Text Processing (Lab-12)

Bibek Sapkota

String Manipulation:

List of String Manipulation Functions

task 1: Assigning value to virables and then printing its type.

```
text <- "san francisco"
typeof(text)
```

```
## [1] "character"
```

```
num <- c("24","34","36")
typeof(num)
```

```
## [1] "character"
```

task 2:Assing value to var3 and printing it.

```
var3 <- paste("Var1","Var2",sep = "-")
var3
```

```
## [1] "Var1-Var2"
```

task 3: Starting with 1 connect with ? and ! and assigning - sep in middle.

```
paste(1:5,c("?","!"),sep = "-")
```

```
## [1] "1-?" "2-!" "3-?" "4-!" "5-?"
```

task 4:Assiging value to variables then assiging sep in middle.

```
text= "England"
cat(text,"USA",sep = "-")
```

```
## England-USA
```

task 5:Assiging sep in middle of month name from 1 Jan to May

```r
cat(month.name[1:5],sep = " ")
```

```
## January February March April May
```

task 6:Changing 1to10 num into string using tostring.

```r
toString (1:10)
```

```
## [1] "1, 2, 3, 4, 5, 6, 7, 8, 9, 10"
```

task 7:Importing the library

```r
library(stringr)
```

task 8:Assign value to variable

```r
str <- "Los Angeles, officially the City of Los Angeles and often known by its
initials L.A., is the second-most populous city in the United States (after New
York City), the most populous city in California and the county seat of Los An
geles County. Situated in Southern California, Los Angeles is known for its Medi
terranean climate, ethnic diversity, sprawling metropolis, and as a major center
of the American entertainment industry."
strwrap(str)
```

```
## [1] "Los Angeles, officially the City of Los Angeles and often known by its"
## [2] "initials L.A., is the second-most populous city in the United States"
## [3] "(after New York City), the most populous city in California and the"
## [4] "county seat of Los An geles County. Situated in Southern California,"
## [5] "Los Angeles is known for its Medi terranean climate, ethnic diversity,"
## [6] "sprawling metropolis, and as a major center of the American"
## [7] "entertainment industry."
```

task 9:count number of characters

```r
nchar(str)
```

```
## [1] 436
```

```r
str_length(str)
```

```
## [1] 436
```

task 10:convert to lower

```r
tolower(str)
```

```
## [1] "los angeles, officially the city of los angeles and often known by its \ninitials l.a., is the s
```

```r
str_to_lower(str)
```

```
## [1] "los angeles, officially the city of los angeles and often known by its \ninitials l.a., is the s
```

task 11:Replace strings

```r
chartr("and","for",x = str) #letters a,n,d get replaced by f,o,r
```

```
## [1] "Los Aogeles, officifully the City of Los Aogeles for ofteo koowo by its \nioitifls L.A., is the s
```

```r
str_replace_all(string = str, pattern = c("City"),replacement = "state") #this is case sensitive
```

```
## [1] "Los Angeles, officially the state of Los Angeles and often known by its \ninitials L.A., is the
```

task 12:Extract parts of string

```r
substr(x = str,start = 5,stop = 11)
```

```
## [1] "Angeles"
```

task 13:Get difference between two vectors

```r
setdiff(c("monday","tuesday","wednesday"),c("monday","thursday","friday"))
```

```
## [1] "tuesday"   "wednesday"
```

task 14:Check if strings are equal

```r
setequal(c("monday","tuesday","wednesday"),c("monday","tuesday","wednesday"))
```

```
## [1] TRUE
```

```r
setequal(c("monday","tuesday","thursday"),c("monday","tuesday","wednesday"))
```

```
## [1] FALSE
```

task 15:Abbreviate strings

```r
abbreviate(c("monday","tuesday","wednesday"),minlength = 3)
```

```
##    monday   tuesday wednesday
##     "mnd"     "tsd"     "wdn"
```

task 16:spliting strings

```r
strsplit(x = c("ID-101","ID-102","ID-103","ID-104"),split = "-")
```

```
## [[1]]
## [1] "ID"  "101"
##
## [[2]]
## [1] "ID"  "102"
##
## [[3]]
## [1] "ID"  "103"
##
## [[4]]
## [1] "ID"  "104"
```

```r
str_split(string = c("ID-101","ID-102","ID-103","ID-104"),pattern = "-",simplify = T)
```

```
##      [,1] [,2]
## [1,] "ID" "101"
## [2,] "ID" "102"
## [3,] "ID" "103"
## [4,] "ID" "104"
```

task 17:find and replace first match

```r
sub(pattern = "L",replacement = "B",x = str,ignore.case = T)
```

```
## [1] "Bos Angeles, officially the City of Los Angeles and often known by its \ninitials L.A., is the s
```

task 18:find and replace all matches

```r
gsub(pattern = "Los",replacement = "Bos",x = str,ignore.case = T)
```

```
## [1] "Bos Angeles, officially the City of Bos Angeles and often known by its \ninitials L.A., is the s
```

Question 1:Write a command to extract the first 5 characters of above-given text (string).

```r
dt <- c("Soloman", "abcdef", "snakjs")
first_5_characters <- substr(dt, start = 1, stop = 5)
print(first_5_characters)
```

```
## [1] "Solom" "abcde" "snakj"
```

Metacharacters

task 1:

```
dt <- c("20","20$")
grep(pattern = "20\\$",x = dt,value = T)
```

```
## [1] "20$"
```

task 2:sub() function to make the replacements.

```
dt <- c("may?","money$","and&")
gsub(pattern = "[\\?-\\$-\\&]",replacement = "",x = dt)
```

```
## [1] "may"   "money" "and"
```

task 3: double backslash in a string, you'll need to prefix it with another double backslash to get detected

```
gsub(pattern = "\\\\",replacement = "-",x = "Barcelona\\Spain")
```

```
## [1] "Barcelona-Spain"
```

Question 2: Write a code to capture the first two elements in dt <- c("may?,"money$" but not "and&").

```
dt <- c("may?", "money$", "and&")
result <- grep("^(may\\?|money\\$)$", dt, value = TRUE)
print(result)
```

```
## [1] "may?"   "money$"
```

Quantifiers

task 1:Assign value to variable

```
names <- c("anna","crissy","puerto","cristian","garcia","steven","alex","rudy")
```

task 2:The symbol .? is known as a non-greedy quantifier. Being non-greedy, for a particular pattern to be matched, it will stop at the first match.

```
#doesn't matter if e is a match
grep(pattern = "e*",x = names,value = T)
```

```
## [1] "anna"     "crissy"   "puerto"   "cristian" "garcia"   "steven"   "alex"
## [8] "rudy"
```

```
#must match t one or more times
grep(pattern = "t+",x = names,value = T)
```

```
## [1] "puerto"   "cristian" "steven"
```

```r
#must match n two times
grep(pattern = "n{2}",x = names,value = T)
```

```
## [1] "anna"
```

Question 3:Write a regex pattern to match 'c' one or more times.

```r
# Use grep to find elements containing 'c' one or more times
matches <- grep("c+", names, value = TRUE)

# Print the matching names
print(matches)
```

```
## [1] "crissy"   "cristian" "garcia"
```

Sequence

task 1:matching a digit

```r
str <- "I have been to Paris 20 times"

gsub(pattern = "\\d+",replacement = "_",x = str)
```

```
## [1] "I have been to Paris _ times"
```

```r
regmatches(str,regexpr(pattern = "\\d+",text = str))
```

```
## [1] "20"
```

task 2:match a non-digit

```r
gsub(pattern = "\\D+",replacement = "_",x = str)
```

```
## [1] "_20_"
```

```r
regmatches(str,regexpr(pattern = "\\D+",text = str))
```

```
## [1] "I have been to Paris "
```

task 3:match a space - returns positions

```r
gregexpr(pattern = "\\s+",text = str)
```

```
## [[1]]
## [1]  2  7 12 15 21 24
## attr(,"match.length")
## [1] 1 1 1 1 1 1
## attr(,"index.type")
## [1] "chars"
## attr(,"useBytes")
## [1] TRUE
```

task 4:match a non space

```
gsub(pattern = "\\S+",replacement = "app",x = str)
```

```
## [1] "app app app app app app app"
```

task 5:match a word character

```
gsub(pattern = "\\w",replacement = "k",x = str)
```

```
## [1] "k kkkk kkkk kk kkkkk kk kkkkk"
```

task 6:match a non-word character

```
gsub(pattern = "\\W",replacement = "k",x = str)
```

```
## [1] "IkhavekbeenktokParisk20ktimes"
```

Character Classes

task 1:extract numbers

```
str <- "20 people got killed in the mob attack. 14 got severely injured"

regmatches(x = str,gregexpr("[0-9]+",text = str))
```

```
## [[1]]
## [1] "20" "14"
```

task 2:extract without digits

```
regmatches(x = str,gregexpr("[^0-9]+",text = str))
```

```
## [[1]]
## [1] " people got killed in the mob attack. "
## [2] " got severely injured"
```

Lecture examples:

task 1:Checks if strings match three characters followed by a period.

```
str<- c("cat.", "896.", "?=+.", "abc1")
grepl(pattern = "...\\.",x=str)
```

```
## [1]  TRUE  TRUE  TRUE FALSE
```

task 2:Checks if strings have 'a's, followed by 'b's, ending with 'c's.

```
str<- c("aaaabc", "aabbbc", "aacc", "a")
grepl(pattern = "a+b*c+",x=str)
```

```
## [1]  TRUE  TRUE  TRUE FALSE
```

task 3:Checks if strings have 'waz' followed by 3-4 'z's and 'up'.

```
str<- c("wazzzzup", "wazzzup", "wazup")
grepl(pattern = "waz{3,4}up",x=str)
```

```
## [1]  TRUE  TRUE FALSE
```

task 4:Selects strings with lowercase, optional period, space, uppercase, and digit.

```
str<-c("A. B","c! d", "e f", "g. H3", "i? J", "k L")
grep(pattern ="[a-z]\\.?\\s+[A-Z]\\d*",x=str)
```

```
## [1] 4 6
```

task 5:Selects strings in mm/dd/yyyy or mm/dd/yy format.

```
str<- c("09/01/2016", "09/21/16", "12/25/2016", "12/05/16")
grep(pattern ="^([0-2][0-9]|(3)[0-1])(\\/)((0)[0-9]|(1)[0-2])(\\/)\\d{4}$", x=str)
```

```
## [1] 1
```

Question 4:Write a code to match only the first digit in dt <- c("75 to 79", "80 to 84", "85 to 89").

```
dt <- c("75 to 79", "80 to 84", "85 to 89")
first_digits <- gsub(pattern = "(\\d+).*", replacement = "\\1", x = dt)
print(first_digits)
```

```
## [1] "75" "80" "85"
```