# class_Oct10.Rmd

Vinay K L

2023-10-10

# Regular Expressions continued

## Sequences

\d - match a digit character - like 0,1,2,3,4 \D - opposite of digit charcter - non-digits \s - match a space character \S - match a non-space character

```r
sub("\\d", "_", "Covid 19")
```

```
## [1] "Covid _9"
```

```r
gsub("\\D", "_", "Covid 19")
```

```
## [1] "______19"
```

```r
sub("\\s", "_", "Covid 19")
```

```
## [1] "Covid_19"
```

```r
sub("\\S", "_", "Covid 19")
```

```
## [1] "_ovid 19"
```

# Character class

[^aeiou] - match anything other than lowercase vowel

```r
d <- c("car", "bike", "plane", "boat", "oct 07", "I-II-III")

#looking for 'e' or 'i'

grep(pattern = "[ei]", x = d, value = TRUE)
```

```
## [1] "bike"  "plane"
```

```
grep(pattern = "[01]", x = d, value = TRUE)
```

```
## [1] "oct 07"
```

## POSIX Character Classes

[[:lower:]] - lower case letters [[:alpha:]] - alphabetic characters [[:digit:]] - Digits [[:alnum:]] - alphanumeric characters [[:punct:]] - puntuation characters

```
gsub(pattern = "[[:blank:]]", replacement = "", x = d )
```

```
## [1] "car"      "bike"     "plane"    "boat"     "oct07"    "I-II-III"
```

```
gsub(pattern = "[[:lower:]]", replacement = "_", x = d)
```

```
## [1] "___"      "____"     "_____"    "____"     "___ 07"   "I-II-III"
```

## Quantifiers

Number of times regex needs to run instead of 1 or all

? - zero or at most once * - zero or more times + - one more more times {n} - exactly n times {n,} - n or more times {n,m} - at least n times but not more than m times.

```
sts <- row.names(USArrests)

sts
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

```
grep(pattern = "n?", x = sts, value = TRUE)
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

```r
grep(pattern = "n{2}", sts, value = TRUE)
```

```
## [1] "Connecticut"  "Minnesota"    "Pennsylvania" "Tennessee"
```

## position of the pattern within a string

ˆ - match start of the string $ - end of the string

```r
# \b - matches the empty string at either edge of a word.
## \B matches the empty string provided it is not at a naedge of a word
```

```r
strings <- c("abcd", "cdab", "cabd", "c abd")
grep("ab", strings, value = TRUE)
```

```
## [1] "abcd"  "cdab"  "cabd"  "c abd"
```

```r
grep( "^ab", strings, value = TRUE )
```

```
## [1] "abcd"
```

```r
grep("ab$", strings, value = TRUE)
```

```
## [1] "cdab"
```

```r
grep("\\bab", strings, value = TRUE)
```

```
## [1] "abcd"  "c abd"
```

```r
grep("\\Bab", strings, value = TRUE)
```

```
## [1] "cdab" "cabd"
```

```r
grep("ab\\B", strings, value = TRUE)
```

```
## [1] "abcd"  "cabd"  "c abd"
```

## Operators

. matches any single character [...] - matches any one of the characters inside the bracket [ˆ...] - matches any other char except those inside the brackets

## Web scrapping