Proyecto - Modulo 1

Sergio Alejandro Paucara Saca

2024-08-03

Contents

Proyecto **1**

Web Scraping de "Los Tiempos" **1**

Web Scraping de "La Razón" **3**

Análisis de la Palabra Clave **4**

Carga de las Bases de Datos 4

Filtrado de Titulares con la Palabra Clave 4

Visualización de los Resultados 4

Cálculo del Porcentaje de Titulares 5

Gráficos **6**

Proyecto

Este proyecto realiza un web scraping de las últimas noticias de los sitios web "Los Tiempos" y "La Razón". Además, analiza la aparición de una palabra clave *(YPFB)* en los titulares de ambos medios.

Web Scraping de "Los Tiempos"

El siguiente código extrae los títulos, resúmenes, fechas y secciones de las últimas noticias de "Los Tiempos".

```
# Cargar las bibliotecas necesarias
library(stringr)
library(rvest)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(knitr)

# Inicialización de variables
todos_titulos <- c()
todos_resumenes <- c()
todas_fechas <- c()
todas_secciones <- c()

# Inicialización de variables
base_url<-"https://www.lostiempos.com/ultimas-noticias"

# Inicialización del contador de páginas
page <- 1

# Bucle para realizar el scraping de las primeras 8 páginas
while (page<9) {
  url<-sprintf("%s?page=%d", base_url, page)
  read_wp<-read_html(url)

  titulo<-read_wp %>% html_nodes("div.views-field-title.term-\\[tid\\]") %>% html_text2()
  resumen<-read_wp %>% html_nodes("div.views-field.views-field-field-noticia-sumario") %>% html_text2()
  fecha <-read_wp %>% html_nodes("span.views-field-field-noticia-fecha") %>% html_text2()
  seccion <-read_wp %>% html_nodes("span.views-field-seccion") %>% html_text2()

  todos_titulos <- c(todos_titulos, titulo)
  todos_resumenes <- c(todos_resumenes, resumen)
  todas_fechas <- c(todas_fechas, fecha)
  todas_secciones <- c(todas_secciones, seccion)

  page<-page+1
}

# Crear un dataframe con los datos extraídos
db_lostiempos<-data.frame(
  titulo=todos_titulos,
  resumen=todos_resumenes,
  fecha=todas_fechas,
  seccion=todas_secciones)

# Guardar el dataframe en un archivo .RData
# save(db_lostiempos, file = "_data/db_lostiempos.RData")
```

Web Scraping de "La Razón"

El siguiente código extrae los títulos, resúmenes, categorías, autores, fechas y enlaces de las noticias de "La Razón".

```
# Inicialización de variables
titulos_larazon <- c()
resumenes_larazon <- c()
categoria_larazon <- c()
autor_larazon <- c()
fechas_larazon <- c()
link_larazon <- c()

# Definir la URL base
base_url_larazon <- "https://www.la-razon.com/economia"

# Inicialización del contador de páginas
page_lr <- 2

while (page_lr<4) {

  url_lr<-sprintf("%s/page/%d", base_url_larazon, page_lr)

  read_ny<-read_html(url_lr)

  titulo_lr<-read_ny %>% html_nodes("div.article-meta > a") %>% html_text2()
  resumen_lr<-read_ny %>% html_nodes("div.article-meta > p") %>% html_text2()
  categoria_lr<-read_ny %>% html_nodes("div.article-meta > div > p > a") %>% html_text2()
  autor_lr<-read_ny %>% html_nodes("span.author > a") %>% html_text2()
  fecha_lr<-read_ny %>% html_nodes("span.date > a") %>% html_text2()
  link_lr<-read_ny %>% html_nodes("div.article-meta > a") %>% html_attr("href")


  titulos_larazon <- c(titulos_larazon, titulo_lr)
  resumenes_larazon <- c(resumenes_larazon, resumen_lr)
  categoria_larazon <- c(categoria_larazon, categoria_lr)
  autor_larazon <- c(autor_larazon, autor_lr)
  fechas_larazon <- c(fechas_larazon, fecha_lr)
  link_larazon <- c(link_larazon, link_lr)

  page_lr<-page_lr+1
}


db_larazon<-data.frame(
  titulo=titulos_larazon[1:10],
  resumen=resumenes_larazon[1:10],
  categoria=categoria_larazon[1:10],
  autor=autor_larazon[1:10],
  fecha=fechas_larazon[1:10],
  link=link_larazon[1:10]
  )

# save(db_larazon, file = "_data/db_larazon.RData")
```

Análisis de la Palabra Clave

En esta sección se analiza la aparición de la palabra clave "YPFB" en los titulares de las noticias extraídas de los sitios web "Los Tiempos" y "La Razón".

Carga de las Bases de Datos

Primero, cargamos los dataframes previamente guardados de "Los Tiempos" y "La Razón".

```
# Eliminar todos los objetos del entorno para evitar conflictos
rm(list = ls())

# Cargar los dataframes de "Los Tiempos" y "La Razón"
load("_data/db_lostiempos.RData")
load("_data/db_larazon.RData")
```

Filtrado de Titulares con la Palabra Clave

Definimos la palabra clave "YPFB" y filtramos los titulares que contienen esta palabra en ambas bases de datos.

```
# Definir la palabra clave a buscar
palabra_clave <- "YPFB"

# Filtrar los titulares que contienen la palabra clave en "Los Tiempos"
titulares_con_palabra_lostiempos <- db_lostiempos[grepl(palabra_clave, db_lostiempos$titulo), ]

# Filtrar los titulares que contienen la palabra clave en "La Razón"
titulares_con_palabra_larazon <- db_larazon[grepl(palabra_clave, db_larazon$titulo), ]
```

Visualización de los Resultados

Mostramos los primeros 3 titulares que contienen la palabra clave en cada uno de los sitios web.

```
# Mostrar los primeros 3 titulares que contienen la palabra clave en "Los Tiempos"
kable(titulares_con_palabra_lostiempos[1:3, ], caption = "Titulares con la palabra clave 'YPFB' en Los
```

Table 1: Titulares con la palabra clave 'YPFB' en Los Tiempos

	titulo	resumen	fecha	seccion
26	YPFB pide que se levanten los bloqueos para que las cisternas con diésel lleguen a los surtidores	El presidente de Yacimientos Petrolíferos Fiscales Bolivianos (YPFB), Armin Dorgathen, afirmó este viernes que el diésel ya está en proceso de descarga en el puerto de Sica Sica, en Arica, por lo que pidió al transporte pesado que levante los bloqueos en las carreteras para que las cisternas puedan llegar a los surtidores.	02/08/2024 - 09:49	Economía

	titulo	resumen	fecha	seccion
84	YPFB dice que cisternas con diésel pueden salir de Arica esta noche y pide a bloqueadores dar paso	Yacimientos Petrolíferos Fiscales de Bolivia (YPFB) confirmó este jueves que un buque ya comenzó el proceso de descarga de diésel en el puerto de Sica Sica, en Arica, y pidió que se dejen pasar a las cisternas que están paradas en los puntos de bloqueo.	01/08/2024 - 14:44	Política
85	Bloqueos en Tambo Quemado impiden el ingreso de 297 cisternas con combustible según YPFB	Dorgathen aclaró que el sábado en la tarde se espera tener las cisternas en las plantas del país, repartiendo el combustible a las estaciones de servicio, siempre que no se presenten nuevos bloqueos. “El combustible está ingresando mediante buque y cisternas desde cuatro países, Paraguay, Argentina, Perú y mediante Chile a partir de las 22:00 de la noche”, explicó.	01/08/2024 - 14:41	Política

```
# Mostrar los primeros 3 titulares que contienen la palabra clave en "La Razón"
kable(titulares_con_palabra_larazon[1:3, ], caption = "Titulares con la palabra clave 'YPFB' en La Razó
```

Table 2: Titulares con la palabra clave ‘YPFB’ en La Razón

	titulo	resumen	categoria	autor	fecha	link
3	YPFB y Vintage suscriben protocolos para exploración en Sayurenda, Yuarenda y Carandaiti	De efectuarse el descubrimiento comercial esperado, se incrementarían las reservas . . .	Economía	Yuri Flores	27 de julio de 2024	https://www.la-razon.com/economia/2024/07/27/ypfb-y-vintage-suscriben-protocolos-para-exploracion-en-sayurenda-yuarenda-y-carandaiti/
14	YPFB aprueba inversión de $us 250 millones para tres nuevos pozos en el Mayaya	Dos de los pozos exploratorios están asociados al proyecto Mayaya . . .	Sociedad	Erika Ibáñez	26 de julio de 2024	https://www.la-razon.com/economia/2024/07/26/ypfb-aprueba-inversion-de-us-250-millones-para-tres-nuevos-pozos-en-el-mayaya/
15	YPFB prevé que ingreso de combustible por Arica e hidrovía genere una autonomía de 25 días	Se prevé que el fin de semana se pueda habilitar . . .	Economía	Yuri Flores	26 de julio de 2024	https://www.la-razon.com/economia/2024/07/26/ypfb-preve-que-ingreso-de-combustible-por-arica-e-hidrovia-genere-una-autonomia-de-25-dias/

Cálculo del Porcentaje de Titulares

Calculamos el porcentaje de titulares que contienen la palabra clave en ambas bases de datos y mostramos los resultados.

```
# Calcular el porcentaje de titulares que contienen la palabra clave en "Los Tiempos"
porcentaje_lostiempos <- (nrow(titulares_con_palabra_lostiempos) / nrow(db_lostiempos)) * 100

# Calcular el porcentaje de titulares que contienen la palabra clave en "La Razón"
porcentaje_larazon <- (nrow(titulares_con_palabra_larazon) / nrow(db_larazon)) * 100

# Mostrar el porcentaje en "Los Tiempos"
sprintf("El porcentaje de titulares que contienen la palabra clave '%s' es: %.2f%% en Los Tiempos", pala
```

```
## [1] "El porcentaje de titulares que contienen la palabra clave 'YPFB' es: 2.63% en Los Tiempos"
```

```
# Mostrar el porcentaje en "La Razón"
sprintf("El porcentaje de titulares que contienen la palabra clave '%s' es: %.2f%% en La Razón", palabra
```

```
## [1] "El porcentaje de titulares que contienen la palabra clave 'YPFB' es: 5.10% en La Razón"
```

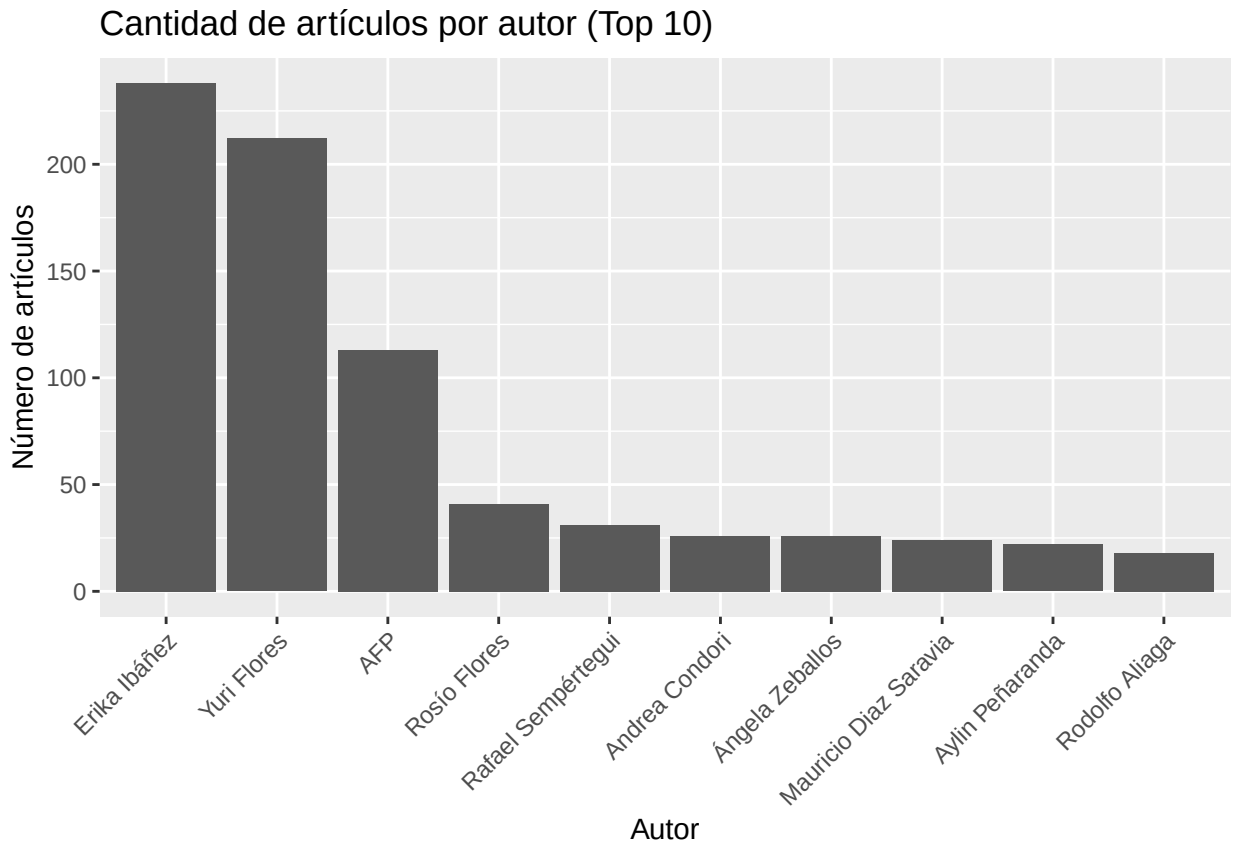
Gráficos

```
library(lubridate)
```

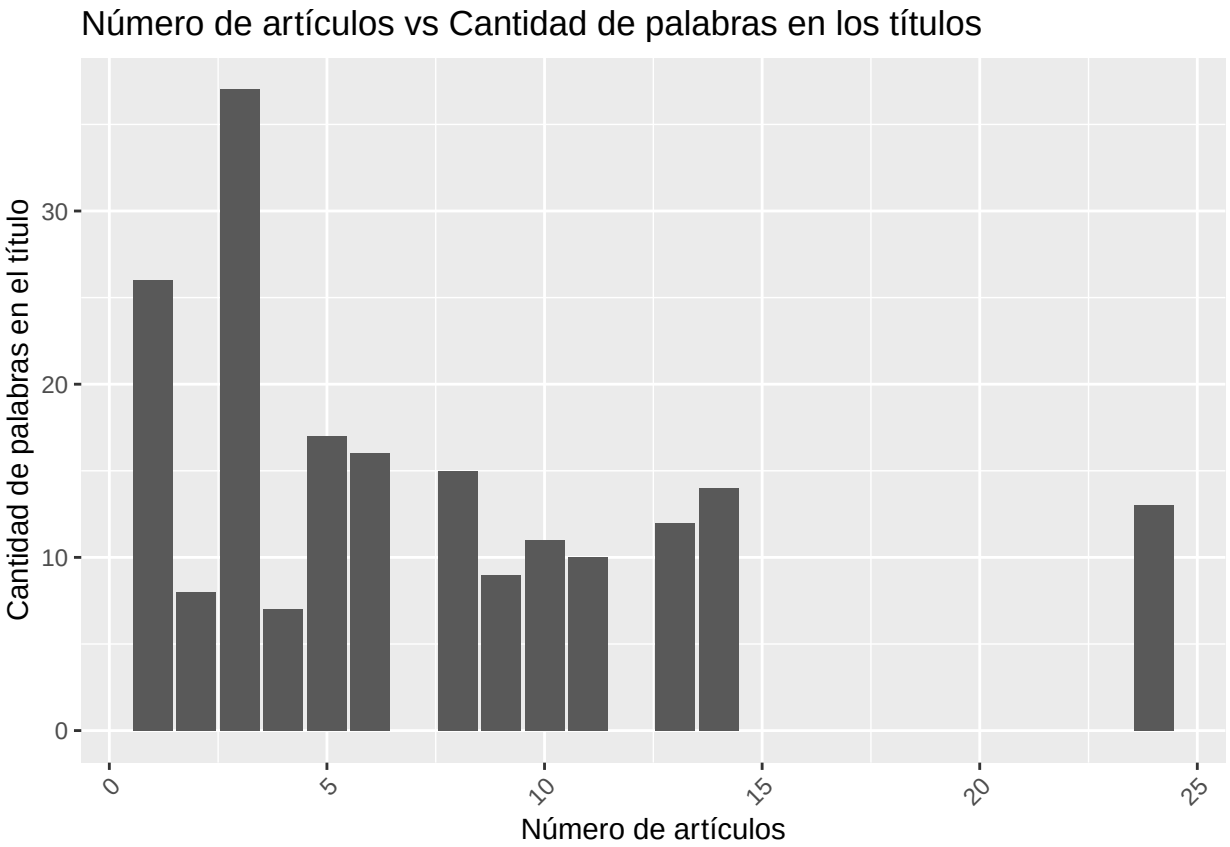
```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)

# Cargar los datos de los periódicos
df1<-db_larazon
df2<-db_lostiempos

# Análisis de los artículos por autor en La Razón
articulos_por_autor <- df1 %>%
  group_by(autor) %>%
  summarise(cantidad_articulos = n()) %>%
  arrange(desc(cantidad_articulos)) %>%
  slice(1:10)

# Crear la gráfica de la cantidad de artículos por los 10 autores más productivos
ggplot(articulos_por_autor, aes(x = reorder(autor, -cantidad_articulos), y = cantidad_articulos)) +
  geom_bar(stat = "identity") +
  labs(title = "Cantidad de artículos por autor (Top 10)",
       x = "Autor",
       y = "Número de artículos") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Análisis de la longitud de los títulos en Los Tiempos
df2 <- df2 %>%
  mutate(num_palabras_titulo = str_count(titulo, "\\S+"))

# Agrupar por número de palabras y contar el número de artículos que tienen esa cantidad de palabras
conteo_palabras <- df2 %>%
  group_by(num_palabras_titulo) %>%
  summarise(num_articulos = n())

# Crear la gráfica del número de artículos por cantidad de palabras en los títulos
ggplot(conteo_palabras, aes(x = num_articulos, y = num_palabras_titulo)) +
  geom_bar(stat = "identity") +
  labs(title = "Número de artículos vs Cantidad de palabras en los títulos",
       x = "Número de artículos",
       y = "Cantidad de palabras en el título") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Número de artículos vs Cantidad de palabras en los títulos
Cantidad de palabras en el título
30
20
10
0
0
5
10
15
20
25
Número de artículos