DATA607_Assignment_3

PeiMing Chen

2023-02-12

Introduction The assignment of this week (week 3) is given to us to practice on regular expressions. Three libraries below were downloaded to solve the four problems.

```
library(tidyverse)
library(openintro)
library(dplyr)
```

Codes that identify the majors that contain either "DATA" or "STATISTICS"

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"

```
UMajors <- read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/majors
UMajors
```

```
##     FOD1P                                                Major
## 1    1100                                GENERAL AGRICULTURE
## 2    1101                AGRICULTURE PRODUCTION AND MANAGEMENT
## 3    1102                            AGRICULTURAL ECONOMICS
## 4    1103                                   ANIMAL SCIENCES
## 5    1104                                      FOOD SCIENCE
## 6    1105                          PLANT SCIENCE AND AGRONOMY
## 7    1106                                      SOIL SCIENCE
## 8    1199                           MISCELLANEOUS AGRICULTURE
## 9    1302                                         FORESTRY
## 10   1303                       NATURAL RESOURCES MANAGEMENT
## 11   6000                                       FINE ARTS
## 12   6001                           DRAMA AND THEATER ARTS
## 13   6002                                           MUSIC
## 14   6003                         VISUAL AND PERFORMING ARTS
## 15   6004                  COMMERCIAL ART AND GRAPHIC DESIGN
## 16   6005                   FILM VIDEO AND PHOTOGRAPHIC ARTS
## 17   6007                                   STUDIO ARTS
## 18   6099                          MISCELLANEOUS FINE ARTS
## 19   1301                            ENVIRONMENTAL SCIENCE
## 20   3600                                          BIOLOGY
## 21   3601                            BIOCHEMICAL SCIENCES
## 22   3602                                          BOTANY
## 23   3603                               MOLECULAR BIOLOGY
```

```
## 24   3604                                                    ECOLOGY
## 25   3605                                                   GENETICS
## 26   3606                                              MICROBIOLOGY
## 27   3607                                              PHARMACOLOGY
## 28   3608                                                PHYSIOLOGY
## 29   3609                                                   ZOOLOGY
## 30   3611                                              NEUROSCIENCE
## 31   3699                                     MISCELLANEOUS BIOLOGY
## 32   4006                       COGNITIVE SCIENCE AND BIOPSYCHOLOGY
## 33   6200                                        GENERAL BUSINESS
## 34   6201                                              ACCOUNTING
## 35   6202                                         ACTUARIAL SCIENCE
## 36   6203                  BUSINESS MANAGEMENT AND ADMINISTRATION
## 37   6204                     OPERATIONS LOGISTICS AND E-COMMERCE
## 38   6205                                     BUSINESS ECONOMICS
## 39   6206                       MARKETING AND MARKETING RESEARCH
## 40   6207                                              FINANCE
## 41   6209             HUMAN RESOURCES AND PERSONNEL MANAGEMENT
## 42   6210                                 INTERNATIONAL BUSINESS
## 43   6211                                 HOSPITALITY MANAGEMENT
## 44   6212           MANAGEMENT INFORMATION SYSTEMS AND STATISTICS
## 45   6299         MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION
## 46   1901                                       COMMUNICATIONS
## 47   1902                                          JOURNALISM
## 48   1903                                          MASS MEDIA
## 49   1904                       ADVERTISING AND PUBLIC RELATIONS
## 50   2001                            COMMUNICATION TECHNOLOGIES
## 51   2100                       COMPUTER AND INFORMATION SYSTEMS
## 52   2101             COMPUTER PROGRAMMING AND DATA PROCESSING
## 53   2102                                        COMPUTER SCIENCE
## 54   2105                                   INFORMATION SCIENCES
## 55   2106         COMPUTER ADMINISTRATION MANAGEMENT AND SECURITY
## 56   2107           COMPUTER NETWORKING AND TELECOMMUNICATIONS
## 57   3700                                            MATHEMATICS
## 58   3701                                     APPLIED MATHEMATICS
## 59   3702                       STATISTICS AND DECISION SCIENCE
## 60   4005                       MATHEMATICS AND COMPUTER SCIENCE
## 61   2300                                     GENERAL EDUCATION
## 62   2301           EDUCATIONAL ADMINISTRATION AND SUPERVISION
## 63   2303                            SCHOOL STUDENT COUNSELING
## 64   2304                                 ELEMENTARY EDUCATION
## 65   2305                          MATHEMATICS TEACHER EDUCATION
## 66   2306                PHYSICAL AND HEALTH EDUCATION TEACHING
## 67   2307                               EARLY CHILDHOOD EDUCATION
## 68   2308                SCIENCE AND COMPUTER TEACHER EDUCATION
## 69   2309                           SECONDARY TEACHER EDUCATION
## 70   2310                              SPECIAL NEEDS EDUCATION
## 71   2311           SOCIAL SCIENCE OR HISTORY TEACHER EDUCATION
## 72   2312                    TEACHER EDUCATION: MULTIPLE LEVELS
## 73   2313                         LANGUAGE AND DRAMA EDUCATION
## 74   2314                              ART AND MUSIC EDUCATION
## 75   2399                              MISCELLANEOUS EDUCATION
## 76   3501                                      LIBRARY SCIENCE
## 77   1401                                         ARCHITECTURE
```

```
## 78   2400                                                GENERAL ENGINEERING
## 79   2401                                              AEROSPACE ENGINEERING
## 80   2402                                             BIOLOGICAL ENGINEERING
## 81   2403                                          ARCHITECTURAL ENGINEERING
## 82   2404                                             BIOMEDICAL ENGINEERING
## 83   2405                                               CHEMICAL ENGINEERING
## 84   2406                                                  CIVIL ENGINEERING
## 85   2407                                               COMPUTER ENGINEERING
## 86   2408                                              ELECTRICAL ENGINEERING
## 87   2409                        ENGINEERING MECHANICS PHYSICS AND SCIENCE
## 88   2410                                          ENVIRONMENTAL ENGINEERING
## 89   2411                           GEOLOGICAL AND GEOPHYSICAL ENGINEERING
## 90   2412                         INDUSTRIAL AND MANUFACTURING ENGINEERING
## 91   2413                       MATERIALS ENGINEERING AND MATERIALS SCIENCE
## 92   2414                                            MECHANICAL ENGINEERING
## 93   2415                                          METALLURGICAL ENGINEERING
## 94   2416                                   MINING AND MINERAL ENGINEERING
## 95   2417                        NAVAL ARCHITECTURE AND MARINE ENGINEERING
## 96   2418                                             NUCLEAR ENGINEERING
## 97   2419                                           PETROLEUM ENGINEERING
## 98   2499                                          MISCELLANEOUS ENGINEERING
## 99   2500                                           ENGINEERING TECHNOLOGIES
## 100  2501                          ENGINEERING AND INDUSTRIAL MANAGEMENT
## 101  2502                              ELECTRICAL ENGINEERING TECHNOLOGY
## 102  2503                            INDUSTRIAL PRODUCTION TECHNOLOGIES
## 103  2504                       MECHANICAL ENGINEERING RELATED TECHNOLOGIES
## 104  2599                            MISCELLANEOUS ENGINEERING TECHNOLOGIES
## 105  5008                                                MATERIALS SCIENCE
## 106  4002                                              NUTRITION SCIENCES
## 107  6100                          GENERAL MEDICAL AND HEALTH SERVICES
## 108  6102                    COMMUNICATION DISORDERS SCIENCES AND SERVICES
## 109  6103                       HEALTH AND MEDICAL ADMINISTRATIVE SERVICES
## 110  6104                                    MEDICAL ASSISTING SERVICES
## 111  6105                                 MEDICAL TECHNOLOGIES TECHNICIANS
## 112  6106                          HEALTH AND MEDICAL PREPARATORY PROGRAMS
## 113  6107                                                       NURSING
## 114  6108                 PHARMACY PHARMACEUTICAL SCIENCES AND ADMINISTRATION
## 115  6109                                    TREATMENT THERAPY PROFESSIONS
## 116  6110                                      COMMUNITY AND PUBLIC HEALTH
## 117  6199                         MISCELLANEOUS HEALTH MEDICAL PROFESSIONS
## 118  1501                            AREA ETHNIC AND CIVILIZATION STUDIES
## 119  2601                 LINGUISTICS AND COMPARATIVE LANGUAGE AND LITERATURE
## 120  2602     FRENCH GERMAN LATIN AND OTHER COMMON FOREIGN LANGUAGE STUDIES
## 121  2603                                          OTHER FOREIGN LANGUAGES
## 122  3301                                ENGLISH LANGUAGE AND LITERATURE
## 123  3302                                       COMPOSITION AND RHETORIC
## 124  3401                                                   LIBERAL ARTS
## 125  3402                                                     HUMANITIES
## 126  4001                          INTERCULTURAL AND INTERNATIONAL STUDIES
## 127  4801                                 PHILOSOPHY AND RELIGIOUS STUDIES
## 128  4901                                 THEOLOGY AND RELIGIOUS VOCATIONS
## 129  5502                                      ANTHROPOLOGY AND ARCHEOLOGY
## 130  6006                                        ART HISTORY AND CRITICISM
## 131  6402                                                         HISTORY
```

```
## 132  6403                                                UNITED STATES HISTORY
## 133  2201                                COSMETOLOGY SERVICES AND CULINARY ARTS
## 134  2901                                         FAMILY AND CONSUMER SCIENCES
## 135  3801                                                MILITARY TECHNOLOGIES
## 136  4101                          PHYSICAL FITNESS PARKS RECREATION AND LEISURE
## 137  5601                                                CONSTRUCTION SERVICES
## 138  5701 ELECTRICAL, MECHANICAL, AND PRECISION TECHNOLOGIES AND PRODUCTION
## 139  5901                              TRANSPORTATION SCIENCES AND TECHNOLOGIES
## 140  4000                                     MULTI/INTERDISCIPLINARY STUDIES
## 141  3201                                                   COURT REPORTING
## 142  3202                                         PRE-LAW AND LEGAL STUDIES
## 143  5301                                  CRIMINAL JUSTICE AND FIRE PROTECTION
## 144  5401                                                PUBLIC ADMINISTRATION
## 145  5402                                                     PUBLIC POLICY
## 146 bbbb                                    N/A (less than bachelor's degree)
## 147  5000                                                 PHYSICAL SCIENCES
## 148  5001                                           ASTRONOMY AND ASTROPHYSICS
## 149  5002                                  ATMOSPHERIC SCIENCES AND METEOROLOGY
## 150  5003                                                        CHEMISTRY
## 151  5004                                          GEOLOGY AND EARTH SCIENCE
## 152  5005                                                      GEOSCIENCES
## 153  5006                                                     OCEANOGRAPHY
## 154  5007                                                          PHYSICS
## 155  5098                                MULTI-DISCIPLINARY OR GENERAL SCIENCE
## 156  5102        NUCLEAR, INDUSTRIAL RADIOLOGY, AND BIOLOGICAL TECHNOLOGIES
## 157  5200                                                       PSYCHOLOGY
## 158  5201                                             EDUCATIONAL PSYCHOLOGY
## 159  5202                                                CLINICAL PSYCHOLOGY
## 160  5203                                             COUNSELING PSYCHOLOGY
## 161  5205                           INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY
## 162  5206                                                SOCIAL PSYCHOLOGY
## 163  5299                                          MISCELLANEOUS PSYCHOLOGY
## 164  5403                          HUMAN SERVICES AND COMMUNITY ORGANIZATION
## 165  5404                                                      SOCIAL WORK
## 166  4007                                   INTERDISCIPLINARY SOCIAL SCIENCES
## 167  5500                                            GENERAL SOCIAL SCIENCES
## 168  5501                                                         ECONOMICS
## 169  5503                                                     CRIMINOLOGY
## 170  5504                                                       GEOGRAPHY
## 171  5505                                              INTERNATIONAL RELATIONS
## 172  5506                                    POLITICAL SCIENCE AND GOVERNMENT
## 173  5507                                                      SOCIOLOGY
## 174  5599                                      MISCELLANEOUS SOCIAL SCIENCES
##                          Major_Category
## 1       Agriculture & Natural Resources
## 2       Agriculture & Natural Resources
## 3       Agriculture & Natural Resources
## 4       Agriculture & Natural Resources
## 5       Agriculture & Natural Resources
## 6       Agriculture & Natural Resources
## 7       Agriculture & Natural Resources
## 8       Agriculture & Natural Resources
## 9       Agriculture & Natural Resources
## 10      Agriculture & Natural Resources
```

```
## 11                            Arts
## 12                            Arts
## 13                            Arts
## 14                            Arts
## 15                            Arts
## 16                            Arts
## 17                            Arts
## 18                            Arts
## 19              Biology & Life Science
## 20              Biology & Life Science
## 21              Biology & Life Science
## 22              Biology & Life Science
## 23              Biology & Life Science
## 24              Biology & Life Science
## 25              Biology & Life Science
## 26              Biology & Life Science
## 27              Biology & Life Science
## 28              Biology & Life Science
## 29              Biology & Life Science
## 30              Biology & Life Science
## 31              Biology & Life Science
## 32              Biology & Life Science
## 33                           Business
## 34                           Business
## 35                           Business
## 36                           Business
## 37                           Business
## 38                           Business
## 39                           Business
## 40                           Business
## 41                           Business
## 42                           Business
## 43                           Business
## 44                           Business
## 45                           Business
## 46          Communications & Journalism
## 47          Communications & Journalism
## 48          Communications & Journalism
## 49          Communications & Journalism
## 50              Computers & Mathematics
## 51              Computers & Mathematics
## 52              Computers & Mathematics
## 53              Computers & Mathematics
## 54              Computers & Mathematics
## 55              Computers & Mathematics
## 56              Computers & Mathematics
## 57              Computers & Mathematics
## 58              Computers & Mathematics
## 59              Computers & Mathematics
## 60              Computers & Mathematics
## 61                           Education
## 62                           Education
## 63                           Education
## 64                           Education
```

```
## 65                          Education
## 66                          Education
## 67                          Education
## 68                          Education
## 69                          Education
## 70                          Education
## 71                          Education
## 72                          Education
## 73                          Education
## 74                          Education
## 75                          Education
## 76                          Education
## 77                        Engineering
## 78                        Engineering
## 79                        Engineering
## 80                        Engineering
## 81                        Engineering
## 82                        Engineering
## 83                        Engineering
## 84                        Engineering
## 85                        Engineering
## 86                        Engineering
## 87                        Engineering
## 88                        Engineering
## 89                        Engineering
## 90                        Engineering
## 91                        Engineering
## 92                        Engineering
## 93                        Engineering
## 94                        Engineering
## 95                        Engineering
## 96                        Engineering
## 97                        Engineering
## 98                        Engineering
## 99                        Engineering
## 100                       Engineering
## 101                       Engineering
## 102                       Engineering
## 103                       Engineering
## 104                       Engineering
## 105                       Engineering
## 106                            Health
## 107                            Health
## 108                            Health
## 109                            Health
## 110                            Health
## 111                            Health
## 112                            Health
## 113                            Health
## 114                            Health
## 115                            Health
## 116                            Health
## 117                            Health
## 118          Humanities & Liberal Arts
```

```
## 119           Humanities & Liberal Arts
## 120           Humanities & Liberal Arts
## 121           Humanities & Liberal Arts
## 122           Humanities & Liberal Arts
## 123           Humanities & Liberal Arts
## 124           Humanities & Liberal Arts
## 125           Humanities & Liberal Arts
## 126           Humanities & Liberal Arts
## 127           Humanities & Liberal Arts
## 128           Humanities & Liberal Arts
## 129           Humanities & Liberal Arts
## 130           Humanities & Liberal Arts
## 131           Humanities & Liberal Arts
## 132           Humanities & Liberal Arts
## 133 Industrial Arts & Consumer Services
## 134 Industrial Arts & Consumer Services
## 135 Industrial Arts & Consumer Services
## 136 Industrial Arts & Consumer Services
## 137 Industrial Arts & Consumer Services
## 138 Industrial Arts & Consumer Services
## 139 Industrial Arts & Consumer Services
## 140                   Interdisciplinary
## 141                Law & Public Policy
## 142                Law & Public Policy
## 143                Law & Public Policy
## 144                Law & Public Policy
## 145                Law & Public Policy
## 146                                <NA>
## 147                  Physical Sciences
## 148                  Physical Sciences
## 149                  Physical Sciences
## 150                  Physical Sciences
## 151                  Physical Sciences
## 152                  Physical Sciences
## 153                  Physical Sciences
## 154                  Physical Sciences
## 155                  Physical Sciences
## 156                  Physical Sciences
## 157           Psychology & Social Work
## 158           Psychology & Social Work
## 159           Psychology & Social Work
## 160           Psychology & Social Work
## 161           Psychology & Social Work
## 162           Psychology & Social Work
## 163           Psychology & Social Work
## 164           Psychology & Social Work
## 165           Psychology & Social Work
## 166                     Social Science
## 167                     Social Science
## 168                     Social Science
## 169                     Social Science
## 170                     Social Science
## 171                     Social Science
## 172                     Social Science
```

```
## 173                       Social Science
## 174                       Social Science
```

1

select column of majors

```
UMajors$Major %>% str_subset(pattern = "DATA")
```

```
## [1] "COMPUTER PROGRAMMING AND DATA PROCESSING"
```

```
UMajors$Major %>% str_subset(pattern = "STATISTICS")
```

```
## [1] "MANAGEMENT INFORMATION SYSTEMS AND STATISTICS"
## [2] "STATISTICS AND DECISION SCIENCE"
```

2

```
DataSet <- data.frame(c("bell pepper", "bilberry", "blackberry", "blood orange", "blueberry", "cantaloup
cat(paste(DataSet))
```

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

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

Examples are run below for each expression discussed. (.)\1\1 The regular expression is to match a pattern in a string that has character repeated in it. For example, in a string like "888kjpf", it match 888

"(.)(.)\2\1" This is to match any character has itself and also followed by its reverse order.

(..)\1: This does not match anything. Instead," (..)\1 "will match any two characters that repeated.

"(.).\1.\1": This expression is to match a character followed by any other character and then followed by the original character again.

"(.)(.)(.).*\3\2\1" : This expression is to match the first three letters and the last three letters that are the same letters with reverse order. The characters in between the first and the last thress characters could be any character in any orders.

```
Test <- c("bluesky", "10000","TTT","hihihi","YaYaYu77","AOOA","QQQQQQ","Alico","Apia","Aka","aacBytfrhh

str_match(Test,"(.)\1\1")
```

```
##       [,1] [,2]
##  [1,] NA   NA
##  [2,] NA   NA
##  [3,] NA   NA
##  [4,] NA   NA
##  [5,] NA   NA
##  [6,] NA   NA
```

```
##  [7,] NA   NA
##  [8,] NA   NA
##  [9,] NA   NA
## [10,] NA   NA
## [11,] NA   NA
## [12,] NA   NA
## [13,] NA   NA
```

```
str_match(Test,"(.)(.)\\2\\1")
```

```
##       [,1]   [,2] [,3]
##  [1,] NA     NA   NA
##  [2,] "0000" "0"  "0"
##  [3,] NA     NA   NA
##  [4,] NA     NA   NA
##  [5,] NA     NA   NA
##  [6,] "AOOA" "A"  "O"
##  [7,] "QQQQ" "Q"  "Q"
##  [8,] NA     NA   NA
##  [9,] NA     NA   NA
## [10,] NA     NA   NA
## [11,] NA     NA   NA
## [12,] NA     NA   NA
## [13,] NA     NA   NA
```

```
str_match(Test,"(..)\\1")
```

```
##       [,1]   [,2]
##  [1,] NA     NA
##  [2,] "0000" "00"
##  [3,] NA     NA
##  [4,] "hihi" "hi"
##  [5,] "YaYa" "Ya"
##  [6,] NA     NA
##  [7,] "QQQQ" "QQ"
##  [8,] NA     NA
##  [9,] NA     NA
## [10,] NA     NA
## [11,] NA     NA
## [12,] NA     NA
## [13,] NA     NA
```

```
str_match(Test,"(.).\\1.\\1")
```

```
##       [,1]    [,2]
##  [1,] NA      NA
##  [2,] NA      NA
##  [3,] NA      NA
##  [4,] "hihih" "h"
##  [5,] "YaYaY" "Y"
##  [6,] NA      NA
##  [7,] "QQQQQ" "Q"
```

```
##  [8,] NA      NA
##  [9,] NA      NA
## [10,] NA      NA
## [11,] NA      NA
## [12,] NA      NA
## [13,] NA      NA
```

```
str_match(Test,"(.)(.)(.).*\\3\\2\\1")
```

```
##       [,1]            [,2] [,3] [,4]
##  [1,] NA              NA   NA   NA
##  [2,] NA              NA   NA   NA
##  [3,] NA              NA   NA   NA
##  [4,] NA              NA   NA   NA
##  [5,] NA              NA   NA   NA
##  [6,] NA              NA   NA   NA
##  [7,] "QQQQQQ"        "Q"  "Q"  "Q"
##  [8,] NA              NA   NA   NA
##  [9,] NA              NA   NA   NA
## [10,] NA              NA   NA   NA
## [11,] "aacBytfrhhcaa" "a"  "a"  "c"
## [12,] NA              NA   NA   NA
## [13,] NA              NA   NA   NA
```

4

Construct regular expressions to match words that: Start and end with the same character.: "^(.).*\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.) : "(.).$\vert$*1*.\1"

```
Data4 <- c("bluesky", "100001","TTT","hihihi","YaYaYu77","AOOA","QQQaQQQ","Alico","Apia","Aka","aacBytfi

result <- str_subset(Data4, "^(.).*\\1$")
result
```

```
## [1] "100001"        "TTT"           "AOOA"          "QQQaQQQ"
## [5] "aacBytfrhhcaa" "9989"
```

```
Data4 <- c("bluesky", "100001","TTT","hihihi","YaYaYu77","AOOA","QQQaQQQ","Alico","Apia","Aka","aacBytfi

result <- str_subset(Data4,"^(..).*\\1$" )
result
```

```
## [1] "hihihi"        "QQQaQQQ"       "aacBytfrhhcaa"
```

```
Data4 <- c("bluesky", "100001","TTT","hihihi","YaYaYu77","AOOA","QQQaQQQ","Alico","Apia","Aka","aacBytfi

result <- str_subset(Data4,"(.).*\\1.*\\1" )
result
```

```
## [1] "100001"        "TTT"           "hihihi"        "YaYaYu77"
## [5] "QQQaQQQ"       "aacBytfrhhcaa" "9989"
```