

Programación y Uso de Librerías en R: Herramientas de Análisis y Visualización de Datos en la Enseñanza y la Investigación Científica

Juan Luis Peñaloza Figueroa
Universidad Complutense de Madrid

Milagros Dones Tacero
Universidad Autónoma de Madrid

Carmen Gladys Vargas Pérez
Universidad Complutense de Madrid

AÑO: 2025

SCRIPT_6: CAPITULO VIII: SCRAPING AND EXTRACTING DATA

```
#Instalar librerías necesarias
> library(tidyverse)
> library(polite)
> library(lubridate)
> library(wordcloud)
> library(tidytext)
> library(tm)
> library(jcolors)
> install.packages("rvest")
> library(rvest)

# Usamos la función rvest para obtener la HTML de un sitio web
> url <- "https://scrapingclub.com/exercise/detail_basic/"
> webpage <- rvest::read_html(url)
> print(webpage)

# Extraer datos
img_src <- webpage %>% html_nodes("img.card-img-top") %>% html_attr("src")
# Fuente de la imagen
> title <- webpage %>% html_nodes("h3") %>% html_text() # Título del producto
> price <- webpage %>% html_nodes("h4") %>% html_text() # Precio del producto
> description <- webpage %>% html_nodes("p.card-description") %>% html_text()

# Mostrar los datos extraídos
> print(img_src)
> print(title)
> print(price)
> print(description)

# Scraping o raspado web de la URL de Amazon
# R la versión actualizada
> Rvest actualizada
> tidyverse actualizada
> HtmlLink<-"https://www.amazon.co.uk/Xbox-Elite-Wireless-
> Controller-2/dp/B07SR4R8K1/ref=sr_1_1_sspa?crid=3F4M36E0LDQF3"
# Extrae el número de identificación estándar de Amazon
> ASIN <- str_match(HtmlLink, "/dp/([A-Za-z0-9]+)/")[,2]
# Descarga el contenido HTML de la página web
> HTMLContent <- read_html(HtmlLink)
> review_title <- HTMLContent %>%
```

```

> html_nodes("div:nth-child(2)a.a-size-base.a-link-
ormal.reviewtitle.a-color-base.review-title-content.a-text-bold
span") %>% html_text()
# Extracción del cuerpo de la reseña
> review_body <- HTMLContent %>% html_nodes("div.a-row.a-spacing-
small.review-data span div.a-expander-content.reviewText.review-text-
content.aexpander-partial-collapse-content span") %>% html_text()
# Valoración de la reseña
> review_rating <- HTMLContent %>% html_nodes("div:nth-child(2)
a:nth-child(1) i.review-rating span") %>% html_text()
# Escalado a múltiples URLs
> install.packages("rcrossref")
> library(rcrossref)
# Ver como se cita un determinado artículo
> my_doi <- "10.1111/j.1467-6486.2012.01072.x"
> cr_cn(dois = my_doi, format = "text", style = "apa")
> cr_cn(dois = my_doi, format = "bibtex", style = "apa",
locale = "en-US", raw = FALSE, progress = "none")
# Ver el número de citas de un artículo/DOI
> aa <- cr_citation_count(doi = my_doi)
# Ver el abstract de los artículos
> aa <- cr_abstract(doi = "10.1109/TASC.2010.2088091")
# Ver los journals
> aa <- cr_journals(query = "economics", limit = 100) %>%
.$data %>% as.tibble()
# Obtener más información de un artículo
> aa <- cr_works(dois = my_doi) %>% .$data %>% as.tibble()

```