Programming Assignment 4

Computing for Data Analysis

Introduction

This Programming Assignment will focus on regular expressions and the various regular expression functions implemented in R. The ability to handle and construct regular expressions is important when dealing with unstructured text data, often obtained from the Web or other online source.

First download the `Baltimore_homicides.zip` file from Courseplus and unzip it into your working directory. You can read the file `homicides.txt` with `readLines` via

```
> homicides <- readLines("homicides.txt")
```

The data were downloaded from the Baltimore Sun web site (`http://goo.gl/hSofH`) and contain information about homicides occurring between 2007 and the middle of 2012. There is one homicide record per line of text. The data contain various HTML and JavaScript markup and so are not easily read into R without some manipulation/processing.

1 How many of each cause of homicide?

The goal of this problem is to count the number of different types of homicides that are in this dataset. In each record there is a field with the word "Cause" in it indicating the cause of death (e.g. "Cause: shooting"). The basic goal is to extract this field and count the number of instances of each cause.

Write a function named `count` that takes one argument, a character string indicating the cause of death. The function should then return an integer representing the number of homicides from that cause in the dataset. If no cause of death is specified, then the function should return an error message via the `stop` function.

- Your function should read the homicides dataset in the manner indicated above.
- The options for cause of death are "asphyxiation", "blunt force", "other", "shooting", "stabbing", "unknown". No other causes are allowed. If a cause of death is specified that is not one of these, then the function should throw an error with the `stop` function.
- Note that some homicides in the dataset do not have a cause of death listed and those records should be ignored.
- Your function should deal with some irregularities in the dataset like capitalization. For example "Shooting" and "shooting" should be counted as the same cause of death.

- Do not worry about spelling errors in the dataset (records with spelling errors can be ignored)

The function should use the following template.

```
count <- function(cause = NULL) {
        ## Check that "cause" is non-NULL; else throw error
        ## Check that specific "cause" is allowed; else throw error

        ## Read "homicides.txt" data file

        ## Extract causes of death

        ## Return integer containing count of homicides for that cause
}
```

The function should execute as follows:

```
> count("other")

[1] 6

> num <- count("unknown")
> print(num)

[1] 10
```

Save your code for this function to a file named `count.R`.

Use the submit script provided to submit your solution to this part. There are 3 tests that need to be passed for this part of the assignment.

2 Ages of homicide victims

The goal of this part is to write a function called `agecount` that returns the number of homicide victims of a given age. For most (but not all) records there is an indication of the age of the victim. Your function should take one argument, the age of the victim(s), extract the age of the victim from each record and then return a count of the number of victims of the specified age.

- The argument passed to `agecount` should be a positive integer, but you do not need to check for this.
- If a record does not contain any age information, the record should be ignored.
- The function should return an integer indicating the number of victims of a given age.
- Your function should read the homicides dataset in the manner indicated above.

The function should use the following template.

```
agecount <- function(age = NULL) {
        ## Check that "age" is non-NULL; else throw error

        ## Read "homicides.txt" data file

        ## Extract ages of victims; ignore records where no age is
        ## given

        ## Return integer containing count of homicides for that age
}
```

The function should execute as follows:

```
> agecount(3)

[1] 0

> num <- agecount(21)
> print(num)

[1] 60
```

Save your code for this function to a file named `agecount.R`.

Use the submit script provided to submit your solution to this part. There are 2 tests that need to be passed for this part of the assignment.