mygetchar() whenever you would ordinarily call getchar(), on both PC and Mac.

Because most of you work on PCs (including in the labs), the test files that are provided have been created **with** the PC-style CR+LF newlines, and should work correctly when copied (use right-click->"Save as") to a PC. With mygetchar() they can also be used on a Mac, but won't interact sensibly using the standard getchar() function.

To be consistent, the final post-submission testing will also be done using PC-style input files but will be executed **on a Unix machine**, meaning that **all** submitted programs will need to make use of mygetchar().

You can use the "Preferences->Encodings" menu ("screwdrive/hammer Options->Encodings" in the PC version) in jEdit to select whether to use Unix (LF) or DOS/Windows (CR+LF) encodings in any test files that you create with jEdit. Note that this only applies to newly created files. jEdit will by default respect the formatting in any current files.

Note also that jEdit doesn't automatically add a newline after the last line of text files, you need to put it there explicitly yourself (just press enter one more time, so that jEdit thinks there is an empty line at the end of the file). Watch out for this problem if you are creating your own test files on Mac or PC.

If in any doubt, use od -a <file> in a Unix shell to look at the byte-by-byte contents of a file, and check which format is being used, and whether there is a final newline character (or finalCR+LF pair). You can do this on a PC by starting the MinGW shell and then using cd to reach the right directory. There is an od version available within the MinGW shell on the PCs in the labs. On a Mac, Terminal is a Unix shell.

1. *Tabs*: The default in jEdit is for tabs to be aligned every 8 character positions. Some of you have altered that to four (Preferences->Editing->Tab width), to reflect the layout that the programs in the book have. Then, on submission, the tabs have "appeared" in the output as being 8 again, which can make your program spill past the 80-character RH boundary. When I run the programs for marking, they'll **all** get formatted with tabs reflecting 4 character positions, not 8. But don't use any fewer than 4 in your jEdit (or other editor) settings.
2. *Magic numbers*: Here is a summary of the rules about magic numbers:
   * Where a number is totally self-defining, I'm happy for it to be used any number of times without a hash-define, provided the code is commented each time and/or explicitly sensible variable names are used. For example, in
   * /\* compute percentage \*/
   * pcent = 100.0\*count/totcount;

I wouldn't expect 100 to have been hash-defined, since the comment explains the role of the 100, and it isn't going to change, ever, even if other percentages are calculated in the program using 100 too.   
This rule also allows 0 and 1, of course, unless they represent something other than the additive and multiplicative identities, in which case they should be hash-defined.   
This rule also allows while (scanf("%lf%lf", &x, &y)==2), since the 2 is immediately obvious from the adjacent context (two variables to be read).

* + Where a constant is one that is a fact that is in no way ever going to be varied, then provided it only appears once in the program and is explained with a comment, then it need not be hash-defined. The example here is the -32.0 in the temperature conversion computation, assuming that it is entirely within a function called Cels2Fahr or etc and that it isn't used in other places scattered through the program. Anything of this type that appears even twice should be hash-defined.
  + Where a factual constant is used more than once in a program, even if all occurrences are in a single function, it should be hash-defined.
  + Where a constant is one that is clearly an artifact of the problem description or the program that implements the solution (for example, numbers like MAXINT, or the number of variables), then they must be hash-defined, even if only used once in the program.

Make sense?

1. *Test data*: Test data for Assignment 1, and some sample outputs:
   * [alice-eg.txt](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/alice-eg.txt)  
     (use right-click-saveas on the link to make a copy)
   * [pg11.txt](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/pg11.txt)  
     (use right-click-saveas on the link to make a copy)
   * [output for "ass1 ali lat long < alice-eg.txt"](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/test0-out.txt)
   * [output for "ass1 f dow long way < alice-eg.txt"](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/test1-out.txt)
   * [output for "ass1 ali lat long < pg11.txt"](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/test2-out.txt) (very long!)
   * [output for "ass1 f dow long way < pg11.txt"](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/test3-out.txt) (very long!)
2. *Specification*: The specification for the project was provided [here](http://people.eng.unimelb.edu.au/ammoffat/teaching/10002/ass1/ass1.pdf) on Friday 1 September.

*Alistair Moffat, ammoffat@unimelb.edu.au*