Práctica grupo 9 - parte dos

Alberto Barreiro Tasende

2026-01-12

En esta parte se recorre la tabla de URLs (generada en la Parte 1) para visitar cada artículo del Código Penal en la web de Conceptos Jurídicos.
Para cada artículo se extrae el texto normativo, se limpia eliminando contenido explicativo/no normativo, se insertan saltos de línea para mejorar la legibilidad y se genera:

- 'texto_articulo': texto del artículo limpio y formateado
- 'contexto': libro/título/capítulo/sección concatenados
- 'nombre_doc': nombre de documento plano compatible con quanteda

El resultado final se guarda como `articulos_limpios.rds` y `articulos_limpios.csv`. En primer lugar se cargan las librerías necesarias y se lee la tabla de artículos generada en la Parte 1.

```
library(rvest)
library(dplyr)
library(stringr)
library(purrr)

articulos <- read.csv("~/Downloads/tabla_articulos.csv", stringsAsFactors = FALSE)
articulos$articulo <- gsub("/", "", articulos$articulo) #normalizamos a 1 (venía con formato tipo /1)
```

Se define una función que accede a la URL de cada artículo, extrae el texto normativo y elimina contenido explicativo de la web.

```
extraer_texto_articulo <- function(url) {
  pagina <- rvest::read_html(url)

  texto <- pagina %>%
    rvest::html_elements("p") %>%
    rvest::html_text() %>%
    paste(collapse = " ") %>%
    stringr::str_squish()

  texto <- stringr::str_split(texto, "\\bart\\s*\\d+\\s*cp\\b", n = 2)[[1]][1]
  texto <- stringr::str_split(texto, "\\bEl artículo\\b", n = 2)[[1]][1]
  texto <- stringr::str_squish(texto)

  texto
}
```

Para facilitar la lectura del contenido (por ejemplo en inspecciones manuales), se insertan saltos de línea aproximadamente cada 50 caracteres.

```r
insertar_saltos <- function(texto, n = 50) {
  stringr::str_wrap(texto, width = n)
}
```

Como quanteda trabaja con documentos “planos”, se genera:

- ‘contexto’: concatenación de libro, título, capítulo y sección (separados por saltos de línea).
- ‘nombre_doc’: identificador único en una sola cadena que incluye el número de artículo y su jerarquía.

```r
crear_contexto <- function(libro, titulo, capitulo, seccion) {
  partes <- c(libro, titulo, capitulo, seccion)
  partes <- partes[!is.na(partes) & partes != ""]
  paste(partes, collapse = "\n")
}

crear_nombre_doc <- function(libro, titulo, capitulo, seccion, articulo) {
  partes <- c(libro, titulo, capitulo, seccion, paste("Artículo", articulo))
  partes <- partes[!is.na(partes) & partes != ""]
  paste0("art.", articulo, ":", paste(partes, collapse = ":"))
}
```

Antes de lanzar el scraping completo, se realiza una prueba sobre los tres primeros artículos para validar que el texto, el contexto y el nombre del documento se generan correctamente.

```r
test <- articulos %>%
  slice(1:3) %>%
  mutate(
    texto_articulo = map_chr(url, extraer_texto_articulo),
    texto_articulo = map_chr(texto_articulo, insertar_saltos),
    contexto = pmap_chr(list(libro, titulo, capitulo, seccion), crear_contexto),
    nombre_doc = pmap_chr(list(libro, titulo, capitulo, seccion, articulo), crear_nombre_doc)
  )

cat(test$nombre_doc[1], "\n\n")
```

```
## art.1:TÍTULO PRELIMINAR: De las garantías penales y de la aplicación de la Ley penal:Artículo 1
```

```r
cat(test$contexto[1], "\n\n")
```

```
## TÍTULO PRELIMINAR: De las garantías penales y de la aplicación de la Ley penal
```

```r
cat(test$texto_articulo[1], "\n")
```

```
## 1. No será castigada ninguna acción ni omisión que
## no esté prevista como delito por ley anterior a
## su perpetración. 2. Las medidas de seguridad sólo
## podrán aplicarse cuando concurran los presupuestos
## establecidos previamente por la Ley.
```

Se aplica el proceso a todos los artículos y se guarda el resultado tanto en formato RDS (para reutilizar en R) como en CSV (para inspección/compartición).

```r
articulos_limpios <- articulos %>%
  mutate(
    texto_articulo = map_chr(url, extraer_texto_articulo),
    texto_articulo = map_chr(texto_articulo, insertar_saltos),
    contexto = pmap_chr(list(libro, titulo, capitulo, seccion), crear_contexto),
    nombre_doc = pmap_chr(list(libro, titulo, capitulo, seccion, articulo), crear_nombre_doc)
  )

saveRDS(articulos_limpios, "articulos_limpios.rds")
write.csv(articulos_limpios, "articulos_limpios.csv", row.names = FALSE)
```