HW: Week 3

36-350 – Statistical Computing

Week 3 – Spring 2022

Name: Ethan Vertal

Andrew ID: evertal

You must submit **your own** HW as a PDF file on Gradescope.

```
shakespeare.lines = readLines("http://www.andrew.cmu.edu/user/mfarag/shakespeare.txt")
```

Question 1

(10 points)

Display the lines of text in `shakespeare.lines` that contain both of the strings "purse" *and* "gold" (in any order, separated by any amount of text). Do so only using regexp literals.

```
p <- grep("purse", shakespeare.lines, value = TRUE)
pg <- grep("gold", p, value = TRUE)
pg
```

```
## [1] "  HELENA. Take this purse of gold,  "
```

Question 2

(10 points)

Retrieve (but don't display) the lines of text in `shakespeare.lines` that contain "briers". Then break the retrieved lines into individual words (`strsplit(input," ")` to split the character vector `input` into words separated by spaces), and merge those words into a single character vector (`unlist()`!). How many unique words are there? Display the top five most commonly occurring words and how often they occur (combine `sort()` and `table()`!)

```
unlist(strsplit(grep("briers", shakespeare.lines, value = TRUE), " "))
```

```
##  [1] ""       ""       ""       ""       "When"   "briers" "shall"  "have"
##  [9] "leaves" "as"     "well"   "as"     "thorns"
```

```
table(unlist(strsplit(grep("briers", shakespeare.lines, value = TRUE), " ")))
```

```
##
##            as briers   have leaves  shall thorns   well   When
##      4      2      1      1      1      1      1      1      1
```

Question 3

(10 points)

In Q25 of Lab 3, you coded a regex to match all patterns of the following form: any letter (1 or more), then a punctuation mark, then "ve" or "ll" or "t", then a space or a punctuation mark. You called it `my.pattern`. Use `my.pattern`, along with `regexpr()` and `regmatches()`, to extract and display all the occurrences of the pattern in `shakespeare.lines`. Then repeat the exercise using `gregexpr()` instead of `regexpr`; note that here, you'll want to `unlist()` the output from `regmatches()`. Do you get the same vector of character strings? Why or why not?

```
# FILL ME IN
```

FILL ME IN: utilize line breaks for PDF output

Question 4

(10 points)

Come up with a strategy that splits punctuation marks or spaces, except that it keeps intact words like "I've" or "wasn't", that have a punctuation mark in the middle, in between two letters. (Or when the punctuation mark is at the beginning, as in "'em") Apply your strategy to `shakespeare.lines` as defined below such that you display only those words with punctuation marks. (Note that I end up with 704 [not necessarily unique, but total] words when I implement this strategy. Some include'", which we can easily remove in a subsequent post-processing step if we so choose. Hint: `[[:alnum:]]` is a good thing to use here.)

```
# FILL ME IN
```

Below, we read in lines of data from the Advanced National Seismic System (ANSS), on earthquakes of magnitude 6+, between 2002 and 2017. (You don't have to do anything yet.)

```
anss.lines = readLines("http://www.stat.cmu.edu/~mfarag/350/anss.htm")
date.pattern = "[0-9]{4}/[0-9]{2}/[0-9]{2}"
date.lines = grep(date.pattern,anss.lines,value=TRUE)
```

Question 5

(10 points)

Check that all the lines in `date.lines` actually start with a date, of the form YYYY/MM/DD. rather than contain a date of this form somewhere in the middle of the text. (Hint: it might help to note that you can look for non-matches, as opposed to matches, by changing one of `grep()`'s logical arguments.)

```
# FILL ME IN
```

Question 6

(10 points)

Which five days witnessed the most earthquakes, and how many were there, these days? Also, what happened on the day with the most earthquakes: can you find any references to this day in the news?

```
# FILL ME IN
```

```
FILL ME IN: utilize line breaks for PDF output
```

Question 7

(10 points)

Go back to the data in `date.lines`. Following steps similar to the ones you used in the lab to extract the latitude and longitude of earthquakes, extract the depth and magnitude of earthquakes. In the end, you should have one numeric vector of depths, and one numeric vector of magnitudes. Show the first three depths and the first three magnitudes. (Hint: if you use `regexpr()` and `regmatches()`, then the output from the latter will be a vector of strings. Look at this vector. The last four characters always represent the magnitudes. Use a combination of `substr()` and `as.numeric()` to create the numeric vector of magnitudes. Then use the fact that everything but the last four characters represents the depths. There are a myriad of ways to do this exercise, but this suggested way is the most concise.)

```
# FILL ME IN
```

Here we read in text containing the fastest men's 100-meter sprint times. We retain only the lines that correspond to the sprint data, for times 9.99 seconds or better.

```
sprint.lines = readLines("http://www.stat.cmu.edu/~mfarag/350/men_100m.html")
data.lines = grep(" +(9|10)\\.",sprint.lines)
sprint.lines = sprint.lines[min(data.lines):max(data.lines)]
```

Question 8

(10 points)

Extract the years in which the sprint times were recorded. Display them in table format. Do the same for the months. Be sure to extract the month of the sprint, not the birth month of the sprinter! (Hint: the month of the sprint is followed by a four-digit year; other instances of two digits in any given line are not. So you may have to extract more than you need, then apply `strsplit()`.)

```
# FILL ME IN
```

Question 9

(10 points)

Extract the countries of origin (for the sprinters). Note that countries of origin are given as a capitalized three-letter abbreviation. Display the table of country origins. Display the proportion of the list that is accounted for by sprinters from the US and Jamaica.

```
# FILL ME IN
```

Question 10

(10 points)

We conclude with a web scraping exercise. I want you to go to this web site. On it, you see there is a set of 12 bold-faced four-digit numbers: this is the number of submitted astrophysics articles for each month of 2019. I want you to extract these numbers and place them into a single vector, with each vector element having a name: Jan for the first vector element, Feb for the second, etc. You would use `readLines()` to read in the page (pass the URL directly to the function!); this creates a vector of strings. You would then use `regexpr()` and `regmatches()` to extract the numbers (plus potentially some other stuff that you may have to pare off using `substr()`). If necessary, use "view source" to look at the html code for the web page itself to determine how best to extract the 12 numbers and nothing else. You don't want to create a table; you simply want to output the vector of four-digit numbers and add the appropriate names. (Hint: see the documentation for `Constants`. `month.abb` might be helpful here.)

```
# FILL ME IN
```

Question 11

(10 points)

Make a string with "Regards," and then your first name, separated by a newline character. Print it to the console via `print()` and then via `cat()`. Comment on the difference.

```
s <- "Regards, \n Ethan"
print(s)
```

```
## [1] "Regards, \n Ethan"
```

```
cat(s)
```

```
## Regards,
##  Ethan
```

```
The difference is that printing the string prints what it literally says and cat accounts for the escape
```

Question 12

(10 points)

Transform the vector so that the words have all uppercase characters.

```
toupper(s)
```

```
## [1] "REGARDS, \n ETHAN"
```