Segundo Punto - Taller 3

Andrés Díaz - Yilmer Palacios

2024-04-30

## R Markdown

### Preparamos librerías

rm(list = ls())  
  
librerias <- (c("rvest", "purrr", "tidyverse","tokenizers","syuzhet","wordcloud",  
 "tm","cowplot","ggwordcloud","modelsummary","estimatr","sandwich",  
 "writexl","readxl","quanteda","countries","sf"))  
  
if(length(setdiff(librerias, rownames(installed.packages()))) > 0){  
 install.packages(setdiff(librerias, rownames(installed.packages())))}  
invisible(sapply(librerias, require, character.only = TRUE,quietly = TRUE))

## Warning: package 'rvest' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ readr::guess\_encoding() masks rvest::guess\_encoding()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

## Warning: package 'tokenizers' was built under R version 4.3.3

## Warning: package 'syuzhet' was built under R version 4.3.3

## Warning: package 'wordcloud' was built under R version 4.3.3

## Warning: package 'tm' was built under R version 4.3.3

##   
## Attaching package: 'NLP'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## annotate

## Warning: package 'cowplot' was built under R version 4.3.3

##   
## Attaching package: 'cowplot'  
##   
## The following object is masked from 'package:lubridate':  
##   
## stamp

## Warning: package 'ggwordcloud' was built under R version 4.3.3

## Warning: package 'modelsummary' was built under R version 4.3.3

## `modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing  
## backend. Learn more at: https://vincentarelbundock.github.io/tinytable/  
##   
## Revert to `kableExtra` for one session:  
##   
## options(modelsummary\_factory\_default = 'kableExtra')  
##   
## Change the default backend persistently:  
##   
## config\_modelsummary(factory\_default = 'gt')  
##   
## Silence this message forever:  
##   
## config\_modelsummary(startup\_message = FALSE)

## Warning: package 'estimatr' was built under R version 4.3.3

## Warning: package 'sandwich' was built under R version 4.3.3

## Warning: package 'writexl' was built under R version 4.3.3

## Warning: package 'readxl' was built under R version 4.3.3

## Warning: package 'quanteda' was built under R version 4.3.3

## Package version: 4.0.2  
## Unicode version: 15.1  
## ICU version: 74.1  
## Parallel computing: 8 of 8 threads used.  
## See https://quanteda.io for tutorials and examples.  
##   
## Attaching package: 'quanteda'  
##   
## The following object is masked from 'package:tm':  
##   
## stopwords  
##   
## The following objects are masked from 'package:NLP':  
##   
## meta, meta<-

## Warning: package 'countries' was built under R version 4.3.3

## Warning: package 'sf' was built under R version 4.3.3

## Linking to GEOS 3.11.2, GDAL 3.8.2, PROJ 9.3.1; sf\_use\_s2() is TRUE

# SEGUNDO PUNTO

##2.1. Realicen un código de R que realice webscraping de Wikipedia para extraer la información de los conflictos desde 1945–1989. Este código debe tener como resultado un DataFrame que contenga una columna con el nombre oficial del país que tuvo conflicto con USA (recuerde que puede ser vencedor o perdedor).

### Pista: esto lo puede hacer identificando los nombres únicos que aparecen en conflictos y creando un Dataframe posteriormente con esta lista. Por otra parte, pueden obtener una lista de los nombres de los países del mundo en internet, Wikipedia, librerías como “countries” en R. Además, pueden basarse en el código realizado en clase.

# Definimos la URL de interés de Wikipedia (Lista de Guerras de 1945 - 1989)  
url <- "https://en.wikipedia.org/wiki/List\_of\_wars:\_1945%E2%80%931989"  
  
  
batallas <- url %>% read\_html() %>% # Hacemos Webscrapping de las tablas  
 html\_table(fill = TRUE)  
  
# Hacemos procesamiento de la data  
  
df <- batallas %>% reduce(bind\_rows)

## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## • `Belligerents` -> `Belligerents...4`  
## • `Belligerents` -> `Belligerents...5`

df <- df %>% select(c(4,5)) %>% na.omit() # Borramos las columbas innecesarias y las filas con NA  
colnames(df) <- c("Victoria", "Derrota") # Ajustamos nombres de columna  
  
  
convertidor <- function(Vector){  
   
 # Esta función toma un vector con elementos de texto, identifica si los elementos tienen las  
 # Palabras "United" y "States", si si, cambia toda la cadena de texto por la palabra United States para  
 # Eliminar a los aliados de Estados Unidos de la base de datos.  
   
 ##Inputs  
 #bd\_libros (Vector): Un Vector a tratar  
 ##Output  
 #bd\_libros (dataframe): Vector tratado  
   
   
 USA <- c("United", "States")  
 df.corpus <- corpus(Vector)  
 df.tokens <- tokens(df.corpus, what = "word")  
 conflictosUSA <- tokens\_select(df.tokens, pattern = USA, selection = "keep")  
 out <- convert(dfm(conflictosUSA), to = "data.frame") # Convertimos lista como dataframe  
 out[,4] <- ifelse(out[,2] > 0 & out[,3] > 0, TRUE, FALSE) # Creamos booleano que es TRUE si ambas palabras son "United" "States"  
 out[,5] <- Vector # Agregamos columna con texto concatenado  
 out <- replace(out$V5, out$V4 == TRUE, "United States")  
  
return(out)  
  
}  
  
Vector1 <- df$Victoria  
Vector2 <- df$Derrota  
df[,1] <- convertidor(Vector1)  
df[,2] <- convertidor(Vector2)   
  
  
df[,3] <- paste(df$Victoria, df$Derrota)  
colnames(df) <- c("Victoria", "Derrota", "Concatenado")  
  
# Identificamos las filas donde no se mencione a USA  
  
USA <- c("United", "States")  
colnames(df)

## [1] "Victoria" "Derrota" "Concatenado"

df.corpus <- corpus(df, text\_field = "Concatenado" )  
df.tokens <- tokens(df.corpus, what = "word")  
conflictosUSA <- tokens\_select(df.tokens, pattern = USA, selection = "keep")  
  
# Creamos vector que indique en que filas no se menciona a USA   
#(por alguna razón unlist no nos guarda como dataframe los nombres de las filas del vector entonces  
# tuvimos que seguir otra ruta)  
  
out <- convert(dfm(conflictosUSA), to = "data.frame") # Convertimos lista como dataframe  
out[,4] <- ifelse(out[,2] > 0 & out[,3] > 0, TRUE, FALSE) # Creamos booleano que es TRUE si el vector contiene ambas palabras "United" "States"  
out[,5] <- df$Concatenado # Agregamos columna con texto concatenado  
outfiltrado <- out[out$V4 == TRUE, ] # Eliminamos filas con FALSE  
outfiltrado <- outfiltrado %>% select(c(5)) %>% na.omit() # Dejamos solo la fila de interés  
colnames(outfiltrado) <- "Texto Concatenado"  
  
# Volvemos a identificar palabras para detectar a los países con la función country\_name  
  
Outcorpus <- corpus(outfiltrado, text\_field = "Texto Concatenado" )  
out.tokens <- tokens(Outcorpus, what = "word")  
finalnames <- as.data.frame(unlist(out.tokens)) # Listado de todas las palabras  
  
### Con la función identificamos los nombres de los países y los guardamos con notación a ISO 3166 - Lo del bono ;)  
paises <- country\_name(finalnames[,1])

## Some country IDs have no match in one or more of the requested country naming conventions, NA returned.  
## Multiple country IDs have been matched to the same country name.  
## There is low confidence on the matching of some country names, NA returned.  
##   
## Set - verbose - to TRUE for more details

paises <- paises %>% na.omit()  
paises <- as.data.frame(paises)  
paises <- paises %>% distinct(paises, .keep\_all = TRUE) # Borramos duplicados  
  
## Dataframe con el listado de paises que estuvieron   
  
head(paises)

## paises  
## 1 ZAF  
## 2 AND  
## 3 CHN  
## 4 BLR  
## 5 SVK  
## 6 USA

## 2.2. Descarguen el shapefile de las fronteras de los países del mundo geoBoundaries y realicen un pegue (left\_join) de los datos de los países del shapefile con la base consolidada por webscraping. Para el pegue, deben utilizar el nombre del país.

###Nota: Traten de revisar que los nombres coincidan para la mayoría de los países, pero no se penalizará si no ocurre un match perfecto entre las bases de datos o si por escritura omiten alguno.

# Cargamos el ShapeFile  
  
setwd("C:/Users/Yilmer Palacios/Desktop/Solución Talleres R/Taller-3---R/")  
sf\_data <- st\_read(dsn= "Shape")

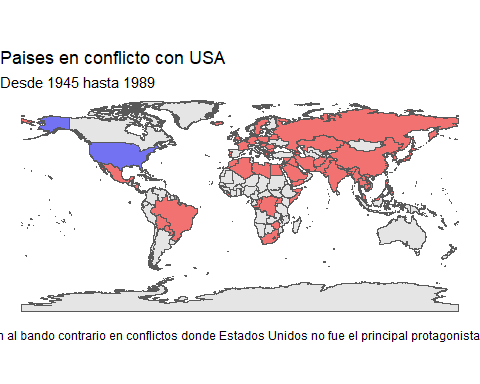
## Reading layer `geoBoundariesCGAZ\_ADM0' from data source   
## `C:\Users\Yilmer Palacios\Desktop\Solución Talleres R\Taller-3---R\Shape'   
## using driver `ESRI Shapefile'  
## Simple feature collection with 218 features and 3 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: -180 ymin: -90 xmax: 180 ymax: 83.63339  
## Geodetic CRS: WGS 84

# Realizamos el left join en la columna "paises"  
colnames(paises) <- "shapeGroup"  
ResLJ2 <- left\_join(paises, sf\_data, by = "shapeGroup")  
head(ResLJ2)

## shapeGroup shapeType shapeName geometry  
## 1 ZAF ADM0 South Africa MULTIPOLYGON (((37.72466 -4...  
## 2 AND ADM0 Andorra MULTIPOLYGON (((1.725802 42...  
## 3 CHN ADM0 China MULTIPOLYGON (((110.7046 20...  
## 4 BLR ADM0 Belarus MULTIPOLYGON (((28.15123 56...  
## 5 SVK ADM0 Slovakia MULTIPOLYGON (((17.16081 48...  
## 6 USA ADM0 United States MULTIPOLYGON (((-168.1579 -...

## 2.3. Presenten una visualización en el mapa del mundo de los países con los que Estados Unidos ha tenido conflictos armados durante esta época. El gráfico debe reflejar estética visual y ser autocontenido. Pueden añadirle información adicional de otros tipos de variables si quieren.

# Creamos un elemento sf a partir de nuestra base con el left joint  
paises\_georef <- st\_as\_sf(ResLJ2)  
  
# Construirmos la gráfica  
ggplot() + geom\_sf(data = sf\_data) + geom\_sf(data = paises\_georef %>% filter(shapeGroup != "USA"), fill = "red", alpha = 0.5 ) +   
 geom\_sf(data = paises\_georef %>% filter(shapeGroup == "USA"), fill = "blue", alpha = 0.5 ) +   
 theme\_void() + labs(title = "Paises en conflicto con USA", subtitle = "Desde 1945 hasta 1989", caption = "Nota: Se incluyen países que apoyaron al bando contrario en conflictos donde Estados Unidos no fue el principal protagonista")

 ## Como se puede apreciar en el mapa, la mayoría de conflictos que tuvo USA en el siglo pasado son principalmente con los países comunistas de la época como Rusia y China, adicionalmente hay una serie de países que apoyaron a bandos contrarios a Estados Unidos por lo que aparecen como países con conflicto con USA, como en el caso de México que apoyo al bando contrario en la guerra civil de Angola, sin embargo la confrontación no fue directa con USA.