Ejercitación-1-NB

Nahuel Bargas

Tabla de contenidos

1 Punto 1

1.1 Consigna

1. Seleccioná un sitio web de la pestaña “Tradicionales” del documento con los comercios electrónicos a scrapear e inspecciona el código fuente. Guarda la precaución que seas el único que eligió ese sitio.

1.2 Respuesta

Selecciono Easy, cuya web es https://www.easy.com.ar/.

2 Punto 2

2.1 Consigna

2. Luego identifica 3 tipos diferentes de etiquetas y respondepara cada una:

a. ¿Para qué se utiliza?
b. ¿Qué atributos posee?
c. ¿Qué característica brinda cada uno de los atributos observados?

2.2 Respuesta

Etiqueta: **link**, línea 16 del código de fuente de la página

a) – Según ‘w3schools’ se utiliza para vincular hojas de estilos externas e íconos para la url en los navegadores.

b) – rel, class, href, as, crossorigin, id, onload, onerror

c) –

- *rel* =‘preload’ indica que el navegador debe tener cómo prioridad cargar este recurso al comienzo de la consulta.
- *class* =‘vtex_io_uncritical_css’ Es el nombre de la clase del elemento y se asocia con un estilo.

- *href* =‘https://...’ Indica la ubicación del elemento a cargar.
- *as* =‘style’ Especifica el contenido cargado por el elemento. Cuando se establece rel=“preload”, es prioritario su definición.
- *crossorigin* = ’ ’ Establece cómo responde el elemento ante consultas de origen cruzado.
- *id* = ‘uncritical_style_0’ Es la clave única de identificación del elemento.
- *onload* =“...” La acción llevada a cabo cuando se carga el elemento.
- *onerror* =“...” El script a ejecutar cuando ocurre un error al cargar el elemento.

Etiqueta: **noscript**, línea 17 del código de fuente de la página

a)– Según ‘w3schools’ define el contenido a mostrar para los usuarios que hayan deshabilitado los scripts en sus navegadores o que posean un navegador que no soporte scripts.

b)– id

c)– id=“styles_overrides” Es la clave única de identificación del elemento.

Etiqueta: **template**, línea 237 del código de fuente de la página

a)– Siguiendo con ‘w3schools’, sirve para mantener cierto contenido HTML escondido para la vista del usuario hasta que la página se cargue. Posteriormente, el contenido puede ser ‘renderizado’ via JavaScript.

b)– data-type, data-varname

c)– Ambos atributos brindan la posibilidad de meter datos personalizados en el elemento y luego pueden ser utilizados en alguna función de JavaScrip para mejorar la experiencia del usuario al navegar por la web.

3 Punto 3

3.1 Consigna

3. Accedé al primer ejemplo de código HTML (muy sencillo) que hemos preparado para este taller y respondé:

a. ¿Qué elementos HTML posee?
b. ¿Qué atributos posee cada elemento?
c. ¿Qué característica brinda cada uno?

3.2 Respuesta

a)– html, head, meta charset=”utf-8 , title, style, body, h1, p, table, thead,tbody, tr, td, a, div y img.

b)– y c)

html :

- lang=“es”. Indica el lenguaje de la página

h1:

- align=“center” . La alineación horizontal del texto del encabezado más grande.
- class=“marca”. El nombre de la clase del elemento.

p:

- align=“center”. La alineación horizontal del texto del elemento p.

table:

- border=“1” El tipo de borde de la tabla. Con el valor ‘1’, el grosor del borde es más fino.
- align=“center” La alineación de la tabla en la página web. En este caso, centrada.

td:

- class=“marca”. El nombre de la clase del elemento.

a:

- href=“ ” El hipervínculo del link, es una dirección hacía un elemento exterior.

div :

- class=‘image-container’. El nombre de la clase del elemento.

img:

- src= “https://www.python.org/static/community_logos/python-logo-master-v3-TM.png” La ubicación del archivo en formato imágen.
- alt= “Logo de Python”. La descripción de la imágen. Es útil para que los navegadores identifiquen el contenido de la imágen y ayuda a las personas con disminución visual que utilizan programas de lectura.

4 Punto 4

4.1 Consigna

4. Por fín, llegamos a trabajar con Python:

a. Instalá e importá la librería requests y descargá el código html.

4.2 Respuesta

```
from bs4 import BeautifulSoup
import requests

### Web
url = "https://raw.githubusercontent.com/jumafernandez/soco-web_scraping/main/data/encuent

### Consulta a la url
response = requests.get(url)

### Obtener el html

print(response.text)
```

```
<!DOCTYPE html>
<html lang="es">
<head>
<meta charset="UTF-8">
<title>Información de Tiendas</title>

<style>
.image-container {
  text-align: center;
}
</style>

</head>
<body>
  <h1 align="center" class="marca">Información de Tiendas</h1>

  <p align="center">A continuación se listan 15 tiendas que comienzan con la letra A:</p>
```

```
<table border="1" align="center">
  <thead>
    <tr>
      <th class="marca">Nombre</th>
      <th>Enlace</th>
    </tr>
  </thead>
  <tbody>
    <tr>
      <td class="marca">Adidas</td>
      <td><a href="https://www.adidas.com.ar/">https://www.adidas.com.ar/</a></td>
    </tr>
    <tr>
      <td class="marca">Akiabara</td>
      <td><a href="https://akiabara.com/">https://akiabara.com/</a></td>
    </tr>
    <tr>
      <td class="marca">Al Mundo</td>
      <td><a href="https://almundo.com.ar/">https://almundo.com.ar/</a></td>
    </tr>
    <tr>
      <td class="marca">Alfabeta</td>
      <td><a href="https://www.alfabeta.net/home/">https://www.alfabeta.net/home/</a></td>
    </tr>
    <tr>
      <td class="marca">Almacen de Pizzas</td>
      <td><a href="https://go.waitry.net/order/2358/almacen-de-pizzas-unicenter/take-away"
    </tr>
    <tr>
      <td class="marca">Aloise</td>
      <td><a href="https://catalogoaloise.com/">https://catalogoaloise.com/</a></td>
    </tr>
    <tr>
      <td class="marca">Amphora</td>
      <td><a href="https://www.amphora.com.ar/">https://www.amphora.com.ar/</a></td>
    </tr>
    <tr>
      <td class="marca">Anavana</td>
      <td><a href="https://www.anavana.com.ar/">https://www.anavana.com.ar/</a></td>
    </tr>
    <tr>
      <td class="marca">Anca & Co</td>
      <td><a href="https://www.ancayco.com.ar/">https://www.ancayco.com.ar/</a></td>
```

```
      </tr>
      <tr>
        <td class="marca">Aquí y Ahora Stand</td>
        <td><a href="https://www.aquiyahora.com.ar/">https://www.aquiyahora.com.ar/</a></td>
      </tr>
      <tr>
        <td class="marca">Arredo</td>
        <td><a href="https://www.arredo.com.ar/">https://www.arredo.com.ar/</a></td>
      </tr>
      <tr>
        <td class="marca">Artisan</td>
        <td><a href="https://artisanbuenosaires.com.ar/">https://artisanbuenosaires.com.ar/<
      </tr>
      <tr>
        <td class="marca">Atomik</td>
        <td><a href="https://atomik.com.ar/">https://atomik.com.ar/</a></td>
      </tr>
      <tr>
        <td class="marca">Atrápalo</td>
        <td><a href="https://www.atrapalo.com.ar/">https://www.atrapalo.com.ar/</a></td>
      </tr>
      <tr>
        <td class="marca">Ay Not Dead</td>
        <td><a href="https://aynotdead.com/">https://aynotdead.com/</a></td>
      </tr>
    </tbody>
  </table>

  <div class="image-container">
    <img src="https://www.python.org/static/community_logos/python-logo-master-v3-TM.png" alt
  </div>

</body>
</html>
```

```
### otra forma utilizando BeautifulSoup

html_bruto = BeautifulSoup(response.text, 'html.parser')

print(html_bruto.prettify())
```

```
<!DOCTYPE html>
<html lang="es">
 <head>
  <meta charset="utf-8"/>
  <title>
   Información de Tiendas
  </title>
  <style>
   .image-container {
  text-align: center;
}
  </style>
 </head>
 <body>
  <h1 align="center" class="marca">
   Información de Tiendas
  </h1>
  <p align="center">
   A continuación se listan 15 tiendas que comienzan con la letra A:
  </p>
  <table align="center" border="1">
   <thead>
    <tr>
     <th class="marca">
      Nombre
     </th>
     <th>
      Enlace
     </th>
    </tr>
   </thead>
   <tbody>
    <tr>
     <td class="marca">
      Adidas
     </td>
     <td>
      <a href="https://www.adidas.com.ar/">
       https://www.adidas.com.ar/
      </a>
     </td>
    </tr>
    <tr>
```

```
 <td class="marca">
  Akiabara
 </td>
 <td>
  <a href="https://akiabara.com/">
   https://akiabara.com/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Al Mundo
 </td>
 <td>
  <a href="https://almundo.com.ar/">
   https://almundo.com.ar/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Alfabeta
 </td>
 <td>
  <a href="https://www.alfabeta.net/home/">
   https://www.alfabeta.net/home/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Almacen de Pizzas
 </td>
 <td>
  <a href="https://go.waitry.net/order/2358/almacen-de-pizzas-unicenter/take-away">
   https://go.waitry.net/order/2358/almacen-de-pizzas-unicenter/take-away
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Aloise
 </td>
```

```
 <td>
  <a href="https://catalogoaloise.com/">
   https://catalogoaloise.com/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Amphora
 </td>
 <td>
  <a href="https://www.amphora.com.ar/">
   https://www.amphora.com.ar/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Anavana
 </td>
 <td>
  <a href="https://www.anavana.com.ar/">
   https://www.anavana.com.ar/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Anca &amp;amp; Co
 </td>
 <td>
  <a href="https://www.ancayco.com.ar/">
   https://www.ancayco.com.ar/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Aquí y Ahora Stand
 </td>
 <td>
  <a href="https://www.aquiyahora.com.ar/">
   https://www.aquiyahora.com.ar/
```

```
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Arredo
 </td>
 <td>
  <a href="https://www.arredo.com.ar/">
   https://www.arredo.com.ar/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Artisan
 </td>
 <td>
  <a href="https://artisanbuenosaires.com.ar/">
   https://artisanbuenosaires.com.ar/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Atomik
 </td>
 <td>
  <a href="https://atomik.com.ar/">
   https://atomik.com.ar/
  </a>
 </td>
</tr>
<tr>
 <td class="marca">
  Atrápalo
 </td>
 <td>
  <a href="https://www.atrapalo.com.ar/">
   https://www.atrapalo.com.ar/
  </a>
 </td>
</tr>
```

```
    <tr>
     <td class="marca">
      Ay Not Dead
     </td>
     <td>
      <a href="https://aynotdead.com/">
       https://aynotdead.com/
      </a>
     </td>
    </tr>
   </tbody>
  </table>
  <div class="image-container">
   <img alt="Logo de Python" src="https://www.python.org/static/community_logos/python-logo-m
  </div>
 </body>
</html>
```

5 Punto 5

5.1 Consignas:

5. Ahora vamos a explorar el contenido HTML de estos documentos.

a. Modifica el script anterior para obtener el título principal del documento y mostrarlo en pantalla.
b. Ahora, cargá la tabla en un DataFrame de la librería Pandas.
c. Iterá en los elementos
del contenido y mostralos en pantalla.
d. A continuación, quedate con los elementos
que sean de clase “marca” y mostrá en pantalla aquellos que terminen con la letra “o”.
e. Guarda la tabla de la página en un archivo xlsx. Tené en cuenta que el archivo sólo deberá tener dos columnas: tienda y sitio web.

5.2 Respuestas

5.2.1 a) Obtener el título del documento:

```
print(html_bruto.find('title').get_text())
```

```
Información de Tiendas
```

5.2.2 b) Cargar la tabla en un data.frame

```
import pandas as pd # importar pandas

tabla= html_bruto.find('table') #obtener la tabla en base al elemento


nombre_lista = []
enlaces_lista =[]

for fila in tabla.tbody.find_all('tr'):
    columnas = fila.find_all('td')

    if(columnas != []):
       nombre = columnas[0].text
       enlace = columnas[1].text
       nombre_lista.append(nombre)
       enlaces_lista.append(enlace)


df = pd.DataFrame({'tienda': nombre_lista, 'sitio_web' : enlaces_lista})
```

5.2.3 c) Iterar los elementos del elemento 'p'

```
elementos_p = html_bruto.find_all('p')

for i in range((len(elementos_p)-1),1):
                            print(elementos_p[i].get_text())
```

```
A continuación se listan 15 tiendas que comienzan con la letra A:
```

5.2.4 d) Elementos 'td' de clase marca y que terminan en 'o'

```
filas_marca = html_bruto.find_all('td', class_='marca')

filas_marca_texto_lista = []
for j in range(0,len(filas_marca),1):
                            texto = filas_marca[j].get_text()
                            filas_marca_texto_lista.append(texto)

marca_que_cumplen_condicion = []

for palabra in filas_marca_texto_lista:
    if palabra.endswith('o'):
        marca_que_cumplen_condicion.append(palabra)

print(marca_que_cumplen_condicion)
```

```
['Al Mundo', 'Anca & Co', 'Arredo', 'Atrápalo']
```

5.2.5 e) Guardar tabla en excel

```
df.to_excel('Tabla_ejercitación_1.xlsx',index=False)
```

6 Punto 6

6.1 Consigna

> Prepará una muy breve presentación para explicar en 7' al inicio del siguiente encuentro cuales son las características del sitio web que escogiste en el punto #1.

6.2 Respuesta

Presentación oral

7 Punto 7

7.1 Consigna

> Por último, documenta todo tu trabajo en un repositorio GitHub y comparte el enlace al equipo docente por Slack hasta el día previo al siguiente encuentro.

7.2 Respuesta

Documentado usando Quarto