DATA607 - Week 3 Assignment

Donald Butler

9/12/2021

```
library(tidyverse)
```

Exercise 1

Using the 173 majors listed in fivethirtyeight.com's College Majors dataset [https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/], provide code that identifies the majors that contain either "DATA" or "STATISTICS"

```
url = 'https://raw.githubusercontent.com/dab31415/DATA607/main/Homework/Assignment_3/majors-list.csv'
majors <- read_csv(url,show_col_types = FALSE)
majors %>%
  filter(grepl('DATA|STATISTICS',Major))
```

```
## # A tibble: 3 x 3
##   FOD1P Major                                          Major_Category
##   <chr> <chr>                                          <chr>
## 1 6212  MANAGEMENT INFORMATION SYSTEMS AND STATISTICS Business
## 2 2101  COMPUTER PROGRAMMING AND DATA PROCESSING      Computers & Mathematics
## 3 3702  STATISTICS AND DECISION SCIENCE               Computers & Mathematics
```

Exercise 2

Write code that transforms the data below:

[1] "bell pepper" "bilberry" "blackberry" "blood orange"

[5] "blueberry" "cantaloupe" "chili pepper" "cloudberry"

[9] "elderberry" "lime" "lychee" "mulberry"

[13] "olive" "salal berry"

Into a format like this:

c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe", "chili pepper", "cloudberry", "elderberry", "lime", "lychee", "mulberry", "olive", "salal berry")

```
raw_input <- '[1] "bell pepper"  "bilberry"     "blackberry"   "blood orange"
[5] "blueberry"    "cantaloupe"   "chili pepper" "cloudberry"
[9] "elderberry"   "lime"         "lychee"       "mulberry"
[13] "olive"        "salal berry"'

expected_result <- c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe"
```

Remove new line characters and positional indicators and trim whitespace.

```
new_input <- raw_input %>%
  str_replace_all('\\n','') %>%
  str_replace_all('(\\[\\d+\\])','') %>%
  str_trim()

new_input
```

```
## [1] "\"bell pepper\"  \"bilberry\"     \"blackberry\"   \"blood orange\" \"blueberry\"    \"cantaloup
```

Introduce delimiter by searching for whitespace between double quotes, then remove the leading and trailing double quote.

```
new_input <- new_input %>%
  str_replace_all('\\"[ ]*\\"',',') %>%
  str_replace('^"','') %>%
  str_replace('"$','')

new_input
```

```
## [1] "bell pepper,bilberry,blackberry,blood orange,blueberry,cantaloupe,chili pepper,cloudberry,elderb
```

Split the input string by the delimiter and convert to a vector.

```
new_input <- new_input %>%
  str_split(pattern = ',') %>%
  unlist()

new_input
```

```
##  [1] "bell pepper"  "bilberry"     "blackberry"   "blood orange" "blueberry"
##  [6] "cantaloupe"   "chili pepper" "cloudberry"   "elderberry"   "lime"
## [11] "lychee"       "mulberry"     "olive"        "salal berry"
```

Compare the converted input to our expected result.

```
identical(expected_result,new_input)
```

```
## [1] TRUE
```

Exercise 3

Describe, in words, what these expressions will match:

(.)\1\1
Will match any character repeated 3 consecutive times, for example 'aaa'.

(.)(.)\2\1
Will match any two characters that are repeated reversed, for example 'abba'.

(..)\1
Will match any two characters that are repeated, for example ‘mama’

(.).\1.\1
Will match any character repeated two times, but separated by a character in between each occurrence, for example ‘anana’ in ‘banana’,

(.)(.)(.).*\3\2\1
Will match any three characters that are later reversed with any number of characters between them. For example, ‘abccba’, ‘abcdddcba’

Exercise 4

Construct regular expressions to match words that:

Start and end with the same character.
ˆ(.).*\1$

Contain a repeated pair of letters (e.g. “church” contains “ch” repeated twice.)
(..).*\1

Contain one letter repeated in at least three places (e.g. “eleven” contains three “e”s.)
(.).*1.\1