Scientific Programming Assignment 1

MPhil in Computational Biology

October 13, 2018

If there are errors found, I will update the assignment on the web at
`http://github.com/sje30/rpc2018`

Due date: to be determined in class.

Please submit your report to Moodle as a PDF: NO OTHER FORMATS ARE ACCEPTED.

This course work will consist of 25% towards your overall mark for this module.

The data files for this assignment are stored in:

`http://github.com/sje30/rpc2018/a1`

Only R packages that come installed by default with a R installation can used for this assignment. If in doubt, check with Stephen whether you can use a particular package.

Include in your report an appendix containing your R code. You can do this in LaTeX with the *listings* package or with the following snippet:

```
\usepackage{verbatim} %% in the preamble
\verbatiminput{myfile.R}
```

1 Words [15 marks]

Read in the file `usr-share-dict-words`; this is a dictionary that comes with Ubuntu linux. Each line contains one word. You should ignore the case of the letters, i.e. treat upper and lower case as the same.

1. Read in the file and convert all words to upper case. Use *unique()* to keep only one copy of each word. (e.g. Zest and zest should count as one word ZEST.) How many unique words are there? [1 mark]

2. How many words contain an apostrophe (′)? Remove these words from the rest of the analysis. [1 mark]

3. How many words contain non-ASCII characters? (Remove these words from the rest of the analysis; the remaining words are called the *database* in the remainder of the question.) [1 mark]

4. Find all the words that are in the database as the two related forms *X*OG and *X*OGUE. For example, CATALOG and CATALOGUE have this pattern. [2 mark]

5. Read in the file `scrabble.txt` from which you create a vector called `scores` where element *i* stores the scrabble score of the *i*th letter of the alphabet. [1 mark]

6. Compute the scrabble score for each word in the database. Plot the distribution of scores. What is the highest-scoring word? [3 marks]

7. The reverse complement of a word is where you reverse the characters in the word, and then replace A with Z, B with Y, C with X and so on. For example, the reverse complement of HILLY is BOORS. Find all words W where both W and its reverse complement are both in the database. [2 marks]

8. Given the following nine letters:

   `F A L U Y P L N I`

   how many words of four or more letters can you find that are in the database AND all contain the letter A? Each letter can be used only once, and you should be able to find a nine-letter word. [4 marks]

2 Examination marking [10 points]

The data for this exercise is in `grading` folder.

Twelve students have sat a multiple-choice exam. The exam had 100 questions, and each answer was one of a,b,c,d,e. The file `crib.dat` stores the correct answer for each question (in order). The students had to answer 30 questions from the 100. Your job is to write a script that will mark each student's performance, and produce a data.frame which stores the results:

```
> results <- data.frame(student=1:num.students, score=correct,
                        grade=alpha.grades, rank=rank)
> print(results)
   student score grade rank
1        1    19     B  6.0
2        2    xx     x  xxx
3        3    xx     x  xxx
4        4    xx     x  xxx
5        5    xx     x  xxx
6        6    xx     x  xxx
7        7    xx     x  xxx
8        8    xx     x  xxx
9        9    xx     x  xxx
10      10    xx     x  xxx
11      11    xx     x  xxx
12      12    xx     x  xxx
```

To help you get started, you can see that student 1 got 19/30 correct, their rank was 6/12 (1st rank for highest) and their grade was B. Grades are determined using the datafile `grade.txt`; convert the score into a percentage, take the floor() to convert percentage to an integer, then find which grade band the score falls in.

Hints: use read.table(, header=TRUE) to read in a student file. The name of the datafile to read in should be generated using paste(). You can use scan() to read in the crib.dat file.

The invigilator of the exam suspects that a student was cheating, but cannot recall which student it was. Write a program that will automatically check whether a pair of students have similar results.