R语言stringr包的使用

### 最近还在适应美帝的生后，并且因为还没有开始新的课题，所以晚上时间会比较充裕一些，一直在看R for data science这本书，前面一直没有记笔记，后面觉得还是需要养成记笔记的习惯。所以从今天开始，即使每天少看一些，也要记下来自己的读书笔记。今天就从字符串处理stringr包开始，也是第一部分。

需要先加载包

library(tidyverse)

## -- Attaching packages ---------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.1 v purrr 0.3.2   
## v tibble 2.1.1 v dplyr 0.8.0.1  
## v tidyr 0.8.3 v stringr 1.4.0   
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(stringr)

### 1 string lenght

如何查看一个字符串中字符的长度

str\_length(c("a", "R for data science", NA))

## [1] 1 18 NA

### 2 合并字符串

str\_c("x", "y")

## [1] "xy"

str\_c("x", "y", "z")

## [1] "xyz"

同样可以使用*sep*参数来控制合并时中间分隔符。

str\_c("x", "y", sep = ",")

## [1] "x,y"

和很多其他函数一样，NA值是一个问题，比如下面的例子，这时候可以使用str\_replace\_na()函数来处理：

x <- c("abc", NA)  
str\_c("|-", x, "-|")

## [1] "|-abc-|" NA

str\_c("|-", str\_replace\_na(x), "-|")

## [1] "|-abc-|" "|-NA-|"

同样从上面的例子也可以看到，str\_c是一个vector函数，他会将长度较短的对象循环，使其与长度最长的对象长度一样。

str\_c("prefix-", c("a", "b", "c"), "-suffix")

## [1] "prefix-a-suffix" "prefix-b-suffix" "prefix-c-suffix"

下面要简单介绍一下在str\_c中非常有用的两个参数，sep和collapse。其中collapse非常形象，坍塌，也就是将一个向量中的多个对象合并为一个字符。可以看下面的例子：

str\_c(c("x", "y", "z"), collapse = ", ")

## [1] "x, y, z"

str\_c(c("x", "y", "z"), sep = ", ")

## [1] "x" "y" "z"

### 3 提取字字符

有时候我们需要一个字符串中的某部分，这时候可以使用str\_sub()。

x <- c("Apple", "Banana", "Pear")  
str\_sub(string = x, start = 1, end = 3)

## [1] "App" "Ban" "Pea"

该函数一共有三个参数，start和end分别是指从哪开始和结束提取字符串。 还可以反过来提取字符串，也就是将参数设置为负数。

str\_sub(x, -3, -1)

## [1] "ple" "ana" "ear"

### 4 区域设置

对于某些规定，不同国家或者地区不太一样，比如排序，默认英语里面的排序是按照英语字母顺序，但是不同地方可能不一样，比如下面的例子：

x <- c("apple", "eggplant", "banana")  
str\_sort(x, locale = "en") # English

## [1] "apple" "banana" "eggplant"

str\_sort(x, locale = "haw") # Hawaiian

## [1] "apple" "eggplant" "banana"

感觉这个功能不会太常用。了解一下就好了。