

Web scraping de páxinas dinámicas. Automatizando o buscador con RSelenium.

Martín García Cebeiro

Universidade de Santiago de Compostela



FACULTADE DE MATEMÁTICAS

Requisitos previos

① Paquetes de R:

- rvest
- RSelenium

② Java 1.8.0_461

③ Firefox (opcional)

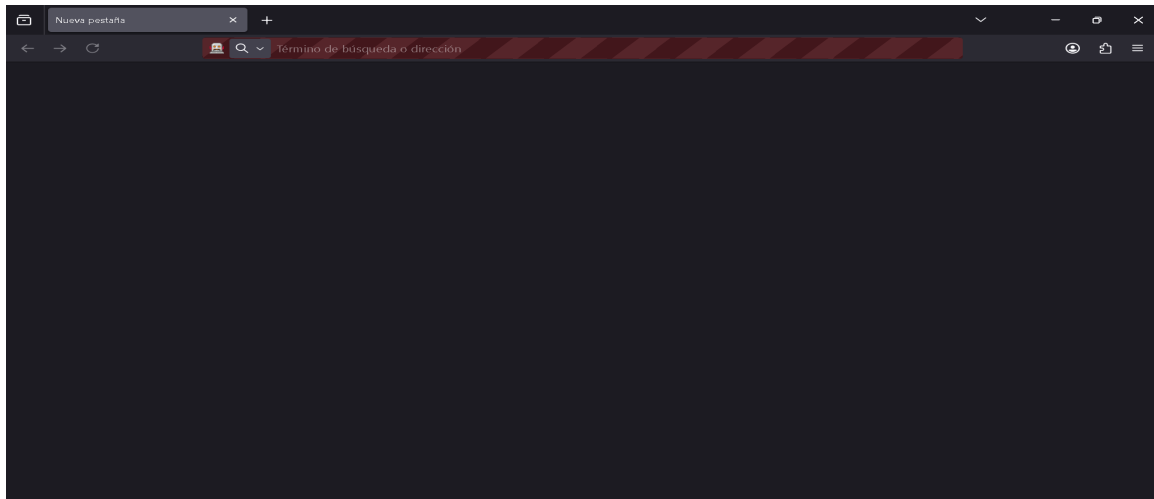
XI XORNADA DE USUARIOS DE EN **GALICIA**



**R, Webscraping y Open Science.
Esquivando balas en Matrix**

Paquete RSelenium

```
1 rD <- RSelenium::rsDriver(browser="firefox",phantomver=NULL)
```



```
1 remDr <- rD[["client"]]
2 remDr$navigate("https://www.bedca.net/bdpub/")
```





Base de Datos Española de
Composición de Alimentos

Spanish Food Composition Database

Inicio

Fuentes

Consulta

[Página principal](#) » Inicio

BIENVENIDO A LA BASE DE DATOS BEDCA

PARA USAR LA BASE DE DATOS HAGA CLICK EN CONSULTA EN LA ESQUINA SUPERIOR DERECHA

La Base de Datos Española de Composición de Alimentos publicada por la Red BEDCA del Ministerio de Ciencia e Innovación y bajo la coordinación y financiación de la Agencia Española de Seguridad Alimentaria y Nutrición del Ministerio de Sanidad, Servicios Sociales e Igualdad. Los valores de composición de alimentos recogidos en esta base de datos han sido obtenidos de distintas fuentes que incluyen laboratorios, industria alimentaria y publicaciones científicas o calculados. El acceso y uso de esta base de datos están sujetos a unas condiciones que puede conocer en el siguiente enlace: [UsoBD.pdf](#)

Navegadores soportados:



Mozilla Firefox (Recomendado)



Internet explorer



Opera



Safari

Copyright © 2007

#navigation :nth-child(4) a

Clear (1)

Toggle Position

XPath

Help

X

```
1 # Click en "Consulta"
```

```
2 consulta <- remDr$findElements(using="css selector",value="#navigation :nth-child(4) a")
```

```
3 consulta$clickElement()
```



#Alfabetica	Clear (1)	Toggle Position	XPath	Help	X
-------------	-----------	-----------------	-------	------	---

```
1 # Click en "Lista alfabetica"
2 lista_alfabetica <- remDr$findElement(using="css selector",value="#
  Alfabetica")
3 lista_alfabetica$clickElement()
```

[Todos](#) A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Otros

[Consulta anterior](#)

LISTADO DE ALIMENTOS DE LA CONSULTA

Id	Nombre	Name
746	Aceite de algodón	Cotton oil
747	Aceite de cacahuete	Peanut oil
748	Aceite de coco	Coconut oil
753	Aceite de colza	Rape oil
749	Aceite de germen de trigo	Wheat germ oil
2541	Aceite de girasol	Sunflower oil
754	Aceite de grano de uva	Grape seed oil
755	Aceite de hígado de bacalao	Cod liver oil
2542	Aceite de lino	Flaxseed oil
750	Aceite de nuez	Walnut oil
2543		
2544		
7545	Aceite de olivos vírgenes extra. producción ecológica	Extra vírgen olive oil. organic

#alphabet a:nth-child(1)

Clear (1)

Toggle Position

XPath

Help

X

```

1 # Click en "Todos"
2 todos <- remDr$findElements(using="css selector",
3 value="#alphabet a:nth-child(1)")
4 todos$clickElement()

```

Todos A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Otros

[Consulta anterior](#)

LISTADO DE ALIMENTOS DE LA CONSULTA

Id	Nombre	Name
746	Aceite de algodón	Cotton oil
747	Aceite de cacahuete	Peanut oil
748	Aceite de coco	Coconut oil
753	Aceite de colza	Rape oil
749	Aceite de germen de trigo	Wheat germ oil
2541	Aceite de girasol	Sunflower oil
754	Aceite de grano de uva	Grape seed oil
755	Aceite de hígado de bacalao	Cod liver oil
2542	Aceite de lino	Flaxseed oil
750	Aceite de nuez	Walnut oil
2543	Aceite de oliva	
2544	Aceite de oliva virgen	
2545	Aceite de oliva virgen extra, producción ecológica	Extra virgin olive oil, organic

#content2

Clear (1)

Toggle Position

XPath

Help

X

CSS Simple Selectors

The simple selectors select elements based on element-name, id, and class. In addition, there is the universal selector (*).

Selector	Example	Example description
<u>element</u>	p	Selects all <p> elements
<u>#id</u>	#firstname	Selects the element with id="firstname"
<u>*</u>	*	Selects all elements
<u>.class</u>	.intro p.intro	Selects all elements with class="intro" Selects all <p> elements with class="intro"

CSS Attribute Selectors

The attribute selector selects HTML elements with a given attribute set.

Selector	Example	Example description
<code>[<u>attribute</u>]</code>	<code>[lang]</code>	Selects all elements with a lang attribute
<code>[<u>attribute</u>=<i>value</i>]</code>	<code>[lang="it"]</code>	Selects all elements with lang="it"
<code>[<u>attribute</u>~=<i>value</i>]</code>	<code>[title~="flower"]</code>	Selects all elements with a title attribute containing the word "flower"
<code>[<u>attribute</u>]=<i>value</i>]</code>	<code>[lang]="en"]</code>	Selects all elements with a lang attribute value equal to "en" or starting with "en-"
<code>[<u>attribute</u>^=<i>value</i>]</code>	<code>[href^="https"]</code>	Selects all elements with a href attribute value that begins with "https"
<code>[<u>attribute</u>\$=<i>value</i>]</code>	<code>[href\$=".pdf"]</code>	Selects all elements with a href attribute value ends with ".pdf"
<code>[<u>attribute</u>*=<i>value</i>]</code>	<code>[href*="w3schools"]</code>	Selects all elements with a href attribute value containing the substring "w3schools"

Base de Datos BEDCA

www.bedca.net/bdpub/

Página principal » Menú de Consulta

Todos A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Otros

Consulta anterior

LISTADO DE ALIMENTOS DE LA CONSULTA

a	18 x 14.6667	Nombre	Name
746		Aceite de algodón	Cotton oil
747		Aceite de cacahuete	Peanut oil
748		Aceite de coco	Coconut oil
753		Aceite de colza	Rape oil
749		Aceite de germen de trigo	Wheat germ oil
2541		Aceite de girasol	Sunflower oil
754		Aceite de grano de uva	Grape seed oil
755		Aceite de hígado de bacalao	Cod liver oil
2542		Aceite de lino	Flaxseed oil

Inspector

Buscar en HTML

```

<tbody>
  <tr>
    <td class="row-a">
      <td align="center">
        <a href="javascript:void(0)"
          onclick="scroll(0,0);document.getElementById('alphabet').style.display='none';new Array(),new Array('746',1,'componentgroup_id','ASC')>746</a>
      </td>
      <td id="Foodname">
      </td>
    </tr>
  </tbody>

```

Filtrar estilos

```

elemento {
  #content table a {
    text-decoration: none;
    color: #000000;
  }
  #content a {
    text-decoration: none;
  }
  Heredado de td
  #content td {
    font-size: 12px;
  }

```

Disposición

Flexbox

Seleccione un contenedor o elemento Flex para continuar.

Rejilla

No hay rejillas CSS en uso en esta página

Modelo de caja

margin: 0

border: 0

```
1 alimento <- remDr$findElement(using="css selector", value="[onclick*=\"746\"
    ]") # [attribute*=value] (contains)
2 foodname$clickElement()
3 library(rvest)
4 html <- read_html(remDr$getPageSource()[[1]])
5 # ... Procesado do html con rvest ...
6 listado_anterior <- remDr$findElement(using="css selector", value="#Todos")
7 bck$clickElement()
8 remDr$close()
9 rD[["server"]]$stop()
```

Accións de navegador con RSelenium

Función	Argumentos	Descrición
goBack	-	Navegar á páxina anterior
goForward	-	Navegar á páxina posterior
refresh	-	Refrescar a páxina actual
getCurrentUrl	-	Devolve a url da páxina actual
screenshot	display, useViewer, file	Saca unha captura de pantalla
getCurrentWindowHandle	-	Devolve o ID da ventá actual
switchToWindow	windowId	Cambia á ventá indicada
mouseMoveToLocation	x, y, webElement	Move o rato (x,y) dende webElement

Cadro: Algúns exemplos de accións de navegador que se poden levar a cabo dende RSelenium.

Interaccións cos elementos web con RSelenium

Función	Argumentos	Descripción
sendKeysToElement	sendKeys	Envía unha secuencia de teclas a un elemento
click	buttonId	Fai click co botón indicado
doubleclick	buttonId	Fai dobre click co botón indicado
buttondown	buttonId	Mantén o botón indicado pulsado
buttonup	buttonId	Deixa de manter pulsado o botón indicado

Cadro: Algúns exemplos de interaccións cos elementos web que se poden levar a cabo dende RSelenium.

- [1] <https://cran.r-project.org/web/packages/R Selenium/R Selenium.pdf>
- [2] https://www.w3schools.com/cssref/css_selectors.php
- [3] <https://www.bedca.net/bdpub/>