Laboratorio Nacional de Políticas Públicas

CIDE CENTRO DE INVESTIGACIÓN Y DOCENCIA ECONÓMICAS A.C.

Web Scraping
(Introducción)
Periodismo de Datos
Abril, 2021

M.C. JORGE JUVENAL CAMPOS FERREIRA.
Investigador Asociado.
Laboratorio Nacional de Políticas Públicas
CIDE

Hoja de Ruta.

1. Revisión de conceptos básicos.

2. Revisión de conceptos legales.

3. Librería y funciones de R.

4. Ejemplo práctico.

Conceptos Básicos

Jorge Juvenal Campos Ferreira

Concepto: Web Scraping

*"Proceso de **extracción de datos almacenados en la web.** Su objetivo es el de recopilar información almacenada en una página web. Podemos "escrapear" artículos web, e-commerce, obtener precios, reseñas, etc."*
(Platzi).

Concepto: Web Scraping

***"La práctica de recolectar datos de manera automatizada a través de internet**, sin recurrir a la interacción con una API o al trabajo de un humano recolectando información".*

(Ryan Mitchell, Web Scraping with Python).

Concepto: Web Scraping

Activity where a party **uses automated software** to go and **crawl the internet and copy data and other content** so that can **compile it together** and make its **own product offering.**
(Evan Brown - Data Lawyer).

Discusión: ¿Qué tan Legal es el WS?

Si tu respuesta es “Si”, replantéate lo que piensas hacer

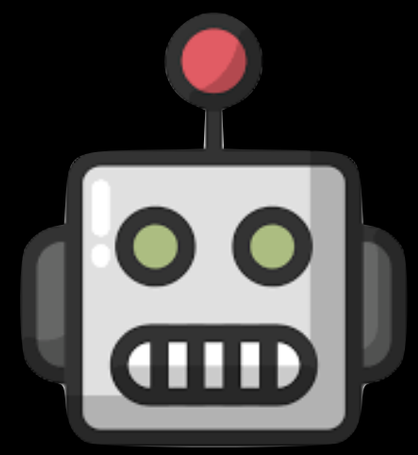
1. ¿Estoy violando alguna reglamentación local?
2. ¿Estoy violando los **Términos y Condiciones** del sitio?
3. ¿Estoy accediendo a lugares **no autorizados,** o a lugares donde se necesita hacer **login**?
4. ¿Es legal el uso que le voy a dar a los datos o **genera algún perjuicio de algún modo**?
5. ¿Estoy violando alguna ley de **derechos de autor**?
6. ¿Estoy **accediendo a datos personales**, o a datos que pudieran violar el **secreto comercial**?

Concepto: Robots.txt

El **robots.txt** es un archivo que define buenas prácticas a la hora de hacer scraping. Nos dice a qué sitios no quiere la página que accedamos.

Es recomendable seguir estas indicaciones, ya que no hacerlo puede ser **no-ético** y puede tener, en algunos casos**, consecuencias legales.**

https://platzi.com/robots.txt
https://www.linkedin.com/robots.txt
https://www.facebook.com/robots.txt
https://www.imdb.com/robots.txt
https://twitter.com/robots.txt



Páginas Web

Árbol

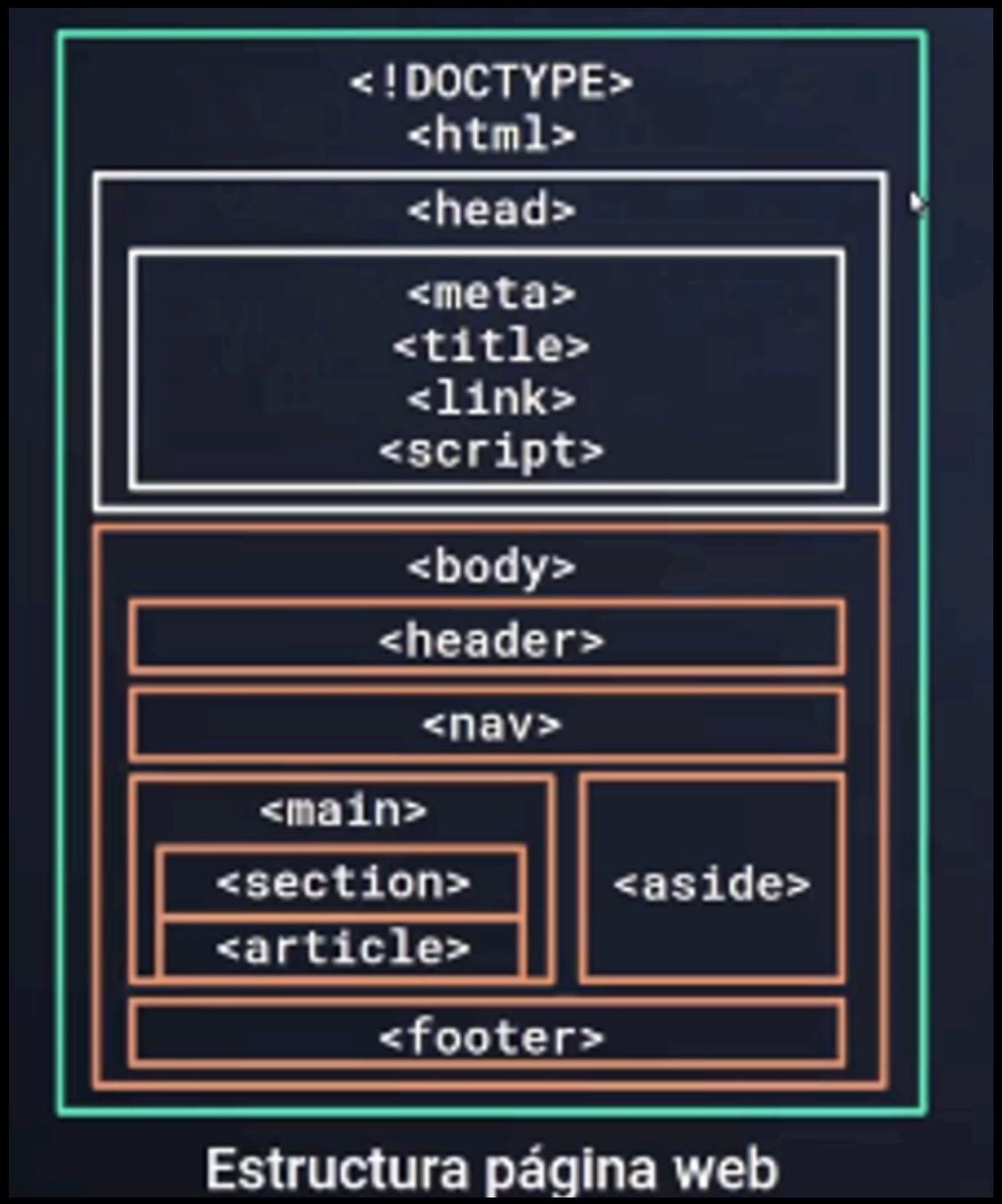
Inspector

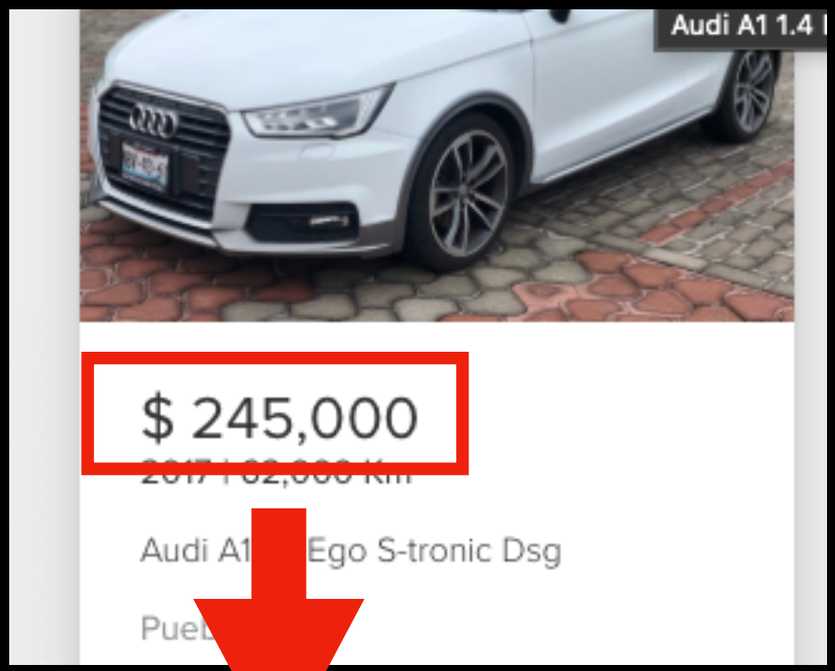
```
<!DOCTYPE html>
▼ <html lang="es-mx">
  ▶ <head>…</head>
  ▼ <body>
    ▼ <div class="wrapper">
      ▼ <header class="header">
        ▼ <nav class="nav">
          ▼ <a href="/" class="nav-logo">
              <img src="/images/logo.png" width="50" height="50" alt=
              "Logo">
            </a>
          ▼ <ul class="nav-links">
            ▼ <li>
                <a href="/sobre_mi/">Sobre mí</a>
              </li>
            ▶ <li>…</li>
            ▶ <li>…</li>
            ▶ <li>…</li>
            </ul>
          </nav>
        </header>
      ▼ <section class="hero">
        ▼ <div class="hero-inner">
            <h1>Juve Campos</h1>
            <h2>Blog personal 😎 👨‍💻 </h2>
          </div>
        </section>
      ▼ <main class="content" role="main">
        ▼ <div class="archive">
            <h2 class="archive-title">2020</h2>
          ▼ <article class="archive-item">
              <a href="/2020/07/26/cortando-islas/" class="archive-item-
              link">Cortando Islas</a>
              <span class="archive-item-date"> 2020-07-26 </span>
            </article>
          ▼ <article class="archive-item">
              <a href="/2020/03/03/ejemplos-de-lo-que-se-puede-hacer-en-
              shiny/" class="archive-item-link">Ejemplos de lo que se
              puede hacer en shiny</a>
              <span class="archive-item-date"> 2020-03-03 </span>
            </article>
          ▶ <article class="archive-item">…</article> = $0
          ▶ <article class="archive-item">…</article>
          ▶ <article class="archive-item">…</article>
          ▶ <article class="archive-item">…</article>
          </div>
        </main>
      ▼ <footer class="footer">
        ▶ <ul class="footer-links">…</ul>
        </footer>
      </div>
    </body>
</html>
```

Páginas Web

Para entrar al inspector, desde el navegador, hay que apretar click-derecho >> inspeccionar elemento

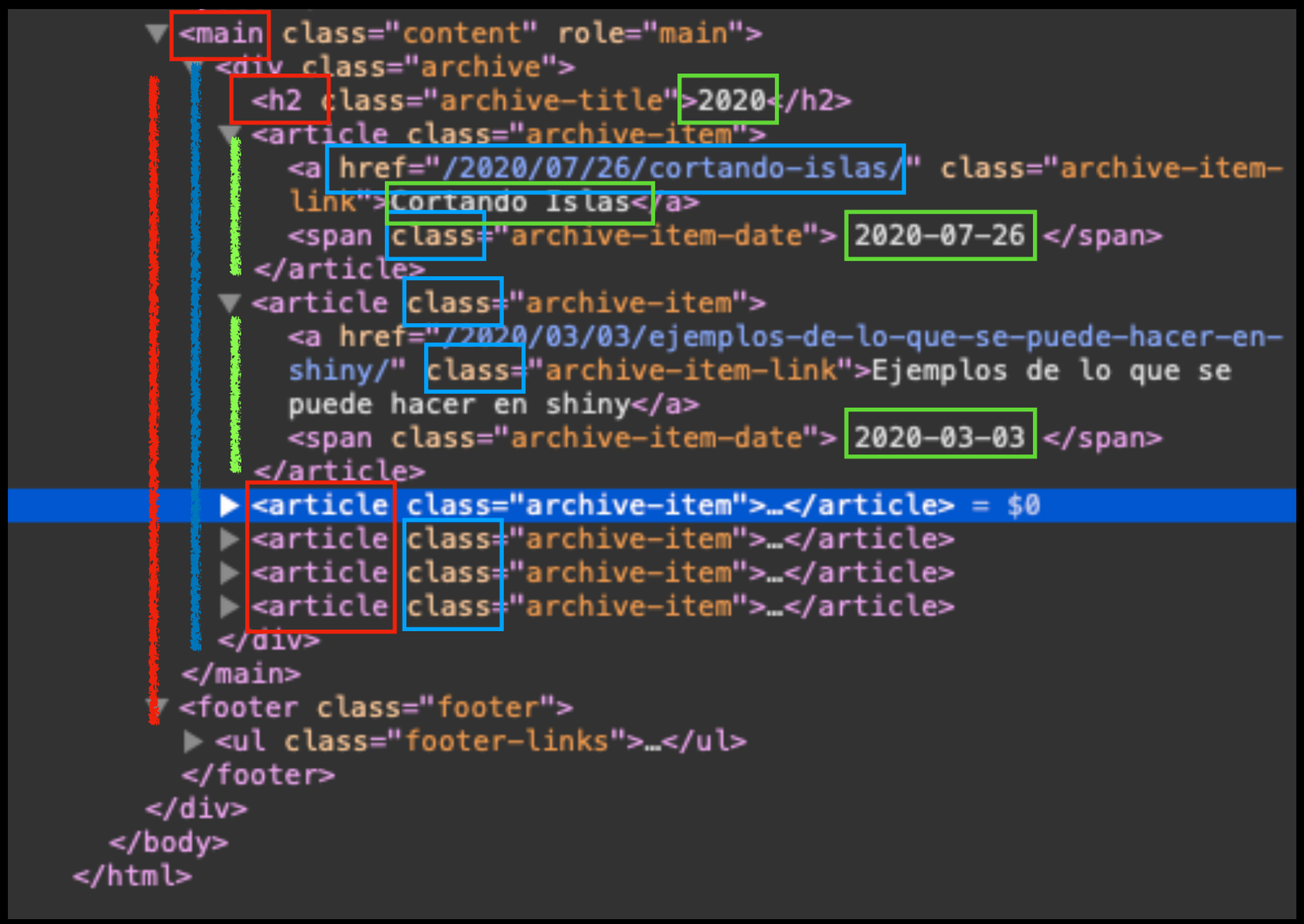
Páginas Web


<!DOCTYPE>
<html>
<head>
<meta>
<title>
<link>
<script>
<body>
<header>
<nav>
<main>
<section>
<article>
<aside>
<footer>
Estructura página web


Audi A1 1.4
$ 245,000
Audi A1 Ego S-tronic Dsg


Dato en la rama correcta

Páginas Web - Tags



Cuando hacemos web-scraping de sitios estáticos, vamos a **navegar en el árbol del HTML**, vamos a **anclarnos de los tags y los atributos,** y vamos a **extraer textos y atributos** como, por ejemplo, enlaces de páginas) **para ponerlos en un objeto de R** (vectores, tibbles o listas)

Texto **Atributos del HTML-Tag** **El HTML-Tag**

Páginas Web - Clases

Las clases **son atributos** de los cuales n**os podemos anclar** para hacer referencia a cierto elemento de la página web.

Cuando hagamos referencia a ellos en el web-scraping, **hay que colocarles un punto** al inicio del nombre de la clase, p. Ej, "*.archive-item*" o "*.archive-item-date*"

Páginas Web - Clases

Las clases **son atributos** de los cuales n**os podemos anclar** para hacer referencia a cierto elemento de la página web.

Cuando hagamos referencia a ellos en el web-scraping, **hay que colocarles un punto** al inicio del nombre de la clase, p. Ej, "*.archive-item*" o "*.archive-item-date*"

Funciones de R

Funciones de R

La librería rvest **nos ayuda a hacer web-scraping.**

Esta librería está **diseñada para trabajar con el tidyverse (%>%)** .

Trata de hacer **fácil las tareas más comunes de web-scraping**, inspirada en la librería *beautifulsoup* de Python.

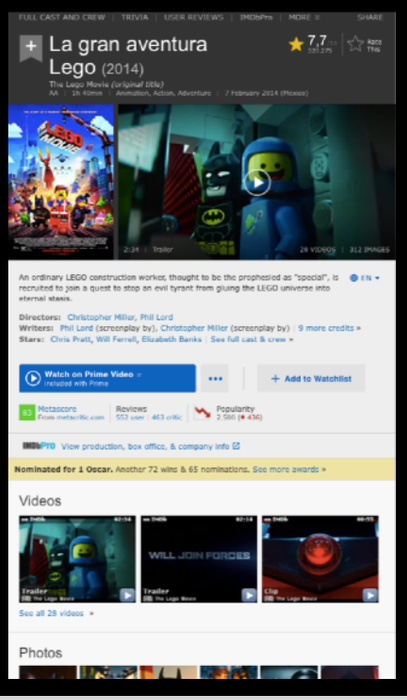
rvest::read_html(url)

Esta función nos permite descargar a nuestra sesión de R el código HTML de la página a la que le queremos sacar información.

```
library(rvest)

lego_movie <- read_html("http://www.imdb.com/title/tt1490017/")
```

Argumento: **url, cadena de texto.**
La dirección web de la página a la cual le queremos extraer información.



rvest::html_nodes()

Esta función extrae un nodo (rama) del código de la página web que tenemos almacenada en un objeto generado por *rvest::read_html()*

```
rating <- lego_movie %>%
  html_nodes("strong span")
```

**Atrapa los nodos del tipo "strong span" dentro del código html.*

El resultado de esta función aún no es legible. Hay que transformarlo (como lo que nos salía de las APIs).

rvest::html_nodes()

Los nodos pueden ser uno de los siguientes datos:

A) Un selector CSS.
B) Un tag de HTML
C) Un XPath (Dirección a un elemento en particular)
D) Una ruta de selección

Y se pueden combinar entre ellos.

rvest::html_text()

Esta función extrae el contenido que se encuentra dentro de un nodo (rama), resultante de la función *rvest::html_node()* o *rvest::html_nodes()*.

```
rating <- lego_movie %>%
  html_nodes("strong span") %>%
  html_text() %>%
  as.numeric()
rating
#> [1] 7.7
```

Acá la rama seleccionada es un tag de HTML

rvest::html_attr()

Esta función extrae el atributo HTML de un nodo que almacene atributos.

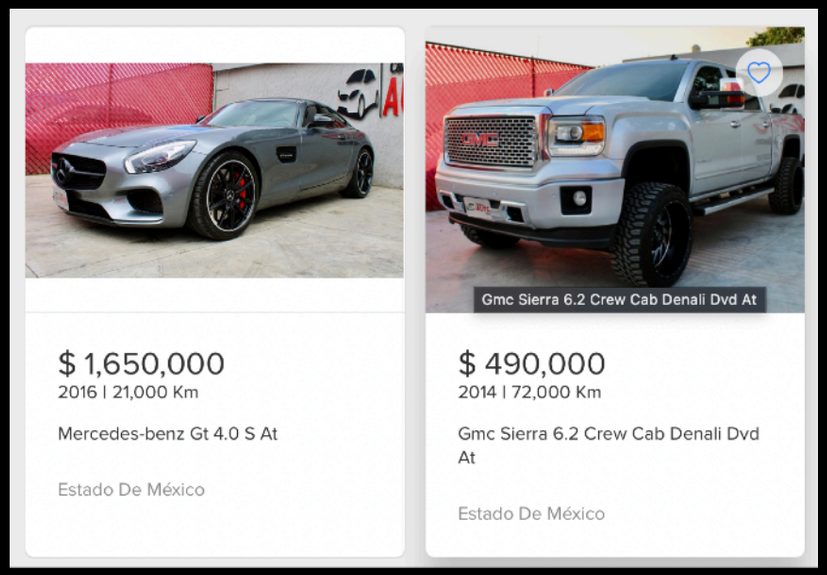
rvest::html_attr()

Ejemplos

```
cast <- lego_movie %>%
  html_nodes("#titleCast .primary_photo img") %>%
  html_attr("alt")
cast
#>  [1] "Will Arnett"     "Elizabeth Banks" "Craig Berry"     "Alison Brie"
#>  [5] "David Burrows"   "Anthony Daniels" "Charlie Day"     "Amanda Farinos"
#>  [9] "Keith Ferguson"  "Will Ferrell"    "Will Forte"      "Dave Franco"
#> [13] "Morgan Freeman"  "Todd Hansen"     "Jonah Hill"
```

```
poster <- lego_movie %>%
  html_nodes(".poster img") %>%
  html_attr("src")
poster
#> [1] "https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15BMl5BanBnXkFtZTgwNzIyNjg3MDE@._V1_U
```

rvest::html_text()



Inspector

```
▼ <a href="https://auto.mercadolibre.com.mx/MLM-903077082-mercedes-benz-clase-
gt-s-am2016-factura-original-tomo-auto-_JM#position=1&type=item&
tracking_id=04316880-e372-4fe8-92e6-faa5eabc400b" class="ui-search-
result__content ui-search-link" title="Mercedes-benz Gt 4.0 S At">
  ▼ <div class="ui-search-result__content-wrapper">
    ▼ <div class="ui-search-item__group ui-search-item__group--price">
      ▼ <div class="ui-search-price ui-search-price--size-medium ui-search-
      item__group__element">
        ▼ <div class="ui-search-price__second-line">
          ▼ <span class="price-tag ui-search-price__part">
              <span class="price-tag-symbol">$</span>
              <span class="price-tag-fraction">1,650,000</span>
            </span>
          </div>
        </div>
      </div>
    ▼ <div class="ui-search-item__group ui-search-item__group--attributes">
      ▼ <ul class="ui-search-card-attributes ui-search-item__group__element">
        ▶ <li class="ui-search-card-attributes__attribute">…</li>
          <li class="ui-search-card-attributes__attribute">21,000 Km</li>
        </ul>
      </div>
    ▼ <div class="ui-search-item__group ui-search-item__group--title">
        <h2 class="ui-search-item__title ui-search-
        item__group__element">Mercedes-benz Gt 4.0 S At</h2>
      </div>
    ▼ <div class="ui-search-item__group ui-search-item__group--location">
        <span class="ui-search-item__group__element ui-search-
        item__location">Estado De México</span>
      </div>
    </div>
  </a>
```

Si ya encontramos el dato que estamos buscando, y ya localizamos la rama correcta, lo podemos extraer con la función html_text()

rvest::html_text()


```r
precios = read_html(url) %>%
  html_nodes(".andes-card") %>%
  html_nodes(".ui-search-result__content-wrapper") %>%
  html_nodes(".ui-search-price") %>%
  html_text() %>%
  str_remove_all(pattern = "\\$|\\,") %>%
  as.numeric()
```

```r
carro = tibble(carro = read_html(url) %>%
                 html_nodes(".andes-card") %>%
                 html_nodes(".ui-search-result__content-wrapper") %>%
                  html_nodes("h2") %>%
                  html_text()) %>%
mutate(marca = str_extract(carro, pattern = "\\w+"))
```

Si ya encontramos el dato que estamos buscando, y ya localizamos la rama correcta, lo podemos extraer con la función html_text()

rvest::html_text()


```r
precios = read_html(url) %>%
  html_nodes(".andes-card") %>%
  html_nodes(".ui-search-result__content-wrapper") %>%
  html_nodes(".ui-search-price") %>%
  html_text() %>%
  str_remove_all(pattern = "\\$|\\,") %>%
  as.numeric()
```

```r
carro = tibble(carro = read_html(url) %>%
                 html_nodes(".andes-card") %>%
                 html_nodes(".ui-search-result__content-wrapper") %>%
                  html_nodes("h2") %>%
                  html_text()) %>%
mutate(marca = str_extract(carro, pattern = "\\w+"))
```

Si ya encontramos el dato que estamos buscando, y ya localizamos la rama correcta, lo podemos extraer con la función html_text()

rvest::html_table()

Country/Territory	Recreational	Medical	Notes
Afghanistan	Illegal	Illegal	*Main article: Cannabis in Afghanistan* Production banned by King Zahir Shah in 1973.[9]
Albania	Illegal	Illegal	*Main article: Cannabis in Albania* Prohibited but plants highly available throughout the country and law often unenforced.[10][11][12]
Algeria	Illegal	Illegal	*Main article: Cannabis in Algeria*
Andorra	Illegal	Illegal	*Main article: Cannabis in Andorra*
Angola	Illegal	Illegal	*Main article: Cannabis in Angola*
Antigua and Barbuda	Decriminalized	Illegal	*Main article: Cannabis in Antigua and Barbuda*
Argentina	Decriminalized	Legal	*Main article: Cannabis in Argentina* Decriminalized for small amounts and private consumption, as ruled by the Supreme Court in 2009.[13] Medicinal cannabis legal nationally since 21 September 2017.[14]
Armenia	Illegal	Illegal	*Main article: Cannabis in Armenia*
Australia	Decriminalized in Northern Territory and South Australia.[15][16] Legal in Australian Capital Territory for personal use but not for sale.	Legal at federal level and in all states.[17] Qualifying conditions and other details vary by state.[18]	*Main article: Cannabis in Australia* In September 2019, the Australian Capital Territory became the first state or territory of Australia to legalize recreational use of cannabis. Since 31 January 2020 residents have been allowed to grow two plants and possess 50 g, though sales or other transfer is prohibited, including cannabis seeds. Federal law also remains enforceable.[19]

Url = https://en.wikipedia.org/wiki/Legality_of_cannabis

Esta función nos sirve si en la página web ya esta ordenada la función que buscamos en forma de tabla. Nos extrae todas las tablas del sitio y nos las acomoda en una lista.

Actividad práctica.

- Repasaremos las funciones vistas previamente y veremos su aplicación en la descarga de datos.

Jorge Juvenal Campos Ferreira

MPPP - PdD2020