CSC 121, Spring 2017 — Small Assignment #1

Worth 5% of the course grade. Due by the start of class on January 31. Instructions for how to hand it in will be posted on the course web page later. This assignment may be handed in late, with a 20% penalty, by start of class on February 3. Assignments will not usually be accepted after that. Contact the instructor as soon as possible if you have a legitimate excuse (such as documented illness) for handing in the assignment late (without penalty).

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. In particular, you shouldn't leave a discussion with someone else with any written notes (either paper or electronic).

In this assignment, you will write and test R functions for locating two-word phrases in a piece of text (a sequence of words), and replacing the words in the phrases that are found with the combined phrase.

This might, for example, be useful if you were doing a statistical analysis of text in documents, perhaps in order to predict the topic of a document from how often various words occur. (This would be useful for web searches.) You might find that documents about navigation contain lots of occurrences of the words "North", "South", "East", and "West". But you could get misleading results in this respect if you counted occurrences that are in the phrases "North Dakota", "South Carolina", "East Timor", etc. So you might want to convert the words making up such known phrases into a single string for the phrase, which would no longer be seen as matching the words it is composed of.

You will write two R functions for this assignment. The first should be called `replace_one_phrase`. It should take as arguments a vector of strings (of any length) and a single two-word phrase (as a vector of two strings). Your definition of this function should start as follows:

```
replace_one_phrase <- function (text, phrase)
```

The value returned by this function should be a modified form of `text`, in which occurrences of `phrase` have been combined into a single string, using `paste`, with the `sep="_"` option.

Here is an example of use of `replace_one_phrase`:

```
> t <- c("I","like","hot","dogs","a","lot","hot","dogs","forever")
> t
[1] "I"       "like"    "hot"     "dogs"    "a"       "lot"     "hot"
[8] "dogs"    "forever"
> replace_one_phrase(t,c("hot","dogs"))
[1] "I"        "like"     "hot_dogs" "a"        "lot"      "hot_dogs" "forever"
```

The phrase replacement should be done sequentially, starting at the beginning of `text`. This matters only if the two words of the phrase are the same, for example:

```
> replace_one_phrase(c("a","xx","xx","xx","b"),c("xx","xx"))
[1] "a"     "xx_xx" "xx"    "b"
```

The second function you write should be called `replace_phrases`. It should take as its first argument a vector of strings, as for `replace_one_phrase`. Its second argument should be a list of phrases, each of which is a string of length two. You should start your definition of this function as follows:

```
replace_phrases <- function (text,phrase_list)
```

The value returned by replace_phrases should be text modified by replacing all occurrences of the phrases in phrase_list with the corresponding combined string. You should make use of your replace_one_phrase function to do the replacements in replace_phrases, calling it multiple times. Replacements for phrases in phrase_list should be done sequentially, which matters if the first word in one phrase is the same as the second word in another phrase, as in this example:

```
> replace_phrases (c("ab","pq","xy"), list(c("ab","pq"),c("pq","xy")))
[1] "ab_pq" "xy"
```

You *must* indent your function definitions properly, as illustrated by the examples in the lecture slides.

You should put your definitions of replace_one_phrase and replace_phrases in an R script file, which contains only these definitions. In another R script file, you should put some R commands that test these functions. These tests should include the examples shown above, plus the following examples (shown with correct output below):

```
> test_text1 <- c(
+ "I","went","South","to","North","Dakota","then","South","to","South","Carolina")
> test_text2 <- c(
+ "North","Dakota","is","North","of","South","Dakota","which","North","Dakota",
+ "is","North","of")
> replace_one_phrase (test_text1,c("North","Dakota"))
 [1] "I"            "went"         "South"        "to"           "North_Dakota"
 [6] "then"         "South"        "to"           "South"        "Carolina"
> replace_one_phrase (test_text1,c("South","Carolina"))
 [1] "I"              "went"           "South"          "to"
 [5] "North"          "Dakota"         "then"           "South"
 [9] "to"             "South_Carolina"
> replace_one_phrase (test_text2,c("North","Dakota"))
 [1] "North_Dakota" "is"           "North"        "of"           "South"
 [6] "Dakota"       "which"        "North_Dakota" "is"           "North"
[11] "of"
> phrases <- list(c("North","Dakota"),    c("South","Dakota"),
+                 c("North","Carolina"),  c("South","Carolina"),
+                 c("New","Brunswick"),   c("British","Columbia"))
> replace_phrases (test_text1, phrases)
[1] "I"              "went"           "South"          "to"
[5] "North_Dakota"   "then"           "South"          "to"
[9] "South_Carolina"
> replace_phrases (test_text2, phrases)
 [1] "North_Dakota" "is"           "North"        "of"           "South_Dakota"
 [6] "which"        "North_Dakota" "is"           "North"        "of"
```

All of these required tests are in an R script file on the course web page, which you should download as the starting point for your script file of tests. You should add additional tests of your own to these tests, that test whether the functions work in both typical and unusual situations.

You should hand in the two script files you produce (of function definitions and tests) and the output of the test script. Details about how to hand these in will be provided later.