Parte 5

Adrián Martín Marín

2026-01-10

```
# Cargar el corpus
corpus_cp <- readRDS("corpus_codigo_penal.rds")
# Cargar librerías
library(quanteda)
library(spacyr)
library(dplyr)
spacy_initialize(model = "es_core_news_sm")
library(udpipe)
```

```
## Warning: package 'udpipe' was built under R version 4.5.2
```

```
modelo_ud <- udpipe_download_model(language = "spanish")
ud_model <- udpipe_load_model(modelo_ud$file_model)
```

Primero extraemos los documentos del corpus y buscamos en ellos las entidades nombradas con spacyr.

```
# Extraer los documentos del corpus
textos <- as.character(corpus_cp)
# Buscamos entidades nombradas con spacyr
ent_spacy <- spacy_extract_entity(textos, type = "named")
```

Ahora mostramos las 20 entidades más frecuentes que se han encontrado.

```
top20_spacy <- count(ent_spacy, text, sort = TRUE)
top20_spacy <- slice_head(top20_spacy, n = 20)
top20_spacy
```

```
##                text   n
## 1                 1 299
## 2           España 112
## 3           Código 104
## 4             Juez  74
## 5             Será  68
## 6         Argentina  45
## 7           Bolivia  45
## 8             Chile  45
## 9          Colombia  45
## 10          Ecuador  45
## 11        Guatemala  45
## 12        Instagram  45
```

```
## 13             Perú  45
## 14           TikTok  45
## 15              USA  45
## 16        Venezuela  45
## 17 Seguridad Social  43
## 18         Tribunal  42
## 19         Capítulo  39
## 20   Administración  28
```

Ahora procesamos los documentos con udpipe y buscamos las entidades nombradas con esta librería.

```
# Procesamos los documentos
texto_udpipe <- udpipe_annotate(ud_model, x = textos)
texto_udpipe_df <- as.data.frame(texto_udpipe)
# Buscamos entidades nombradas con udpipe
ent_udpipe <- keywords_rake(x = texto_udpipe_df,term = "lemma", group = "doc_id",
                            relevant = texto_udpipe_df$upos == "PROPN")
```

Mostramos las 20 entidades nombradas más frecuentes encontradas por udpipe.

```
top20_udpipe <- count(ent_udpipe, keyword, sort = TRUE)
top20_udpipe <- slice_head(top20_udpipe, n = 20)
top20_udpipe
```

```
##               keyword n
## 1             Cookies 1
## 2           Instagram 1
## 3              Leyes 1
## 4             Oculte 1
## 5             TikTok 1
## 6             abogado 1
## 7              agosto 1
## 8           argentina 1
## 9            artículo 1
## 10           asturias 1
## 11             ataque 1
## 12            bolivia 1
## 13              chile 1
## 14           colombia 1
## 15            cometer 1
## 16 comunidad autónoma 1
## 17          condición 1
## 18           congreso 1
## 19            consejo 1
## 20    consejo general 1
```

Por último buscamos qué entidades coinciden en los resultados dados por ambas librerías. Veremos que son pocas, y esto se debe a que la manera de filtrar el texto de cada librería es distinta, ya que por un lado spacyr sí que está entrenado para encontrar entidades nombradas como tal (y por tanto es el modelo más adecuado) mientras que udpipe solo hace la búsqueda según la categoría gramatical de la palabra (en este caso nombres propios, aunque incluso comete algún error en el filtrado e incluye palabras de otras categorías).

```r
spacy_vec <- tolower(top20_spacy$text) # Lo convertimos todo a minúsculas para evitar errores
udpipe_vec <- tolower(top20_udpipe$keyword)
comunes <- intersect(spacy_vec, udpipe_vec) # Encontramos las entidades comunes
comunes
```

```
## [1] "argentina" "bolivia"   "chile"     "colombia"  "instagram" "tiktok"
```