Assignment03

Ted Kim

2022-09-16

Team Member

Seung Min Song

Load library

```
library(tidyverse)
```

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 <- paste0(
        'https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/',
        'majors-list.csv'
      )
df <- read.csv(url)
```

Find matched majors and display the results.

The pattern to be used in the regular expression is defined as '**DATA|STATISTICS**' and the match is returned using the str_detect() function. Only values with *true* results are filtered from the data frame and displayed.

```
##   FOD1P                                         Major          Major_Category
## 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
```

Find matched majors, create new column using mutate and display the results.

Using the lapply() and the str_detect() functions to store in a new field "is_include" whether the value in the major column of the data frame matches the pattern and display.

```
regex1 = 'DATA|STATISTICS'

df <- df %>%
        mutate(is_include = lapply(df$Major, function(str) {
                                   str_detect(str, regex1)
                                 }))

#The following error occurred when trying to sort the column created using the lapply()
#function.
#  unimplemented type 'list' in 'orderVector1'
#This is caused by the inclusion of a non-vector list in the data frame, which should be
#converted to classical format using as.data.frame.

df2 <- as.data.frame(lapply(df, unlist))

print(df2[order(-df2$is_include), c('Major', 'is_include')],
      row.names = FALSE, right = FALSE)
```

```
##  Major                                                                 is_include
##  MANAGEMENT INFORMATION SYSTEMS AND STATISTICS                          TRUE
##  COMPUTER PROGRAMMING AND DATA PROCESSING                               TRUE
##  STATISTICS AND DECISION SCIENCE                                        TRUE
##  GENERAL AGRICULTURE                                                   FALSE
##  AGRICULTURE PRODUCTION AND MANAGEMENT                                 FALSE
##  AGRICULTURAL ECONOMICS                                                FALSE
##  ANIMAL SCIENCES                                                       FALSE
##  FOOD SCIENCE                                                          FALSE
##  PLANT SCIENCE AND AGRONOMY                                            FALSE
##  SOIL SCIENCE                                                          FALSE
##  MISCELLANEOUS AGRICULTURE                                             FALSE
##  FORESTRY                                                              FALSE
##  NATURAL RESOURCES MANAGEMENT                                          FALSE
##  FINE ARTS                                                          FALSE
##  DRAMA AND THEATER ARTS                                             FALSE
##  MUSIC                                                              FALSE
##  VISUAL AND PERFORMING ARTS                                         FALSE
##  COMMERCIAL ART AND GRAPHIC DESIGN                                  FALSE
##  FILM VIDEO AND PHOTOGRAPHIC ARTS                                   FALSE
##  STUDIO ARTS                                                        FALSE
##  MISCELLANEOUS FINE ARTS                                            FALSE
##  ENVIRONMENTAL SCIENCE                                              FALSE
##  BIOLOGY                                                            FALSE
##  BIOCHEMICAL SCIENCES                                               FALSE
##  BOTANY                                                             FALSE
##  MOLECULAR BIOLOGY                                                  FALSE
##  ECOLOGY                                                            FALSE
##  GENETICS                                                           FALSE
##  MICROBIOLOGY                                                       FALSE
##  PHARMACOLOGY                                                       FALSE
##  PHYSIOLOGY                                                         FALSE
##  ZOOLOGY                                                            FALSE
##  NEUROSCIENCE                                                       FALSE
##  MISCELLANEOUS BIOLOGY                                              FALSE
##  COGNITIVE SCIENCE AND BIOPSYCHOLOGY                                FALSE
##  GENERAL BUSINESS                                                   FALSE
##  ACCOUNTING                                                         FALSE
##  ACTUARIAL SCIENCE                                                  FALSE
##  BUSINESS MANAGEMENT AND ADMINISTRATION                             FALSE
##  OPERATIONS LOGISTICS AND E-COMMERCE                                FALSE
##  BUSINESS ECONOMICS                                                 FALSE
##  MARKETING AND MARKETING RESEARCH                                   FALSE
##  FINANCE                                                            FALSE
##  HUMAN RESOURCES AND PERSONNEL MANAGEMENT                           FALSE
##  INTERNATIONAL BUSINESS                                             FALSE
##  HOSPITALITY MANAGEMENT                                             FALSE
##  MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION                    FALSE
##  COMMUNICATIONS                                                     FALSE
##  JOURNALISM                                                         FALSE
## MASS MEDIA                                                         FALSE
## ADVERTISING AND PUBLIC RELATIONS                                   FALSE
## COMMUNICATION TECHNOLOGIES                                         FALSE
## COMPUTER AND INFORMATION SYSTEMS                                   FALSE
## COMPUTER SCIENCE                                                   FALSE
## INFORMATION SCIENCES                                               FALSE
## COMPUTER ADMINISTRATION MANAGEMENT AND SECURITY                    FALSE
## COMPUTER NETWORKING AND TELECOMMUNICATIONS                         FALSE
## MATHEMATICS                                                        FALSE
## APPLIED MATHEMATICS                                                FALSE
## MATHEMATICS AND COMPUTER SCIENCE                                   FALSE
## GENERAL EDUCATION                                                  FALSE
## EDUCATIONAL ADMINISTRATION AND SUPERVISION                         FALSE
## SCHOOL STUDENT COUNSELING                                          FALSE
## ELEMENTARY EDUCATION                                               FALSE
## MATHEMATICS TEACHER EDUCATION                                      FALSE
## PHYSICAL AND HEALTH EDUCATION TEACHING                             FALSE
## EARLY CHILDHOOD EDUCATION                                          FALSE
## SCIENCE AND COMPUTER TEACHER EDUCATION                             FALSE
## SECONDARY TEACHER EDUCATION                                        FALSE
## SPECIAL NEEDS EDUCATION                                            FALSE
## SOCIAL SCIENCE OR HISTORY TEACHER EDUCATION                        FALSE
## TEACHER EDUCATION: MULTIPLE LEVELS                                 FALSE
## LANGUAGE AND DRAMA EDUCATION                                       FALSE
## ART AND MUSIC EDUCATION                                            FALSE
## MISCELLANEOUS EDUCATION                                            FALSE
## LIBRARY SCIENCE                                                    FALSE
## ARCHITECTURE                                                       FALSE
## GENERAL ENGINEERING                                                FALSE
## AEROSPACE ENGINEERING                                              FALSE
## BIOLOGICAL ENGINEERING                                             FALSE
## ARCHITECTURAL ENGINEERING                                          FALSE
## BIOMEDICAL ENGINEERING                                             FALSE
## CHEMICAL ENGINEERING                                               FALSE
## CIVIL ENGINEERING                                                  FALSE
## COMPUTER ENGINEERING                                               FALSE
## ELECTRICAL ENGINEERING                                           FALSE
## ENGINEERING MECHANICS PHYSICS AND SCIENCE                        FALSE
## ENVIRONMENTAL ENGINEERING                                        FALSE
## GEOLOGICAL AND GEOPHYSICAL ENGINEERING                           FALSE
## INDUSTRIAL AND MANUFACTURING ENGINEERING                         FALSE
## MATERIALS ENGINEERING AND MATERIALS SCIENCE                      FALSE
## MECHANICAL ENGINEERING                                           FALSE
## METALLURGICAL ENGINEERING                                        FALSE
## MINING AND MINERAL ENGINEERING                                   FALSE
## NAVAL ARCHITECTURE AND MARINE ENGINEERING                        FALSE
## NUCLEAR ENGINEERING                                              FALSE
## PETROLEUM ENGINEERING                                            FALSE
## MISCELLANEOUS ENGINEERING                                        FALSE
## ENGINEERING TECHNOLOGIES                                         FALSE
## ENGINEERING AND INDUSTRIAL MANAGEMENT                            FALSE
## ELECTRICAL ENGINEERING TECHNOLOGY                                FALSE
## INDUSTRIAL PRODUCTION TECHNOLOGIES                               FALSE
## MECHANICAL ENGINEERING RELATED TECHNOLOGIES                      FALSE
## MISCELLANEOUS ENGINEERING TECHNOLOGIES                           FALSE
## MATERIALS SCIENCE                                                FALSE
## NUTRITION SCIENCES                                               FALSE
## GENERAL MEDICAL AND HEALTH SERVICES                              FALSE
## COMMUNICATION DISORDERS SCIENCES AND SERVICES                    FALSE
## HEALTH AND MEDICAL ADMINISTRATIVE SERVICES                       FALSE
## MEDICAL ASSISTING SERVICES                                       FALSE
## MEDICAL TECHNOLOGIES TECHNICIANS                                 FALSE
## HEALTH AND MEDICAL PREPARATORY PROGRAMS                          FALSE
## NURSING                                                          FALSE
## PHARMACY PHARMACEUTICAL SCIENCES AND ADMINISTRATION              FALSE
## TREATMENT THERAPY PROFESSIONS                                    FALSE
## COMMUNITY AND PUBLIC HEALTH                                      FALSE
## MISCELLANEOUS HEALTH MEDICAL PROFESSIONS                         FALSE
## AREA ETHNIC AND CIVILIZATION STUDIES                             FALSE
## LINGUISTICS AND COMPARATIVE LANGUAGE AND LITERATURE              FALSE
## FRENCH GERMAN LATIN AND OTHER COMMON FOREIGN LANGUAGE STUDIES    FALSE
## OTHER FOREIGN LANGUAGES                                          FALSE
##  ENGLISH LANGUAGE AND LITERATURE                                  FALSE
##  COMPOSITION AND RHETORIC                                         FALSE
##  LIBERAL ARTS                                                     FALSE
##  HUMANITIES                                                       FALSE
##  INTERCULTURAL AND INTERNATIONAL STUDIES                          FALSE
##  PHILOSOPHY AND RELIGIOUS STUDIES                                 FALSE
##  THEOLOGY AND RELIGIOUS VOCATIONS                                 FALSE
##  ANTHROPOLOGY AND ARCHEOLOGY                                      FALSE
##  ART HISTORY AND CRITICISM                                        FALSE
##  HISTORY                                                          FALSE
##  UNITED STATES HISTORY                                            FALSE
##  COSMETOLOGY SERVICES AND CULINARY ARTS                           FALSE
##  FAMILY AND CONSUMER SCIENCES                                     FALSE
##  MILITARY TECHNOLOGIES                                            FALSE
##  PHYSICAL FITNESS PARKS RECREATION AND LEISURE                    FALSE
##  CONSTRUCTION SERVICES                                            FALSE
##  ELECTRICAL, MECHANICAL, AND PRECISION TECHNOLOGIES AND PRODUCTION FALSE
##  TRANSPORTATION SCIENCES AND TECHNOLOGIES                         FALSE
##  MULTI/INTERDISCIPLINARY STUDIES                                  FALSE
##  COURT REPORTING                                                  FALSE
##  PRE-LAW AND LEGAL STUDIES                                        FALSE
##  CRIMINAL JUSTICE AND FIRE PROTECTION                             FALSE
##  PUBLIC ADMINISTRATION                                            FALSE
##  PUBLIC POLICY                                                    FALSE
##  N/A (less than bachelor's degree)                                FALSE
##  PHYSICAL SCIENCES                                                FALSE
##  ASTRONOMY AND ASTROPHYSICS                                       FALSE
##  ATMOSPHERIC SCIENCES AND METEOROLOGY                             FALSE
##  CHEMISTRY                                                        FALSE
##  GEOLOGY AND EARTH SCIENCE                                        FALSE
##  GEOSCIENCES                                                      FALSE
##  OCEANOGRAPHY                                                     FALSE
##  PHYSICS                                                          FALSE
##  MULTI-DISCIPLINARY OR GENERAL SCIENCE                            FALSE
##  NUCLEAR, INDUSTRIAL RADIOLOGY, AND BIOLOGICAL TECHNOLOGIES       FALSE
##  PSYCHOLOGY                                                       FALSE
##  EDUCATIONAL PSYCHOLOGY                                             FALSE
##  CLINICAL PSYCHOLOGY                                                FALSE
##  COUNSELING PSYCHOLOGY                                              FALSE
##  INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY                           FALSE
##  SOCIAL PSYCHOLOGY                                                  FALSE
##  MISCELLANEOUS PSYCHOLOGY                                           FALSE
##  HUMAN SERVICES AND COMMUNITY ORGANIZATION                          FALSE
##  SOCIAL WORK                                                        FALSE
##  INTERDISCIPLINARY SOCIAL SCIENCES                                  FALSE
##  GENERAL SOCIAL SCIENCES                                            FALSE
##  ECONOMICS                                                          FALSE
##  CRIMINOLOGY                                                        FALSE
##  GEOGRAPHY                                                          FALSE
##  INTERNATIONAL RELATIONS                                            FALSE
##  POLITICAL SCIENCE AND GOVERNMENT                                   FALSE
##  SOCIOLOGY                                                          FALSE
##  MISCELLANEOUS SOCIAL SCIENCES                                      FALSE
```

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”)

Define strings

```
str <- paste0('[1] "bell pepper"  "bilberry"     "blackberry"   "blood orange"',
              '[5] "blueberry"    "cantaloupe"   "chili pepper" "cloudberry"',
              '[9] "elderberry"   "lime"         "lychee"       "mulberry"',
              '[13] "olive"        "salal berry"')
```

Remove the most unnecessary characters first

[step 1] remove [:number]:whitespace and repeated :whitespaces

```
str1 <- str_replace_all(str, "\\[\\d+\\]\\s|\\s{2,}", "")
cat(str1)
```

```
## "bell pepper""bilberry""blackberry""blood orange""blueberry""cantaloupe""chili pepper" "cloudberry""
```

[step 2] replace ““(no space) or” “(include space) to”, “

```
str1 <- str_replace_all(str1, '\\"\\s?\\"', '\\", \\"')
cat(str1)
```

```
## "bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe", "chili pepper",
```

[step 3] replace start of strings(line) ” to c(“

```
str1 <- str_replace_all(str1, '^\\"', 'c(\\"')
cat(str1)
```

```
## c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe", "chili pepper"
```

[step 4] replace end of strings(line) ” to “)

```
str1 <- str_replace_all(str1, '\\"$', '\\")')
cat(str1)
```

```
## c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe", "chili pepper"
```

[Wrap up]

wrap up step1 thru step4 above by combining them into one command

```r
cat(str_replace_all(str, '\\[\\d+\\]\\s|\\s{2,}', '') %>%
      str_replace_all('\\"\\s?\\"', '\\", \\"') %>%
      str_replace_all('^\\"', 'c(\\"') %>%
      str_replace_all('\\"$', '\\")'))
```

```
## c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe", "chili pepper"
```

Replace string from left to right

[step 1] replace start of strings(line) [:number]:whitespace to c(

```r
str2 <- str_replace_all(str, '^\\[\\d+\\]\\s', 'c(')
cat(str2)
```

```
## c("bell pepper"  "bilberry"     "blackberry"   "blood orange"[5]  "blueberry"    "cantaloupe"   "chil
```

[step 2] replace white space(s) between double quotes such as " " to ", "

```r
str2 <- str_replace_all(str2, '\\"\\s+\\"', '\\", \\"')
cat(str2)
```

```
## c("bell pepper", "bilberry", "blackberry", "blood orange"[5] "blueberry", "cantaloupe", "chili peppe
```

[step 3] replace middle of strings(line) [:number]:whitespace to ',' (comma & whitespace)

```r
str2 <- str_replace_all(str2, '\\[\\d+\\]\\s+', ', ')
cat(str2)
```

```
## c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe", "chili pepper"
```

[step 4] replace end of strings(line) " to ")

```r
str2 <- str_replace_all(str2, '\\"$', '\\")')
cat(str2)
```

```
## c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe", "chili pepper"
```

[Wrap up]

wrap up step1 thru step4 above by combining them into one command

```
cat(str_replace_all(str, '^\\[\\d+\\]\\s', 'c(') %>%
      str_replace_all('\\"\\s+\\"', '\\", \\"') %>%
      str_replace_all('\\[\\d+\\]\\s+', ', ') %>%
      str_replace_all('\\"$', '\\")'))
```

```
## c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloupe", "chili pepper"
```

String to Array

Use the str_split() function to make a string array

```
str1 <- str_replace_all(str, '\\[\\d+\\]\\s|\\s{2,}', '') %>%
       str_replace_all('\\"\\s?\\"', '\\", \\"')

print(str_split(str_replace_all(str1, '\\"', ''), ', ')[[1]])
```

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

3 Describe, in words, what these expressions will match:

```
## (.)\1\1              : a capturing group of any character repeats three times in a row

## (.)(.)\2\1           : Two capturing groups consisting of one character each and the next content of ca
##                        connected by the reverse order. Four letters are palindrome.

## (..)\1               : a capturing group of any two-characters repeats two times

## (.).\1.\1            : A capturing group of any character is repeated three times.
##                       First, third, and fifth character shold be same, but second, and
##                       forth can be any other character. Furthermore, all five can be the
##                       same character.
```

```
## (.)(.)(.).*\3\2\1 : the first three any characters(capturing groups) and the last three characters a
```

4 Construct regular expressions to match words that:

Define test strings

```
arr <- c('church', 'buddy', 'tomato', 'eleven', 'bahama',
         '12345612', '1234', 'seventeen', 'mom')
```

Start and end with the same character.

```
# ^: start of string(line)
# $: end of string(line)
# .: any character except line break
# *: zero or more times
# (): capturing group
# \\1: contents of group 1
regex4_1 = '^(.).*\\1$'
str_detect(arr, regex4_1)
```

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

Contain a repeated pair of letters (e.g. "church" contains "ch" repeated twice.)

```
# Start and end with the same (allow letter only)
# ^: start of string(line)
# $: end of string(line)
# [a-zA-Z]: only letter
# *: zero or more times
# (): capturing group
# \\1: contents of group 1
# {2}: exactly two times
regex4_2_1 = '^([a-zA-Z]{2})[a-zA-Z]*\\1$'
```

```r
str_detect(arr, regex4_2_1)
```

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

```r
# any position (allow letter only)
regex4_2_2 = '([a-zA-Z]{2})[a-zA-Z]*\\1'
str_detect(arr, regex4_2_2)
```

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

```r
# Start and end with the same (allow any character)
regex4_2_3 = '^(.{2}).*\\1$'
str_detect(arr, regex4_2_3)
```

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

```r
# any position (allow any character)
regex4_2_4 = '(.{2}).*\\1'
str_detect(arr, regex4_2_4)
```

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

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

```r
# *: zero or more times
# (): capturing group#
# .: any character except line break
# \\1: contents of group 1
# {2}: exactly two times
regex4_3 = '(.).*\\1.*\\1'
str_detect(arr, regex4_3)
```

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

- GitHub - https://github.com/blacksmilez/DATA607/tree/main/Assignment03
- RPubs - https://rpubs.com/blacksmilez/