Class Activity 12

Your name here

April 18 2024

In this tutorial, we will learn about string manipulations using regular expressions and the `stringr` library in R. We will cover different examples and use cases to help you understand the concepts and functions related to string manipulation.

Group Activity 1

```
x <- "My SSN is 593-29-9502 and my age is 55"
y <- "My phone number is 612-643-1539"
z <- "My old SSN number is 39532 9423."
out <- str_flatten(c(x,y,z), collapse = ". ")
```

a. What characters in `x` will `str_view_all(x, "-..-")` find?

answer:

b. What pattern will `str_view_all(x, "-\\d{2}-")` find?

answer:

c. What pattern will `str_view_all(out, "\\d{2}\\.*")` find?

answer:

d. Use `str_view_all` to determine the correct regex pattern to identify all SSN in `out`

answer:

This misses the oddly formatted SSN in the third entry. Rather than use a dash, we can specify the divider as `[-\\s]?` which allows either 0 or 1 occurrences of either a dash or space divider:

e. Write a regular expression to extract dates in the format YYYY-MM-DD from a given text.

```
date_pattern <- ""
text <- "The event will take place on 2023-07-20 and end on 2023-07-22."
str_extract_all(text, date_pattern)
[[1]]
 [1] "T" "h" "e" " " "e" "v" "e" "n" "t" " " "w" "i" "l" "l" " " "t" "a" "k" "e"
[20] " " "p" "l" "a" "c" "e" " " "o" "n" " " "2" "0" "2" "3" "-" "0" "7" "-" "2"
[39] "0" " " "a" "n" "d" " " "e" "n" "d" " " "o" "n" " " "2" "0" "2" "3" "-" "0"
[58] "7" "-" "2" "2" "."
```

answer:

f. Write a regular expression to extract all words that start with a capital letter in a given text.

answer:

```
capital_pattern <- ""
text <- "Alice and Bob went to the Market to buy some Groceries."
str_extract_all(text, capital_pattern)
[[1]]
 [1] "A" "l" "i" "c" "e" " " "a" "n" "d" " " "B" "o" "b" " " "w" "e" "n" "t" " "
[20] "t" "o" " " "t" "h" "e" " " "M" "a" "r" "k" "e" "t" " " "t" "o" " " "b" "u"
[39] "y" " " "s" "o" "m" "e" " " "G" "r" "o" "c" "e" "r" "i" "e" "s" "."
```

Group Activity 2

Consider the following string:

```
string1 <- "100 dollars 100 pesos"
```

a. Explain why the following matches the first 100 and not the second.

answer:

```
str_view(string1, "\\d+(?= dollars)")
[1] | <100> dollars 100 pesos
```

b. Explain why the following matches the second 100 and not the first.

answer:

```
str_view(string1, "\\d+(?!\\d| dollars)")
[1] | 100 dollars <100> pesos
```

Please take a look at `string2`:

```
string2 <- "USD100 PESO100"
```

c. Explain why the following matches the first 100 and not the second.

answer:

```
str_view(string2, "(?<=USD)\\d{3}")
[1] | USD<100> PESO100
```

d. Explain why the following matches the second 100 and not the first.

answer:

```
str_view(string2, "(?<!USD)\\d{3}")
[1] | USD100 PESO<100>
```

Group Activity 3

```
tweets<- read_csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/TrumpTweetData.csv")
```

a. What proportion of tweets (text) mention "America"?

```
tweets %>%
  summarize(prop = mean(str_detect(str_to_title(text), "America")))
# A tibble: 1 x 1
    prop
   <dbl>
1 0.0926
```

b. What proportion of these tweets include "great"?

```
tweets %>%

Error: <text>:3:0: unexpected end of input
1: tweets %>%
2:
  ^
```

c. What proportion of the tweets mention @?

```
tweets %>%
Error: <text>:2:0: unexpected end of input
1: tweets %>%
   ^
```

d. Remove the tweets having mentions @.

```
Mentions <- c("@[^\\s]+")

tw_noMentions <- tweets %>%
Error: <text>:4:0: unexpected end of input
2:
3: tw_noMentions <- tweets %>%
  ^
```

e. What poportion of tweets originated from an iPhone?

```
tweets %>%
Error: <text>:2:0: unexpected end of input
1: tweets %>%
   ^
```