# Practica_PLN_Grupo08

2025-12-04

## Práctica 8

**CUIDADO AL EJECUTAR (Limpia el Global Enviroment)**

```
rm(list = ls())
```

### Lectura del archivo 10000 palabras

```
## Lectura del html y transformación a dataframe

# Leer el archivo
lineas <- readLines("10000_formas_ortograficas.txt",
                    encoding = "UTF-8")
lineas[0:4]
```

```
## [1] "Forma\tFrecuencia\tFrec. norm." ",\t27823866\t56483.65"
## [3] "de\t26286396\t53362.52"         ".\t19226627\t39030.88"
```

```
# Buscar donde empieza la priemra linea
linea_inicio <- grep("^[^0-9\t\n]+\t[0-9]+\t[0-9]+", lineas)[1]

# Extraer 10000 líneas a partir del inicio
datos <- lineas[linea_inicio:(linea_inicio + 9999)]

# Separar cada línea por tabulador (\t)
partes <- strsplit(datos, "\t")
partes[0:4]
```

```
## [[1]]
## [1] ","        "27823866" "56483.65"
##
## [[2]]
## [1] "de"       "26286396" "53362.52"
##
## [[3]]
## [1] "."        "19226627" "39030.88"
##
## [[4]]
## [1] "la"       "15799962" "32074.6"
```

```r
# Extraer las columnas
formas <- sapply(partes, function(x) x[1])
frecuencias <- as.numeric(sapply(partes, function(x) x[2]))
frec_norm <- as.numeric(sapply(partes, function(x) x[3]))

# Crear el dataframe
tabla <- data.frame(Forma = formas,
                    Frecuencia = frecuencias,
                    Frec.norm = frec_norm,
                    stringsAsFactors = FALSE)
head(tabla)
```

```
##   Forma Frecuencia Frec.norm
## 1     ,   27823866  56483.65
## 2    de   26286396  53362.52
## 3     .   19226627  39030.88
## 4    la   15799962  32074.60
## 5   que   13350795  27102.69
## 6     y   11562228  23471.82
```

```r
tail(tabla)
```

```
##                 Forma Frecuencia Frec.norm
## 9995       militancia       3422      6.94
## 9996         perciben       3421      6.94
## 9997         católico       3421      6.94
## 9998        convertía       3421      6.94
## 9999  especialización       3421      6.94
## 10000          golpeó       3421      6.94
```

## Apartado 1.1

## Apartado 1.2

```r
## Apartado 1.2

# Pasar todo a minúsculas
tabla$min_forma <- tolower(tabla$Forma)

# Contar cuántas formas básicas tienen variantes
conteo <- table(tolower(tabla$Forma))
duplicados <- conteo[conteo > 1]

cat("Formas básicas con duplicados no exactos:", length(duplicados), "\n")
```

```
## Formas básicas con duplicados no exactos: 760
```

```r
# Mostrar primeros 5 ejemplos
cat("\nPrimeros 5 ejemplos:\n")
```

```
##
## Primeros 5 ejemplos:
```

```
for (i in 1:5) {
  forma <- names(duplicados)[i]
  variantes <- unique(tabla$Forma[tolower(tabla$Forma) == forma])
  cat(i, " '", forma, "' con las varientes: ", paste(variantes, collapse = ", "), "\n", sep = "")
}
```

```
## 1 'a' con las varientes: a, A
## 2 'abierto' con las varientes: abierto, Abierto
## 3 'acá' con las varientes: acá, Acá
## 4 'academia' con las varientes: Academia, academia
## 5 'acaso' con las varientes: acaso, Acaso
```

### Apartado 1.3

```
# Función para detectar si la palabra contiene caracter no español
tiene_no_espanoles <- function(texto) {
  grepl("[^A-Za-zñÑáéíóúÁÉÍÓÚüÜ .,;:¡!¿?--]", texto)
}

no_espanol <- tiene_no_espanoles(tabla$Forma)

# Número de palabras con caracteres no españoles (según nuestro criterio)
num_no_espanol <- sum(no_espanol)
cat("El nº de formas no españolas según nuestro criterio es: ", num_no_espanol)
```

```
## El nº de formas no españolas según nuestro criterio es:  224
```

```
#Ejemplos
head(tabla$Forma[no_espanol],20)
```

```
##  [1] "\"" ")"  "("  "»"  "«"  "'"  "1"  "2"  """  """  "3"  "4"  "-"  "10" "5"
## [16] "6"  "20" "15" "7"  "12"
```