DATA607: HW3

Neil Shah

2/10/2020

Introduction:

This assignment will demonstrate the use of string manipulation and regex in R.

Problem 1

First let's read in the data (which can be found here (https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/majors-list.csv))

```
> df <-read.csv('https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/ma
jors-list.csv')
> head(df)
  FOD1P                                Major                  Major_Category
1  1100                   GENERAL AGRICULTURE Agriculture & Natural Resources
2  1101 AGRICULTURE PRODUCTION AND MANAGEMENT Agriculture & Natural Resources
3  1102                AGRICULTURAL ECONOMICS Agriculture & Natural Resources
4  1103                       ANIMAL SCIENCES Agriculture & Natural Resources
5  1104                         FOOD SCIENCE Agriculture & Natural Resources
6  1105            PLANT SCIENCE AND AGRONOMY Agriculture & Natural Resources
> names(df)
[1] "FOD1P"          "Major"          "Major_Category"
> df$Major
  [1] GENERAL AGRICULTURE
  [2] AGRICULTURE PRODUCTION AND MANAGEMENT
  [3] AGRICULTURAL ECONOMICS
  [4] ANIMAL SCIENCES
  [5] FOOD SCIENCE
  [6] PLANT SCIENCE AND AGRONOMY
  [7] SOIL SCIENCE
  [8] MISCELLANEOUS AGRICULTURE
  [9] FORESTRY
 [10] NATURAL RESOURCES MANAGEMENT
 [11] FINE ARTS
 [12] DRAMA AND THEATER ARTS
 [13] MUSIC
 [14] VISUAL AND PERFORMING ARTS
 [15] COMMERCIAL ART AND GRAPHIC DESIGN
 [16] FILM VIDEO AND PHOTOGRAPHIC ARTS
 [17] STUDIO ARTS
 [18] MISCELLANEOUS FINE ARTS
 [19] ENVIRONMENTAL SCIENCE
 [20] BIOLOGY
 [21] BIOCHEMICAL SCIENCES
 [22] BOTANY
 [23] MOLECULAR BIOLOGY
 [24] ECOLOGY
 [25] GENETICS
 [26] MICROBIOLOGY
 [27] PHARMACOLOGY
 [28] PHYSIOLOGY
 [29] ZOOLOGY
 [30] NEUROSCIENCE
 [31] MISCELLANEOUS BIOLOGY
 [32] COGNITIVE SCIENCE AND BIOPSYCHOLOGY
 [33] GENERAL BUSINESS
 [34] ACCOUNTING
 [35] ACTUARIAL SCIENCE
 [36] BUSINESS MANAGEMENT AND ADMINISTRATION
 [37] OPERATIONS LOGISTICS AND E-COMMERCE
 [38] BUSINESS ECONOMICS
 [39] MARKETING AND MARKETING RESEARCH
 [40] FINANCE
```

```
[41] HUMAN RESOURCES AND PERSONNEL MANAGEMENT
[42] INTERNATIONAL BUSINESS
[43] HOSPITALITY MANAGEMENT
[44] MANAGEMENT INFORMATION SYSTEMS AND STATISTICS
[45] MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION
[46] COMMUNICATIONS
[47] JOURNALISM
[48] MASS MEDIA
[49] ADVERTISING AND PUBLIC RELATIONS
[50] COMMUNICATION TECHNOLOGIES
[51] COMPUTER AND INFORMATION SYSTEMS
[52] COMPUTER PROGRAMMING AND DATA PROCESSING
[53] COMPUTER SCIENCE
[54] INFORMATION SCIENCES
[55] COMPUTER ADMINISTRATION MANAGEMENT AND SECURITY
[56] COMPUTER NETWORKING AND TELECOMMUNICATIONS
[57] MATHEMATICS
[58] APPLIED MATHEMATICS
[59] STATISTICS AND DECISION SCIENCE
[60] MATHEMATICS AND COMPUTER SCIENCE
[61] GENERAL EDUCATION
[62] EDUCATIONAL ADMINISTRATION AND SUPERVISION
[63] SCHOOL STUDENT COUNSELING
[64] ELEMENTARY EDUCATION
[65] MATHEMATICS TEACHER EDUCATION
[66] PHYSICAL AND HEALTH EDUCATION TEACHING
[67] EARLY CHILDHOOD EDUCATION
[68] SCIENCE AND COMPUTER TEACHER EDUCATION
[69] SECONDARY TEACHER EDUCATION
[70] SPECIAL NEEDS EDUCATION
[71] SOCIAL SCIENCE OR HISTORY TEACHER EDUCATION
[72] TEACHER EDUCATION: MULTIPLE LEVELS
[73] LANGUAGE AND DRAMA EDUCATION
[74] ART AND MUSIC EDUCATION
[75] MISCELLANEOUS EDUCATION
[76] LIBRARY SCIENCE
[77] ARCHITECTURE
[78] GENERAL ENGINEERING
[79] AEROSPACE ENGINEERING
[80] BIOLOGICAL ENGINEERING
[81] ARCHITECTURAL ENGINEERING
[82] BIOMEDICAL ENGINEERING
[83] CHEMICAL ENGINEERING
[84] CIVIL ENGINEERING
[85] COMPUTER ENGINEERING
[86] ELECTRICAL ENGINEERING
[87] ENGINEERING MECHANICS PHYSICS AND SCIENCE
[88] ENVIRONMENTAL ENGINEERING
[89] GEOLOGICAL AND GEOPHYSICAL ENGINEERING
[90] INDUSTRIAL AND MANUFACTURING ENGINEERING
[91] MATERIALS ENGINEERING AND MATERIALS SCIENCE
[92] MECHANICAL ENGINEERING
[93] METALLURGICAL ENGINEERING
[94] MINING AND MINERAL ENGINEERING
```

```
 [95] NAVAL ARCHITECTURE AND MARINE ENGINEERING
 [96] NUCLEAR ENGINEERING
 [97] PETROLEUM ENGINEERING
 [98] MISCELLANEOUS ENGINEERING
 [99] ENGINEERING TECHNOLOGIES
[100] ENGINEERING AND INDUSTRIAL MANAGEMENT
[101] ELECTRICAL ENGINEERING TECHNOLOGY
[102] INDUSTRIAL PRODUCTION TECHNOLOGIES
[103] MECHANICAL ENGINEERING RELATED TECHNOLOGIES
[104] MISCELLANEOUS ENGINEERING TECHNOLOGIES
[105] MATERIALS SCIENCE
[106] NUTRITION SCIENCES
[107] GENERAL MEDICAL AND HEALTH SERVICES
[108] COMMUNICATION DISORDERS SCIENCES AND SERVICES
[109] HEALTH AND MEDICAL ADMINISTRATIVE SERVICES
[110] MEDICAL ASSISTING SERVICES
[111] MEDICAL TECHNOLOGIES TECHNICIANS
[112] HEALTH AND MEDICAL PREPARATORY PROGRAMS
[113] NURSING
[114] PHARMACY PHARMACEUTICAL SCIENCES AND ADMINISTRATION
[115] TREATMENT THERAPY PROFESSIONS
[116] COMMUNITY AND PUBLIC HEALTH
[117] MISCELLANEOUS HEALTH MEDICAL PROFESSIONS
[118] AREA ETHNIC AND CIVILIZATION STUDIES
[119] LINGUISTICS AND COMPARATIVE LANGUAGE AND LITERATURE
[120] FRENCH GERMAN LATIN AND OTHER COMMON FOREIGN LANGUAGE STUDIES
[121] OTHER FOREIGN LANGUAGES
[122] ENGLISH LANGUAGE AND LITERATURE
[123] COMPOSITION AND RHETORIC
[124] LIBERAL ARTS
[125] HUMANITIES
[126] INTERCULTURAL AND INTERNATIONAL STUDIES
[127] PHILOSOPHY AND RELIGIOUS STUDIES
[128] THEOLOGY AND RELIGIOUS VOCATIONS
[129] ANTHROPOLOGY AND ARCHEOLOGY
[130] ART HISTORY AND CRITICISM
[131] HISTORY
[132] UNITED STATES HISTORY
[133] COSMETOLOGY SERVICES AND CULINARY ARTS
[134] FAMILY AND CONSUMER SCIENCES
[135] MILITARY TECHNOLOGIES
[136] PHYSICAL FITNESS PARKS RECREATION AND LEISURE
[137] CONSTRUCTION SERVICES
[138] ELECTRICAL, MECHANICAL, AND PRECISION TECHNOLOGIES AND PRODUCTION
[139] TRANSPORTATION SCIENCES AND TECHNOLOGIES
[140] MULTI/INTERDISCIPLINARY STUDIES
[141] COURT REPORTING
[142] PRE-LAW AND LEGAL STUDIES
[143] CRIMINAL JUSTICE AND FIRE PROTECTION
[144] PUBLIC ADMINISTRATION
[145] PUBLIC POLICY
[146] N/A (less than bachelor's degree)
[147] PHYSICAL SCIENCES
[148] ASTRONOMY AND ASTROPHYSICS
```

```
[149] ATMOSPHERIC SCIENCES AND METEOROLOGY
[150] CHEMISTRY
[151] GEOLOGY AND EARTH SCIENCE
[152] GEOSCIENCES
[153] OCEANOGRAPHY
[154] PHYSICS
[155] MULTI-DISCIPLINARY OR GENERAL SCIENCE
[156] NUCLEAR, INDUSTRIAL RADIOLOGY, AND BIOLOGICAL TECHNOLOGIES
[157] PSYCHOLOGY
[158] EDUCATIONAL PSYCHOLOGY
[159] CLINICAL PSYCHOLOGY
[160] COUNSELING PSYCHOLOGY
[161] INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY
[162] SOCIAL PSYCHOLOGY
[163] MISCELLANEOUS PSYCHOLOGY
[164] HUMAN SERVICES AND COMMUNITY ORGANIZATION
[165] SOCIAL WORK
[166] INTERDISCIPLINARY SOCIAL SCIENCES
[167] GENERAL SOCIAL SCIENCES
[168] ECONOMICS
[169] CRIMINOLOGY
[170] GEOGRAPHY
[171] INTERNATIONAL RELATIONS
[172] POLITICAL SCIENCE AND GOVERNMENT
[173] SOCIOLOGY
[174] MISCELLANEOUS SOCIAL SCIENCES
174 Levels: ACCOUNTING ACTUARIAL SCIENCE ADVERTISING AND PUBLIC RELATIONS ... ZOOLOGY
```

Ok–so this is our dataframe that contains our 173 majors.

#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/ (https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/)], provide code that identifies the majors that contain either "DATA" or "STATISTICS"

First, let's install stringR

```
> install.packages('stringr')
Installing package into ‘C:/Users/Neil/Documents/R/win-library/3.6’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/stringr_1.4.0.zip'
Content type 'application/zip' length 216755 bytes (211 KB)
downloaded 211 KB

package ‘stringr’ successfully unpacked and MD5 sums checked

> library(stringr)
```

Now to slice out dataframe via matching "DATA" and "STATISTICS"

```
> str_subset(df$Major,pattern='DATA')
[1] "COMPUTER PROGRAMMING AND DATA PROCESSING"
> str_subset(df$Major,pattern='STATISTICS')
[1] "MANAGEMENT INFORMATION SYSTEMS AND STATISTICS" "STATISTICS AND DECISION SCIENCE"
```

There is one major that matches “DATA”: Data Processing, and two that match statistics: “Management Information System & Statistics” and “Statistics and Decision Science”.

To manually verify for the heck of it I use CTRL-F on the master file and came up with the same results.

Problem 2

```
> data='[1] "bell pepper"  "bilberry"     "blackberry"   "blood orange"
+
+ [5] "blueberry"    "cantaloupe"   "chili pepper" "cloudberry"
+
+ [9] "elderberry"   "lime"         "lychee"       "mulberry"
+
+ [13] "olive"        "salal berry"
```

Used regex to extract all the characters (and white space A-Z, NO numbers) between quotes/word boundary

The resultant is a list–that I unlisted (using unlist) and stored as a column vector.

```
> fruits <- c(unlist(str_extract_all(data, "\\b[ a-z]+\\b")))
> fruits
 [1] "bell pepper"  "bilberry"     "blackberry"   "blood orange" "blueberry"    "cantaloupe"
"chili pepper"
 [8] "cloudberry"   "elderberry"   "lime"         "lychee"       "mulberry"     "olive"
"salal berry"
> fruits[5]
[1] "blueberry"
```

Problem 3

```
(.)\1\1
```

Would be

```
  '(.)\\1\\1'
```

This would match a character three times in a row.

```
str_view('abcbbbcaf','(.)\\1\\1')
```

abc**bbb**caf

```
"(.)(.)\\2\\1"
```

This will match two of the same characters (back to back) surrounded by the another pair of character. TWo wild cards (.) followed by the 2nd wild card, then the first wildcard. Essentially–first charcter, second, followed by second and first character.

For example; XYYX or POOP

Example:

```
> str_view(fruit,"(.)(.)\\2\\1",match=TRUE)
```

bell p**eppe**r chili p**eppe**r

```
(..)\1
```

If we put this in quotes it would be:

```
"(..)\\1
str_view(fruits, "(..)\\1")
```

It would match two wild cards followed by the pair again–basically two characters repeated twice. From the above example it would match

s**alal** berry

```
"(.).\\1.\\1"
```

This would extract three of the same wildcard characters with another character in between–for example:

```
> str_view('afabacad',"(.).\\1.\\1")
```

Extracts **afaba**cad

```
'(.)(.)(.).*\\3\\2\\1'
```

This will extract an expression (6 or more characters) that starts off with three wild card characters and ends with the reverse three characters, and whatever is in between them. Three wild cards (.) followed by .* (everything) and then the reverse. (3rd wildcard, 2nd and then 1)

For example:

```
> str_view('12abctestcba',"(.)(.)(.).*\\3\\2\\1")
```

Extracts

12**abctestcba**

Problem 4

Start and end with the same character

```
str_view(eludciate, "^(.).*\\1$")
```

Highlights **elucidate**

The ^ and $ to anchor it to the entire string, the .* to capture everything, and the the \1 makes sure it matches the same wildcard (.) from the beginnng.

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

```
str_view('church','([A-Za-z][A-Za-z]).*\\1')
```

[A-za-z] twice for the two letters we want the match, the .* for everything else, and then the \1 to match the initial two letters again.

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

```
 str_view('eleven', "([A-Za-z]).*\\1.*\\1*")
```

eleven

Here we use the a-z to get all chraracters, the .* for everything after, and the \1 (twice) to match it two more times, the * to catch everything else.

Conclusion

This was a suprsingly challenging assignment but after decoding regex, I understand the power and useful of it! Cool tool!