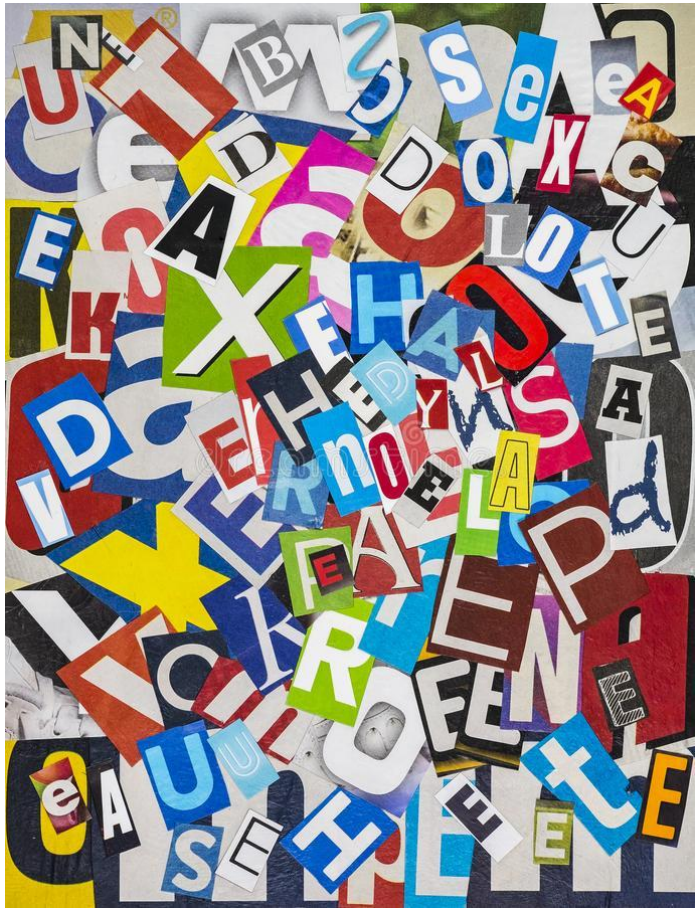
Manipular una cadena de texto con stringr

El paquete stringr proporciona un conjunto cohesivo de funciones diseñadas para hacer que trabajar con strings sea lo más fácil posible.

Las variables string también son llamadas variables alfanuméricas, cadena de texto/palabras/caracteres.

Son variables QUE SON TRATADAS COMO TEXTO (Recordar: un número no siempre es un número).

Lectura recomendada: Capítulo sobre Strings. R for Data Science.



Mg. Christian Chiroque Ruiz | christian.chiroque@pucp.edu.pe

Manipular una cadena de texto con stringr

El trabajar con cadenas de texto no es la parte más “amigable” de la tarea de preprocesamiento. Esto debido a que el principal mecanismo para trabajar con strings es nuestra capacidad para visualizar patrones. Para ello existen las EXPRESIONES REGULARES.

Qué son expresiones regulares? (regex)

En cómputo teórico y teoría de lenguajes formales, una expresión regular, o expresión racional, también son conocidas como regex o regexp, por su contracción de las palabras inglesas regular expression, es una secuencia de caracteres que conforma un patrón de búsqueda.

Se utilizan principalmente para la búsqueda de patrones de cadenas de caracteres u operaciones de sustituciones.

Qué son expresiones regulares? (regex)

Símbolo	Cualquier texto que…
[:alpha:]	… contenga una letra
.x.	… contenga una x entre dos caracteres.
\\.	…contenga un punto. Ojo: Se necesita dos backslash (alt+92 en Windows)
\\\\	…contenga un backslash (Se necesitan 4 backslash)
^x	… inicie con x (^ alt+94)
$x	… termine con x
\\d	…contenga dígitos
\\s	…contenga espacios
[abc]	…contenga los elementos a, b o c
[^abc]	…contenga cualquier elemento menos a, b o c. Ten cuidado: Recuerda que ^ significa también “al principio”. Sin embargo, si está dentro de [] equivale a negación.
(?<=x)	…esté precedido por x
(?=x)	…. Esté seguido por x

Qué son expresiones regulares? (regex)

Repeticiones: Puedes identificar el patrón de repetición dentro de la cadena. De esa manera:

Símbolo	Significado
?	0 o 1 repeticiones
+	1 o más repeticiones
*	0 o más repeticiones

También puedes identificar un número específico de veces:

Símbolo	Significado
{n}	n veces
{n,}	n o más
{,m}	Al menos m
{n,m}	Entre n y m veces

Qué son expresiones regulares? (regex)

Ejemplos:

Patrón	Todos los caracteres que..
“^a”	…comiencen con la letra a
“$a”	…terminen con la letra a
".a."	…tengan dentro del caracter la letra a
"\\.com"	…contenga la frase “.com”
"gr(e|a)y"	…contenga la frase “grey” o “gray”
"CC+"	…contengan 1 vez o más veces el frase CC
"C{4}"	…contengan la letra C cuatro veces seguidas
“\\s”	…contengan espacios en blanco

Manipular una cadena de texto con stringr

Dentro del paquete strings hay más de 20 funciones, sin embargo, se podría resaltar las siguientes:

Función	Detalle
str_extract	Extrae el fragmento de string indicado.
str_split	Parte la cadena en un patrón determinado
str_subset	Extrae una porción de la cadena de acuerdo a la ubicación de ciertos caracteres.
str_view	Muestra los matches que se realicen con el patrón indicado.
str_to_lower	Cambia todo a minúscula
str_to_upper	Cambia todo a mayúscula

Manipular una cadena de texto con stringr

Practiquemos con la data de titanic, la variable Name:

```
> head(df$Name)
[1] "Braund, Mr. Owen Harris"
[2] "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
[3] "Heikkinen, Miss. Laina"
[4] "Futrelle, Mrs. Jacques Heath (Lily May Peel)"
[5] "Allen, Mr. William Henry"
[6] "Moran, Mr. James"
```

Para un mayor detalle de las funciones y su utilidad te recomiendo revisar: Cheat Sheet de Stringr

Manipular una cadena de texto con stringr

Práctica:

- Cadena que empiece con cualquier vocal.

"[aeiou]"

- Cadena que consiste sólo en consonantes.

"^[^aeiou]+$"

- Cadena que termine en x.

“x$”