a_3

Jie Zou

2021-02-18

```
library(tidyverse)
```

identify the majors that contain wither "DATA" or "STATISTICS" from https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/

```
major<-read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/majors-lis
major<-data.frame(major)
data_or_stat<- str_subset(major$Major, "DATA|STATISTICS")
data_or_stat
```

```
## [1] "MANAGEMENT INFORMATION SYSTEMS AND STATISTICS"
## [2] "COMPUTER PROGRAMMING AND DATA PROCESSING"
## [3] "STATISTICS AND DECISION SCIENCE"
```

write code to transforms the data

```
idk <- c("bell pepper", "bilberry", "blackberry","blood orange", "blueberry", "cantaloupe", "chili peppe

writeLines(str_c("c(",str_c("\"",idk, "\"", collapse = ", "), ")"))
```

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

describe these expressions

(.)\1\1

It should match any strings that have one character that repeats three times if the regex is presented correctly as "(.)\1\1" in other languages like Python. However, it won't find any matches in R.

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

The regex backreferencing is right, it will present the string with 1221 format such as "anna", "eppe", etc.

(..)\1

The backreference is not coreect, therefore it will not match any string so far. However, if regex is "(..)\1", it will represent the string with 2 letters and repeats twice, such as "anan", "haha", "lolo", etc. (Note: not working in R)

“(.).\1.\1”

The backreference is fine. It will match strings like any one character(1) followed by any another character(*) followed by 1, then followed by * then followed by 1. For example, “anana”, “apaya”, etc.

“(.)(.)(.).*\3\2\1”

The backreference is good. It will match strings like one character(1), followed by a second random character(2), followed by a third random character(3), then followed by random characters then followed by 3, 2, 1. Examples: “paragrap”, “abcccba”

construct regular expressions

start and end with the same character

"^(.).*\1$"

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

"(.)(.).*\1\2"

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

"(.).*\1.*\1"