Assignmnet 3 Seung Min Song

2022-09-18

Contents

Team member: Ted Kim
GitHub: https://github.com/seung-m1nsong/607
rpub: https://rpubs.com/seungm1nsong/943748

Question 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"

```
##   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
```

```
##   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
```

```
regex1 = 'DATA|STATISTICS'

df <- dfmajordata %>%
        mutate(is_include = lapply(dfmajordata$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
```

Question 2

Write code that transforms the data below:

[1] "bell pepper" "bilberry" "blackberry" "blood orange" [5] "blueberry" "cantaloupe" "chili pepper" "cloud-berry"
[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")

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

```
## [1] "bell pepper"  "bilberry"     "blackberry"   "blood orange"[5] "blueberry"    "cantaloupe"   "chi
```

```r
#step 1. remove '[#] 'and multiple white space
str <- str_replace_all(str, '\\[\\d+?\\]\\s|\\s{2,}', '')
cat(str)
```

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

```r
#step 2. replace '""' to '", "'
str <- str_replace_all(str, '\\"\\s?\\"', '\\", \\"')
cat(str)
```

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

```r
forVec <- str

#step 3. replace '"' start of strings(line) to 'c("'
str <- str_replace_all(str, '^\\"', 'c(\\"')
cat(str)
```

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

```r
#step 4. replace '"' end of strings(line) to '")'
str <- str_replace_all(str, '\\"$', '\\")')
cat(str)
```

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

```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"
```

```r
#extra) string to vector
print(str_split(str_replace_all(forVec, '\\"', ''), ', ', simplify = FALSE)[[1]][2])
```

```
## [1] "bilberry"
```

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

```
## [1] "bell pepper"  "bilberry"     "blackberry"   "blood orange"[5] "blueberry"    "cantaloupe"   "ch
```

```r
#step 1. replace '[#] ' start of strings(line) to 'c('
str <- str_replace_all(str, '^\\[\\d+?\\]\\s', 'c(')
cat(str)
```

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

```r
#step 2. replace '"   "' to '", "'
str <- str_replace_all(str, '\\"\\s\\"', '\\", \\"')
cat(str)
```

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

```r
#step 3. replace '[#] ' in the middle of strings(line) to ', '
str <- str_replace_all(str, '\\[\\d+?\\]\\s', ', ')
cat(str)
```

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

```r
#step 4. replace '"' end of strings(line) to '")'
str <- str_replace_all(str, '\\"$', '\\")')
cat(str)
```

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

```r
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"
```

Question 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 capturing are connected by the reverse order. Four letters are palindrome.
- (..)\1 : A group of two-characters repeats two times.
- "(.).\1.\1": A group of a syllable(character) is repeated three times. The first, third, and fifth characters should be the same, but the second and fourth can be any other character. Furthermore, all five can be the same character.
- "(.)(.)(.).*\3\2\1": The first three letters and the last three letters as a palindrome.

Question 4

Construct regular expressions to match words that:

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

- Start and end with the same character.

```r
# ^: 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.)

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