# Parte 4

Ada Moral

2026-01-09

```
library(quanteda)
```

```
## Package version: 4.3.1
## Unicode version: 15.1
## ICU version: 74.1
```

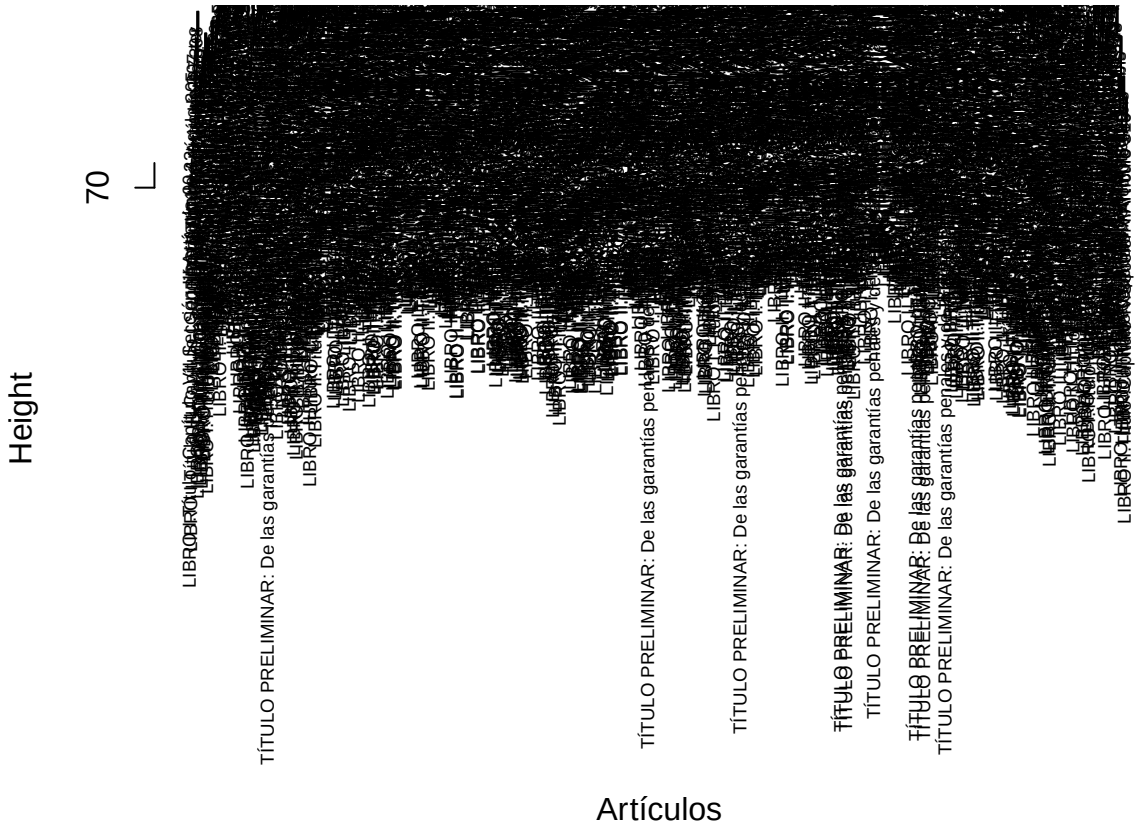
```
## Parallel computing: 12 of 12 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```
# 1. Cargar el corpus generado en la parte anterior
corpus_cp <- readRDS("corpus_codigo_penal.rds")

# --- FUNCIÓN AUXILIAR PARA OBTENER LOS MÁS SIMILARES ---
obtener_mas_similares <- function(dist_matrix) {
  dist_df <- as.data.frame(as.matrix(dist_matrix))
  diag(dist_df) <- NA # Quitamos la distancia a sí mismo

  # Encontrar el valor mínimo (excluyendo NA)
  min_dist <- min(dist_df, na.rm = TRUE)
  pos <- which(dist_df == min_dist, arr.ind = TRUE)

  art1 <- rownames(dist_df)[pos[1, 1]]
  art2 <- rownames(dist_df)[pos[1, 2]]

  return(list(art1 = art1, art2 = art2, valor = min_dist))
}
```

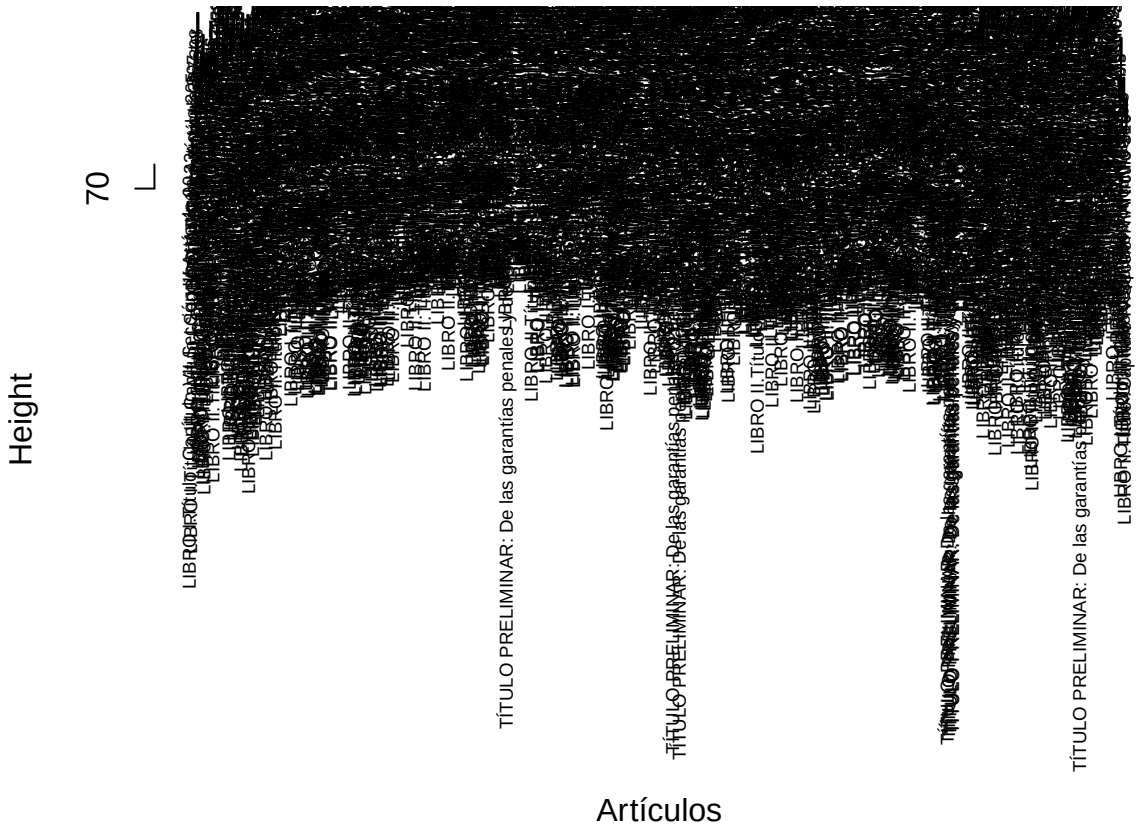
```
# ======================================================
# 2. CASO A: Solo texto del artículo
# ======================================================

# Crear DFM (limpieza básica: quitar puntuación y stopwords)
dfm_solo_texto <- tokens(corpus_cp, remove_punct = TRUE) %>%
  tokens_remove(stopwords("spanish")) %>%
  dfm()

# Calcular distancias euclídeas
dist_solo_texto <- dist(as.matrix(dfm_solo_texto), method = "euclidean")
```

```r
# Generar y guardar Dendrograma
plot(hclust(dist_solo_texto), main = "Dendrograma: Solo Texto del Artículo",
     xlab = "Artículos", sub = "", cex = 0.6)
```




```r
sim_A <- obtener_mas_similares(dist_solo_texto)
```

```r
# ======================================================
# 3. CASO B: Texto del artículo + contexto concatenado
# ======================================================

# Crear un nuevo corpus con el texto combinado
texto_combinado <- paste(as.character(corpus_cp), docvars(corpus_cp, "contexto"))
corpus_combinado <- corpus(texto_combinado, docnames = docnames(corpus_cp))

# Crear DFM
dfm_contexto <- tokens(corpus_combinado, remove_punct = TRUE) %>%
  tokens_remove(stopwords("spanish")) %>%
  dfm()

# Calcular distancias euclídeas
dist_contexto <- dist(as.matrix(dfm_contexto), method = "euclidean")

# Generar y guardar Dendrograma
plot(hclust(dist_contexto), main = "Dendrograma: Texto + Contexto",
     xlab = "Artículos", sub = "", cex = 0.6)
```




```
sim_B <- obtener_mas_similares(dist_contexto)
```

```
# ========================================================
# 4. SALIDA: Archivo .txt y Consola
# ========================================================

resultados_txt <- paste0(
  "RESULTADOS DE SIMILARIDAD\n",
  "==========================\n\n",
  "CASO A: Solo texto del artículo\n",
  "Artículos más similares: ", sim_A$art1, " y ", sim_A$art2, "\n",
  "Distancia euclídea mínima: ", round(sim_A$valor, 4), "\n\n",
  "--------------------------\n\n",
  "CASO B: Texto + Contexto concatenado\n",
  "Artículos más similares: ", sim_B$art1, " y ", sim_B$art2, "\n",
  "Distancia euclídea mínima: ", round(sim_B$valor, 4), "\n"
)

writeLines(resultados_txt, "articulos_parecidos.txt")
cat(resultados_txt) # Mostrar en consola
```

```
## RESULTADOS DE SIMILARIDAD
## ==========================
##
## CASO A: Solo texto del artículo
```

```
## Artículos más similares: LIBRO I.Título III.Capítulo I.Sección V.Artículo 55.1 y LIBRO I.Título III.C
## Distancia euclídea mínima: 0
##
## --------------------------
##
## CASO B: Texto + Contexto concatenado
## Artículos más similares: LIBRO I.Título III.Capítulo I.Sección V.Artículo 55.1 y LIBRO I.Título III.C
## Distancia euclídea mínima: 0
```