vafonsor_prac1

Víctor Elías Afonso Rodríguez y Saúl Santomé Rúa

6/4/2021

9. Código

9.1 Cargamos librerias necesarias

```
library(robotstxt)
library(rvest)
library(readr)
```

```
##
## Attaching package: 'readr'

## The following object is masked from 'package:rvest':
##
##     guess_encoding
```

```
library(stringr)
```

La libreria "**robotstxt**" nos permite comprobar el archivo robot de las diferentes webs para asegurarnos que podemos realizar scraping. La libreria **rvest** permite extraer información de una web.
La libreria **readr** permite exportar dataset a csv. La libreria **stringr** permite trabajar con cadenas de caracteres.

9.2 Elegimos la url de la que queremos extraer la información

Amazon

```
url_ama <- "https://www.amazon.es/s?k=paraguas&__mk_es_ES=%C3%85M%C3%85%C5%BD%C3%95%C3%91&crid=2TPM4DDOU
```

Ebay

```
url_eba <- "https://www.ebay.es/sch/i.html?_from=R40&_trksid=p2334524.m570.l1313&_nkw=paraguas&_sacat=0&
```

9.3 Comprobamos que se puede realizar scraping

Amazon

```
paths_allowed(paths=c(url_ama))
```

```
##  www.amazon.es
```

```
## [1] TRUE
```

Ebay

```
paths_allowed(paths=c(url_eba))
```

```
##  www.ebay.es
```

```
## [1] TRUE
```

El valor "TRUE" indica que podemos realizar scraping.

9.4 Convertimos en objeto

Amazon

```
web_ama <- read_html(url_ama)
```

Ebay

```
web_eba <- read_html(url_eba)
```

"read_html" permite leer el código html.

9.5 Escogemos la información que queremos extraer de cada web

9.5.1 Nombre de artículos

Amazon

```
clase_nombre_articulo_ama <- ".s-line-clamp-2"

articulo_ama <- html_nodes(web_ama,clase_nombre_articulo_ama)

nombre_articulo_ama <- html_text(articulo_ama)

#Se eliminan los saltos de linea
nombre_articulo_ama <- str_replace_all(nombre_articulo_ama, "[\n]", "")

nombre_articulo_ama <- nombre_articulo_ama[0:58]
```

Ebay

```
clase_nombre_articulo_eba <- ".s-item__title"

articulo_eba <- html_nodes(web_eba,clase_nombre_articulo_eba)

nombre_articulo_eba <- html_text(articulo_eba)

nombre_articulo_eba <- nombre_articulo_eba[0:58]
```

"html_nodes" permite extaer las partes deseadas del código HTML. "html_text" transformamos a texto.

9.5.2 Precio de artículos

Amazon

```
clase_precio_articulo_ama <- ".a-price-whole"

articulo_ama <- html_nodes(web_ama,clase_precio_articulo_ama)

precio_articulo_ama <- html_text(articulo_ama)

precio_articulo_ama <- gsub(",",".",precio_articulo_ama)

precio_articulo_ama <- as.numeric(precio_articulo_ama)

precio_articulo_ama <- precio_articulo_ama[0:58]
```

Ebay

```
clase_precio_articulo_eba <- ".s-item__price"

articulo_eba <- html_nodes(web_eba,clase_precio_articulo_eba)

precio_articulo_eba <- html_text(articulo_eba)

precio_articulo_eba <- gsub("EUR","",precio_articulo_eba)

precio_articulo_eba <- gsub(",",".",precio_articulo_eba)

precio_articulo_eba <- as.numeric(precio_articulo_eba)
```

```
## Warning: NAs introducidos por coerción
```

```
precio_articulo_eba <- precio_articulo_eba[0:58]
```

9.5.2 Link de compra de los artículos

Amazon

```
claseIdentificacionArticulo_ama <- ".s-line-clamp-2"

articulo_ama <- html_nodes(web_ama, claseIdentificacionArticulo_ama)

finLinkCompra_ama <- html_nodes(articulo_ama, ".a-link-normal") %>% html_attr('href')

#Al finLinkCompra_ama hay que añadirle antes "https://www.amazon.es" para que funcione
linkCompra_ama <- str_c("https://www.amazon.es", finLinkCompra_ama)

linkCompra_ama <- linkCompra_ama[0:58]
```

Ebay

```
claseIdentificacionArticulo_eba <- ".s-item__info"

articulo_eba <- html_nodes(web_eba, claseIdentificacionArticulo_eba)

linkCompra_eba <- html_nodes(articulo_eba, ".s-item__link") %>% html_attr('href')

linkCompra_eba <- linkCompra_eba[0:58]
```

9.5.2 Link de imagenes de artículos

Amazon

```
claseIdentificacionArticulo_ama <- ".s-image-tall-aspect"

articulo_ama <- html_nodes(web_ama, claseIdentificacionArticulo_ama)

linkImagen_ama <- html_nodes(articulo_ama, ".s-image") %>% html_attr('src')


linkImagen_ama <- linkImagen_ama[0:58]
```

Ebay

```
claseIdentificacionArticulo_eba <- ".s-item__image"

articulo_eba <- html_nodes(web_eba, claseIdentificacionArticulo_eba)

linkImagen_eba <- html_nodes(articulo_eba, ".s-item__image-img") %>% html_attr('src')

linkImagen_eba <- linkImagen_eba[0:58]
```

10. Dataset

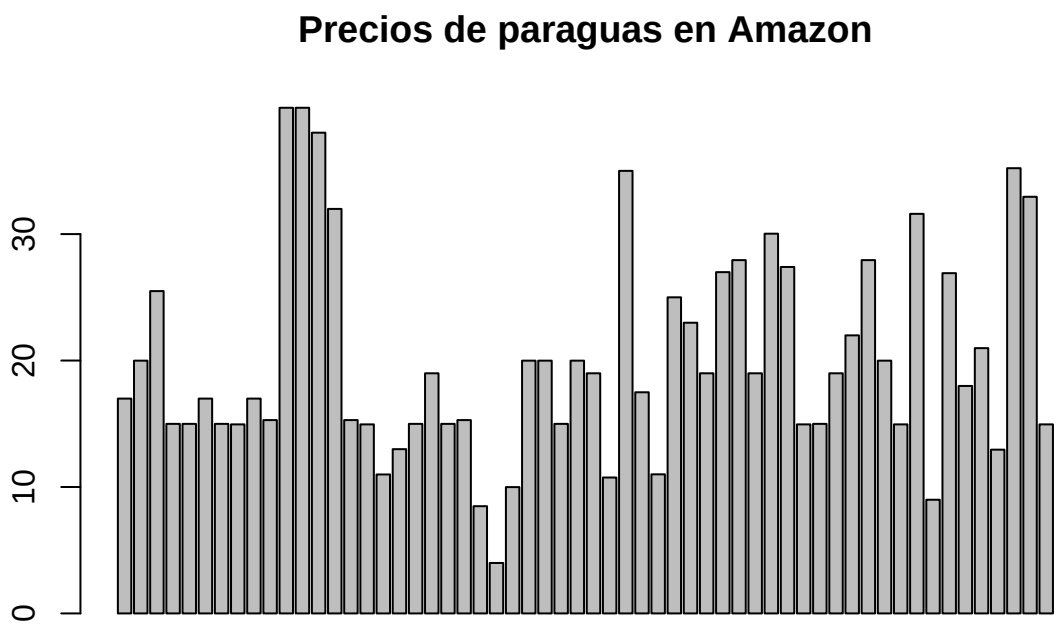
Unimos todos los atributos de los diferentes artículos

Amazon

```
Amazon <- data.frame(Producto = nombre_articulo_ama, Precio = precio_articulo_ama, Link = linkCompra_ama
```

Representamos el comportamiento de los precios en Amazon

```
barplot(precio_articulo_ama, main = "Precios de paraguas en Amazon")
```

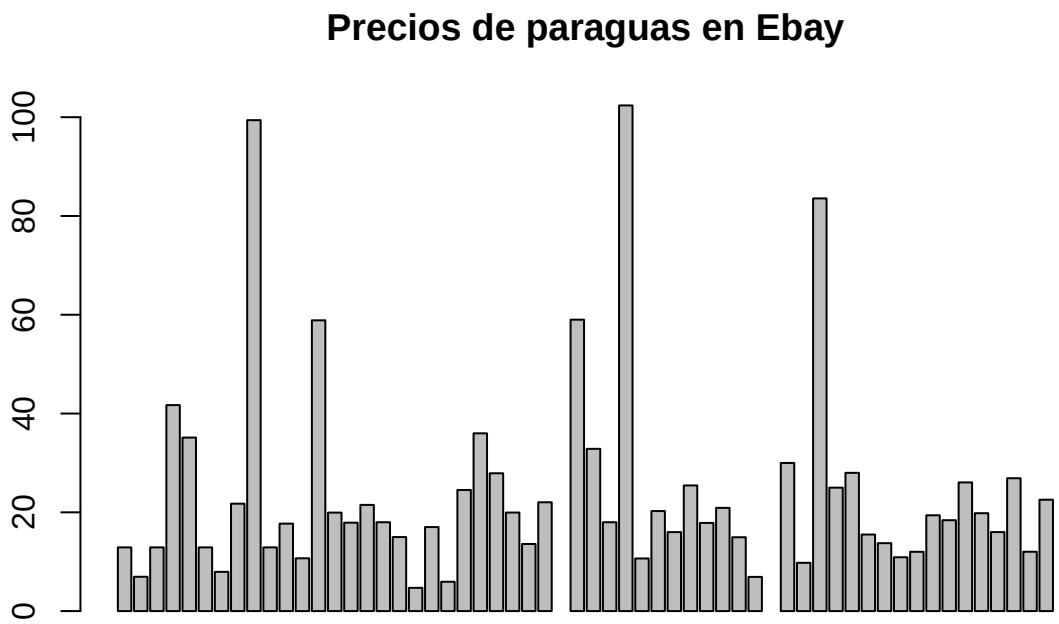


Ebay

```
Ebay <- data.frame(Producto = nombre_articulo_eba, Precio = precio_articulo_eba, Link = linkCompra_eba,
```

Representamos el comportamiento de precios en Ebay

```
barplot(precio_articulo_eba, main = "Precios de paraguas en Ebay")
```




```
summary(precio_articulo_ama)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.99   14.99   18.49   20.09   25.37   39.99
```

```
summary(precio_articulo_eba)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    4.71   12.90   18.20   24.12   25.59  102.38       2
```

Observamos que los precios son menores en Ebay.

Exportamos a csv

```
write_csv(Amazon, file = "../datasets/Amazon.csv")
write_csv(Ebay, file= "../datasets/Ebay.csv")
```