

Guide 5 strings in R 2

September 10, 2019

1 Guía 4: Strings en R (2)

Computación 2, IES. Profesor: Eduardo Jorquera, eduardo.jorquera@postgrado.uv.cl

1.1 Librerías!

Usaremos los siguientes paquetes para manipular cadenas de caracteres:

```
In [4]: library(tidyverse)
        library(stringr)

Attaching packages: tidyverse 1.2.1
  ggplot2 3.2.1      purrr   0.3.2
  tibble  2.1.3      dplyr   0.8.3
  tidyr   0.8.3      stringr 1.4.0
  readr   1.3.1      forcats 0.4.0
Conflicts: tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()
```

2 Subconjunto de strings

Puedes extraer partes de un string usando `str_sub()`. Así, la función toma un `start` y un `end` como argumentos los cuales dan la posición (inclusiva) de la cadena de caracteres:

```
In [6]: x <- c("Manzana", "Naranja", "Pera")
        str_sub(x, 1, 3)

        str_sub(x, -3, -1)
```

1. 'Man' 2. 'Nar' 3. 'Per'

1. 'ana' 2. 'nja' 3. 'era'

Note que `str_sub()` no fallará si el string es demasiado corto, simplemente retornará tanto como sea posible:

```
In [7]: str_sub("a", 1, 5)
```

'a'

También puedes asignar a `str_sub()` para modificar strings:

```
In [8]: str_sub(x, 1, 1) <- str_to_lower(str_sub(x, 1, 1))
      x
```

1. 'manzana' 2. 'naranja' 3. 'pera'

3 Locales

Antes usamos `str_to_lower()` para cambiar las letras minúsculas. También puedes usar `str_to_upper()` ó `str_to_title()`. De cualquier manera, cambiar el caso es más complicado de lo que parece si trabajas con más de un idioma. A la configuración de cada idioma, es a lo que se le llama "locale". Los distintos lenguajes usan diferentes reglas de gramática y escritura para usar mayúsculas y minúsculas. Puedes seleccionar cualquier conjunto de reglas para usar especificando un locale:

```
In [9]: str_to_upper(c("i", ""))
      str_to_upper(c("i", ""), locale = "tr") #turco
```

1. 'I' 2. 'I'

1. '' 2. 'I'

El locale es especificado como un código de language ISO 639, el cual es una abreviación de dos o tres letras. Wikipedia tiene un buen listado de códigos de language (https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes). Si dejas el campo en blanco, usará el locale almacenado en tu computador, proveído por tu sistema operativo.

Otra operación importante que es afectado por el locale es el orden. La base de R tiene las funciones `order()` y `sort()`, que usan el locale del sistema. Si quieres un comportamiento más robusto para diferentes computadores, quizás estés interesado en usar `str_sort()` y `str_order()`, que toman el argumento `locale` adicionalmente:

```
In [12]: x <- c("apple", "eggplant", "banana")
```

```
      str_sort(x, locale = "en") # inglés
```

```
      str_sort(x, locale = "haw") # hawaii
```

1. 'apple' 2. 'banana' 3. 'eggplant'

1. 'apple' 2. 'eggplant' 3. 'banana'

4 Ejercicios

- En varios códigos que no usan `stringr`, frecuentemente verás las funciones `paste` y `paste0`.Cuál es la diferencia entre las dos funciones? Qué función de `stringr` es equivalente? Cómo difieren en el uso de `NA`?
- Con tus palabras, describe la diferencia entre los argumentos `sep` y `collapse` en `str_c()`.

- Usa `str_length()` y `str_sub()` para extraer el caracter del medio de un string. Qué harías si el string tiene número par de caracteres?
- Qué hace `str_wrap()`? Cuándo querrías usarlo?
- Qué hace `str_trim()`?Cuál es el opuesto de esta función?
- Escribe una función que convierte (por ejemplo) un vector `c("a", "b", "c")` en un string `a`, `b`, y `c`, sin importar el largo del input. Piensa cuidadosamente sobre qué debería hacer si el input es un vector de tamaño 0, 1, ó 2.

In []:

5 Emparejar patrones con expresiones regulares

Las expresiones regulares te permiten describir patrones en strings. Toma un poco de tiempo digerirlo en la cabeza, pero una vez que las entiendes, son sumamente útiles.

Para aprender expresiones regulares, usaremos `str_view()` y `str_view_all()`. Estas funciones toman un vector de caracteres y expresiones regulares, y muestran cómo se emparejan. Comenzaremos con expresiones regulares simples y gradualmente se tornan más y más complicadas. Una vez que seas un maestro de las expresiones regulares, aprenderás cómo aplicar varias ideas con varias funciones de `stringr`.

6 Emparejamientos simples

El emparejamiento de patrones más simples es con strings exactos:

```
In [19]: x <- c("Manzana", "Banana", "Pera")
         str_view(x, "an")
```

HTML widgets cannot be represented in plain text (need html)

7 Tarea

1. Guarda el siguiente texto en un archivo y léelo importándolo a R:

```
In [ ]: Polina was not at all pleased at my questions; I could see that she was doing her best
        "It amuses me to see you grow angry," she continued. "However, inasmuch as I allow you
        "I consider that I have a perfect right to put these questions to you," was my calm re
        Polina giggled.
        "Last time you told mewhen on the Shlangenbergthat at a word from me you would be read
        Then she made a movement to rise. Her tone had sounded very angry. Indeed, of late her
```

"May I ask you who is this Mlle. Blanche?" I **inquired** (since I did not wish Polina to c

"You **KNOW** who she is just Mlle. Blanche. Nothing further has transpired. Probably she w

"And is the General at last in love?"

"That has nothing to do with it. Listen to me. Take these 700 florins, and go and play
So saying, she called Nadia back to her side, and entered the Casino, where she joined

Still, she had charged me with a commission to win what I could at roulette. Yet all the

2. Cuántos párrafos tiene el texto?
3. Cuántos caracteres tiene el texto?
4. Colapse los párrafos en uno y muéstrelo en pantalla (no en una lista)
5. Convierta el texto a mayúsculas y guárdelo en un nuevo archivo "gambler-upper.txt".
6. Al texto original, cambie las letras 'a' y 't' por 'A' y 'T'.
7. El texto contiene la palabra "lucky"?
8. Cuántas palabras hay en el texto original? Asumiendo que las palabras son sub-strings (sub-cadenas de caracteres) separadas por un espacio o el carácter de nueva línea?
9. Cuántas veces la palabra "money" se encuentra en el texto?