Working with text data in R

Learn R online at www.DataCamp.com

> Packages to install for this cheat sheet

Some functionality from this cheat sheet comes with base-R, but the following packages are also used throughout this cheat sheet.

```
library(stringr)
library(snakecase)
library(glue)
```

Functions with names starting `str_` are from `stringr`; those with names starting `to_` are from `snakecase`; those with `glue` in the name are from `glue`.

> Example data

Throughout this cheat sheet, we'll be using this vector containing the following strings.

```
suits <- c("Clubs", "Diamonds", "Hearts", "Spades")
```

> Get string lengths and substrings

```
# Get the number of characters with nchar()
nchar(suits) # Returns 5 8 6 6

# Get substrings by position with str_sub()
stringr::str_sub(suits, 1, 4) # Returns "Club" "Diam" "Hear" "Spad"

# Remove whitespace from the start/end with str_trim()
str_trim(" Lost in Whitespace ") # Returns "Lost in Whitespace"

# Truncate strings to a maximum width with str_trunc()
str_trunc(suits, width = 5) # Returns "Clubs" "Di..." "He..." "Sp..."

# Pad strings to a constant width with str_pad()
str_pad(suits, width = 8) # Returns " Clubs" "Diamonds" " Hearts" " Spades"

# Pad strings on right with str_pad(side="right")
str_pad(suits, width = 8, side = "right", pad = "!")
# Returns "Clubs!!!" "Diamonds" "Hearts!!" "Spades!!"
```

> Changing case

```
# Convert to lowercase with tolower()
tolower(suits) # Returns "clubs" "diamonds" "hearts" "spades"

# Convert to uppercase with toupper()
toupper(suits) # Returns "CLUBS" "DIAMONDS" "HEARTS" "SPADES"

# Convert to title case with to_title_case()
to_title_case("hello, world!") # Returns "Hello, World!"

# Convert to sentence case with to_sentence_case()
to_sentence_case("hello, world!") # Returns "Hello, world!"
```

> Formatting strings

```
# Format numbers with sprintf()
sprintf("%.3e", pi) # "3.142e+00"

# Substitute value in a string with an expression
glue('The answer is {ans}', ans = 30 + 10) # The answer is 40

# Substitute value in a string with an expression
cards <- data.frame(value = c("8", "Queen", "Ace"),
                    suit = c("Diamonds", "Hearts", "Spades"))
cards %>% glue_data("{value} of {suit}")

# 8 of Diamonds
# Queen of Hearts
# Ace of Spades

# Wrap strings across multiple lines
str_wrap('The answer to the universe is 42', width = 25)
# The answer to the
# universe is 42
```

> Splitting strings

```
# Split strings into list of characters with str_split(pattern = "")
str_split(suits, pattern = "")

# "C" "l" "u" "b" "s"
# "D" "i" "a" "m" "o" "n" "d" "s"
# "H" "e" "a" "r" "t" "s"
# "S" "p" "a" "d" "e" "s"

# Split strings by a separator with str_split()
str_split(suits, pattern = "a")

# "Clubs"
# "Di" "monds"
# "He" "rts"
# "Sp" "des"

# Split strings into matrix of n pieces with str_split_fixed()
str_split_fixed(suits, pattern = 'a', n = 2)

# [,1] [,2]
# [1,] "Clubs" ""
# [2,] "Di" "monds"
# [3,] "He" "rts"
# [4,] "Sp" "des"
```

> Joining or concatenating strings

```
# Combine two strings with paste0()
paste0(suits, '5') # "Clubs5" "Diamonds5" "Hearts5" "Spades5"

# Combine strings with a separator with paste()
paste(5, suits, sep = " of ") # "5 of Clubs" "5 of Diamonds" "5 of Hearts" "5 of
Spades"

# Collapse character vector to string with paste() or paste0()
paste(suits, collapse = ", ") # "Clubs, Diamonds, Hearts, Spades"

# Duplicate and concatenate strings with str_dup()
str_dup(suits, 2) # "ClubsClubs" "DiamondsDiamonds" "HeartsHearts"
"SpadesSpades"
```

> Detecting matches

```
# Highlight string matches in HTML widget with str_view_all()
str_view_all(suits, "[ae]")
```

Clubs
Diamonds
Hearts
Spades

```
# Detect if a regex pattern is present in strings with str_detect()
str_detect(suits, "[ae]") # FALSE TRUE TRUE TRUE

# Find the index of strings that match a regex with str_which()
str_which(suits, "[ae]") # 2 3 4

# Count the number of matches with str_count()
str_count(suits, "[ae]") # 0 1 2 2

# Locate the position of matches within strings with str_locate()
str_locate(suits, "[ae]")

# start end
# [1,] NA NA
# [2,] 3 3
# [3,] 2 2
# [4,] 3 3
```

> Extracting matches

```
# Extract matches from strings with str_extract()
str_extract(suits, ".[ae].") # NA "iam" "Hea" "pad"

# Extract matches and capture groups with str_match()
str_match(suits, ".([ae])(.)")

# [,1] [,2] [,3]
# [1,] NA NA NA
# [2,] "iam" "a" "m"
# [3,] "Hea" "e" "a"
# [4,] "pad" "a" "d"

# Get subset of strings that match with str_subset()
str_subset(suits, "d") # "Diamonds" "Spades"
```

> Replacing matches

```
# Replace a regex match with another string with str_replace()
str_replace(suits, "a", "4") # "Clubs" "Di4monds" "He4rts" "Sp4des"

# Remove a match with str_remove()
str_remove(suits, "s") # "Club" "Diamond" "Heart" "Spade"

# Replace a substring with `str_sub<-`()
str_sub(suits, start = 1, end = 3) <- c("Bi", "Al", "Yu", "Hi")
suits # Returns "Bibs" "Almonds" "Yurts" "Hides"
```