

```
library(dplyr)
```

```
rladies_global %>%  
  filter(city == 'Buenos Aires')
```



Intro a Web Scraping

con



en



Spotify

Charts

Paola Prieto



@paoprieto



1. Qué es Web Scraping?

El web scraping es una técnica que sirve para extraer información de páginas web de forma automatizada para su posterior análisis.

2. Web Scraping Ética

Tener en cuenta

- No vulnerar los derechos de propiedad intelectual
- Considerar que no estemos incurriendo en una conducta de competencia desleal
- No incumplir las normativas en protección de datos



3. Web Scraping

Para qué se usa?

- Conocer las Tendencias del Momento
- Para la Compra/Venta de Servicios o Productos – Sector inmobiliario
- Análisis de las reviews de Productos
- Páginas de comparación de precios
- Listado de eventos
- Documentación Pasada, Útil a Día de Hoy - Abogacía
- En Recursos Humanos
- En tiempos de campañas electorales



4. Paquete rvest



Paquete para Web Scraping

rvest



Desarrollado por Hadley Wickham

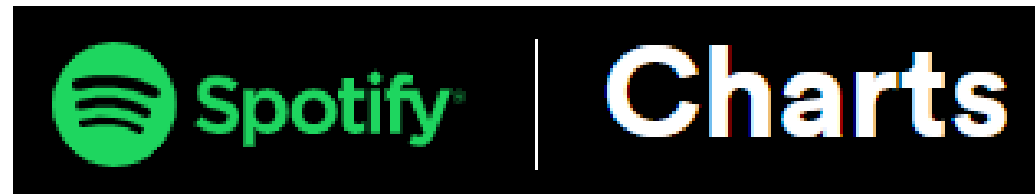
Funciones:

- **read_html()** Analiza gramaticalmente (parse) una página HTML.
- **html_nodes()** Selecciona nodos de un documento HTML.
- **html_text()** Extrae atributos, texto, nombres de tags de un HTML.

<https://CRAN.R-project.org/package=rvest>



4. SpotifyCharts HTML





Charts

↓ DOWNLOAD TO CSV

TOP 200

VIRAL 50

Filter by

ARGENTINA

DAILY

05/25/2019

TRACK

STREAMS ?

	1	Otro Trago by Sech	494,549
	2	Tal Vez by Paulo Londra	343,925
	3	Pa Mí - Remix by Dalex	275,260
	4	Soltera - Remix by Lunay	272,325
	5	La Cobra by j mena	254,160
	6	Con Altura by ROSALÍA	224,963
	7	Solo Pienso En Ti (feat. De La Ghetto & Justin Quiles) by Paulo Londra	208,745
	8	Con Calma by Daddy Yankee	204,618

Breve repaso de HTML

Tags de HTML

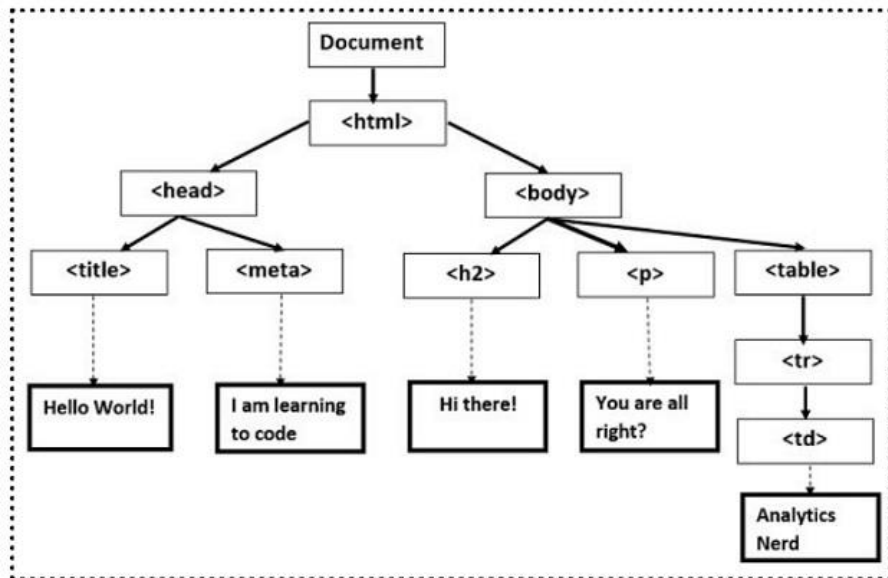


Tabla en HTML

`<TABLE HIEGHT=10 WIDTH=30 BORDER=0>`

	<code><TH> Header1 </TH></code>	<code><TH> Header2 </TH></code>	
<code><TR></code>	<code><TD> </TD></code>	<code><TD> </TD></code>	<code></TR></code>
<code><TR></code>	<code><TD> </TD></code>	<code><TD> </TD></code>	<code></TR></code>
<code><TR></code>	<code><TD> </TD></code>	<code><TD> </TD></code>	<code></TR></code>
<code><TR></code>	<code><TD> </TD></code>	<code><TD> </TD></code>	<code></TR></code>

`</TABLE>`



Charts

[DOWNLOAD TO CSV](#)

TOP 200

VIRAL 50

Filter by

ARGENTINA

DAILY

05/25/2019

TRACK

STREAMS ?

	1	— Otro Trago by Sech	494,549
	2	— Tal Vez by Paulo Londra	343,925
	3	— Pa Mí - Remix by Dalex	
	4	— Soltera - Remix by Lunay	
	5	— La Cobra by j mena	
	6	▲ Con Altura by ROSALÍA	
	7	▼ Solo Pienso En Ti (feat. De L	
	8	— Con Calma by Daddy Yanke	

view-source:https://spotifycharts.com/regional/ar/daily/2019-05-25

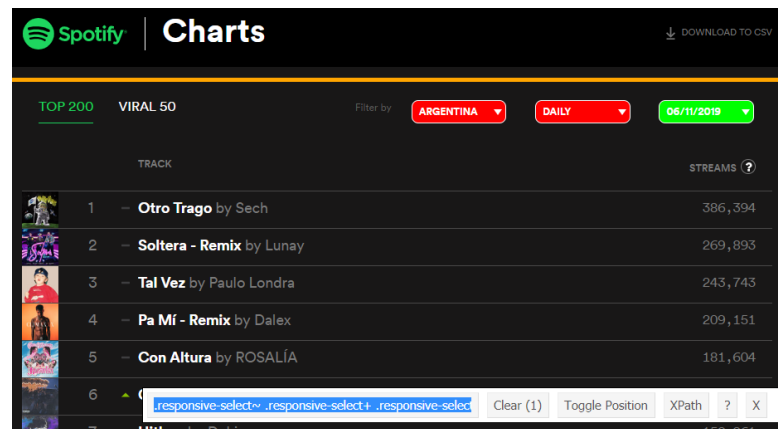
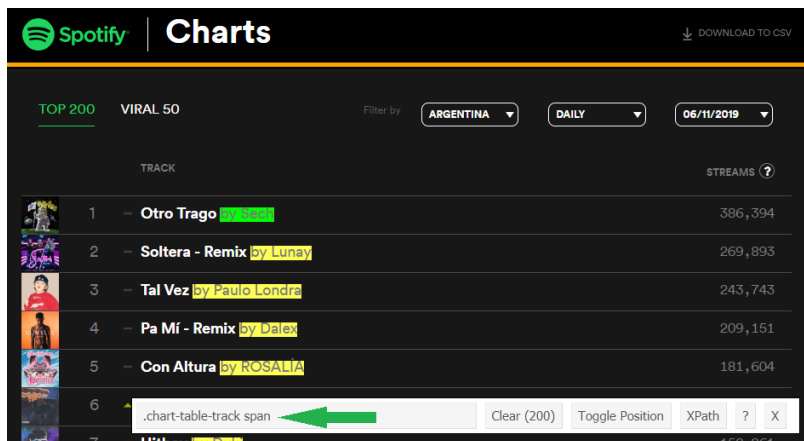
```

</td>
<td class="chart-table-position">1</td>
<td class="chart-table-trend">
  <div class="chart-table-trend_icon">
    <svg fill="#3e3e40" version="1.1" x="0" y="0" viewBox="0 0 12 12"><rect y="4.5" width="12"
height="3"></rect></svg>
  </div>
</td>
<td class="chart-table-track">
  <strong>Otro Trago</strong>
  <span>by Sech</span>
</td>
<td class="chart-table-streams">494,549</td>
</tr>
<tr>
<td class="chart-table-image">
  <a href="https://open.spotify.com/track/5ju3rF4URndK7t02xjSSE1" target="_blank">
    
</td>
<td class="chart-table-position">2</td>
<td class="chart-table-trend">
  <div class="chart-table-trend_icon">
    <svg fill="#3e3e40" version="1.1" x="0" y="0" viewBox="0 0 12 12"><rect y="4.5" width="12"
height="3"></rect></svg>
  </div>
</td>
<td class="chart-table-track">
  <strong>Tal Vez</strong>
  <span>by Paulo Londra</span>
</td>
<td class="chart-table-streams">343,925</td>
</tr>
<tr>
<td class="chart-table-image">
  <a href="https://open.spotify.com/track/7g0YauQABHal0zive7a2ljz" target="_blank">
    
</td>
<td class="chart-table-position">3</td>
<td class="chart-table-trend">
  <div class="chart-table-trend_icon">
    <svg fill="#3e3e40" version="1.1" x="0" y="0" viewBox="0 0 12 12"><rect y="4.5" width="12"
height="3"></rect></svg>
  </div>

```

Ayuda para identificar componentes HTML y CSS

Podemos ayudarnos del plugin para Chrome: **SelectorGadget**
Para identificar los componentes de la pagina web.



<https://cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html>

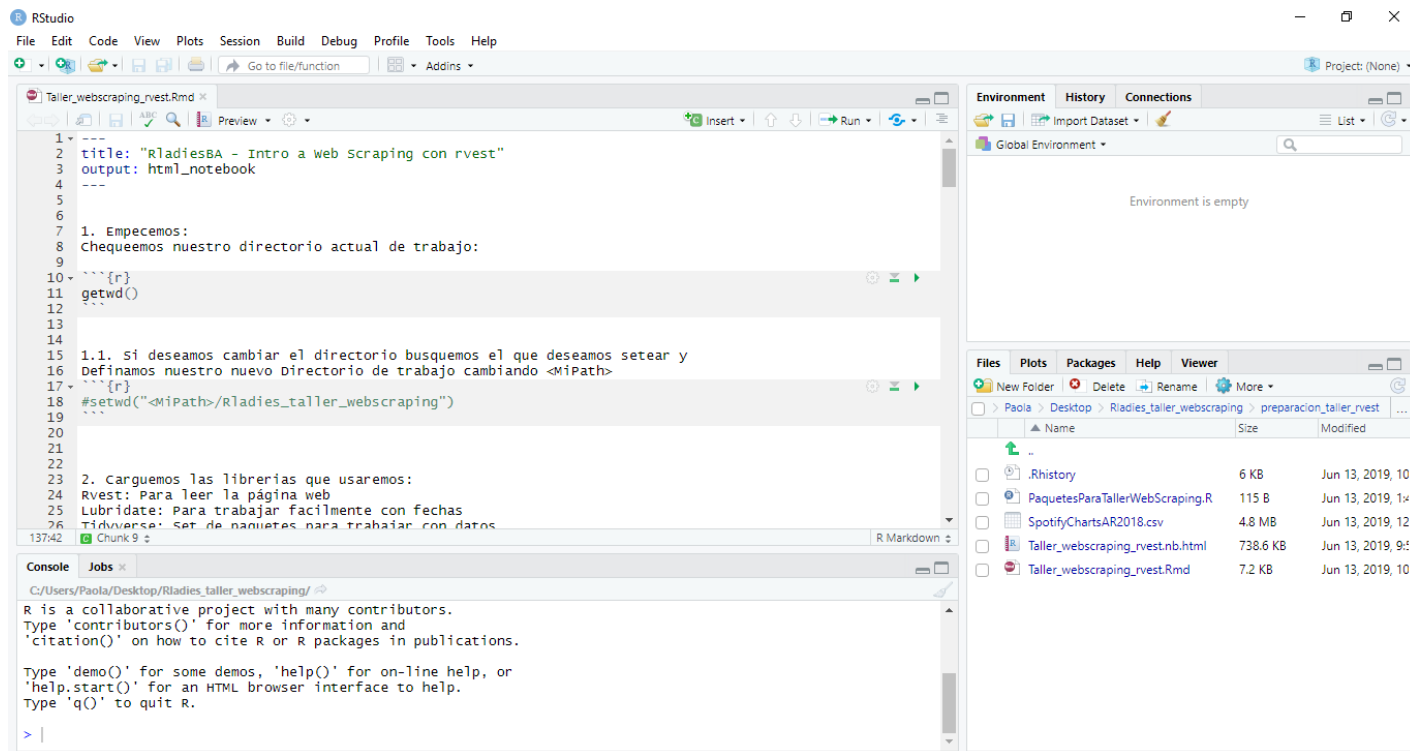


5. Empecemos con el Script



3. Script

Taller_webscrapping_rvest.Rmd



The screenshot shows the RStudio interface with the following components:

- Editor:** Displays the R Markdown script 'Taller_webscrapping_rvest.Rmd'. The script includes a title, output format, and R code for setting the working directory and installing packages.
- Environment:** Shows the 'Global Environment' with the message 'Environment is empty'.
- Files:** A file explorer showing the project structure, including folders like '.Rhistory' and '.Rplots', and files like 'Taller_webscrapping_rvest.Rmd'.
- Console:** Shows the output of the R script, including the message 'R is a collaborative project with many contributors.' and the prompt '> |'.

Script Content:

```

1 ---
2 title: "RladiesBA - Intro a Web Scrapping con rvest"
3 output: html_notebook
4 ---
5
6
7 1. Empecemos:
8 Chequeemos nuestro directorio actual de trabajo:
9
10 ```{r}
11 getwd()
12
13
14 1.1. Si deseamos cambiar el directorio busquemos el que deseamos setear y
15 Definamos nuestro nuevo Directorio de trabajo cambiando <MiPath>
16
17 ```{r}
18 #setwd("<MiPath>/Rladies_taller_webscrapping")
19
20
21
22 2. Carguemos las librerías que usaremos:
23 Rvest: Para leer la página web
24 Lubridate: Para trabajar fácilmente con fechas
25 Tidverse: Set de paquetes para trabajar con datos
26
27 13742 Chunk 9

```

Console Output:

```

C:/Users/Paola/Desktop/Rladies_taller_webscrapping/
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |

```

Files List:

Name	Size	Modified
.Rhistory	6 KB	Jun 13, 2019, 10
.Rplots	115 B	Jun 13, 2019, 10
PaquetesParaTallerWebScrapping.R	4.8 MB	Jun 13, 2019, 12
SpotifyChartsAR2018.csv	738.6 KB	Jun 13, 2019, 9:
Taller_webscrapping_rvest.nb.html	7.2 KB	Jun 13, 2019, 10
Taller_webscrapping_rvest.Rmd		

Muchas Gracias!

Preguntas?



WE ARE ALL WONDERWOMEN!



*John
C. Smith*

Referencias

Ética en Web Scraping

- ▶ Tener en cuenta: <https://ecija.com/web-scraping-legalhttps://ecija.com/web-scraping-legal-ilegal/-ilegal/>
- ▶ Meme: <https://stevenmortimer.com/scraping-responsibly-with-r/>

Paquete Rvest

<https://CRAN.R-project.org/package=rvest>

Plugin para Chrome

<https://cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html>

Referencias

- ▶ Jake Daniels, Scrape it yourself: Spotify Charts
<https://datacritics.com/2018/03/20/scrape-it-yourself-spotify-charts/>
- ▶ Scraping Spotify Data
https://rpubs.com/argdata/web_scraping
- ▶ Iterating rvest scrape function gives: “Error in open.connection(x, ”rb“) : Timeout was reached”
<https://stackoverflow.com/questions/39056103/iterating-rvest-scrape-function-gives-error-in-open-connectionx-rb-time>