



Web Scraping



DigitalHouse >
Coding School

DATA SCIENCE

UNIDAD 1
MÓDULO 2

Web Scraping con Python

Marzo 2017

El web scraping es la extracción automática de información de la web mediante el uso de bots o crawlers.

Web Crawler

El web crawler o *spider* recorre e indexa las páginas web para el procesamiento de su contenido.



1

Describir los fundamentos del web scraping

2

Conocer cómo implementar web scraping usando Python

- Requests
- BeautifulSoup
- Regex

<https://regexr.com/>

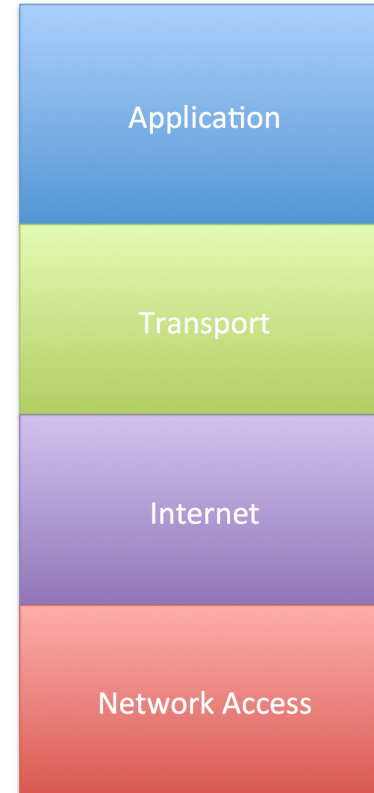
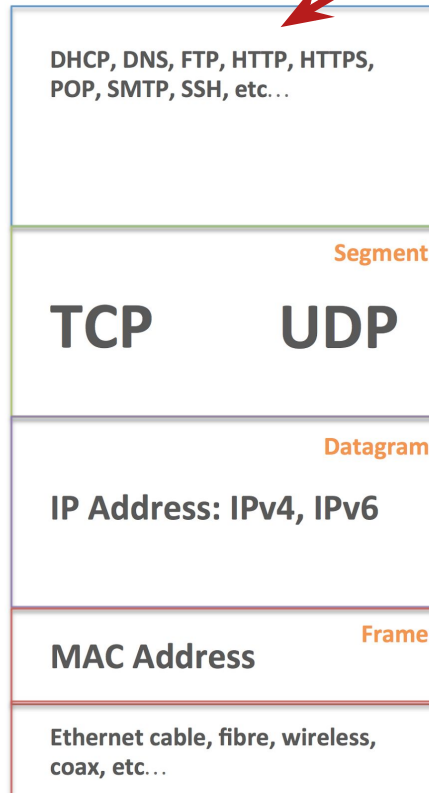
Internet



The OSI Model



The TCP/IP Model



headers

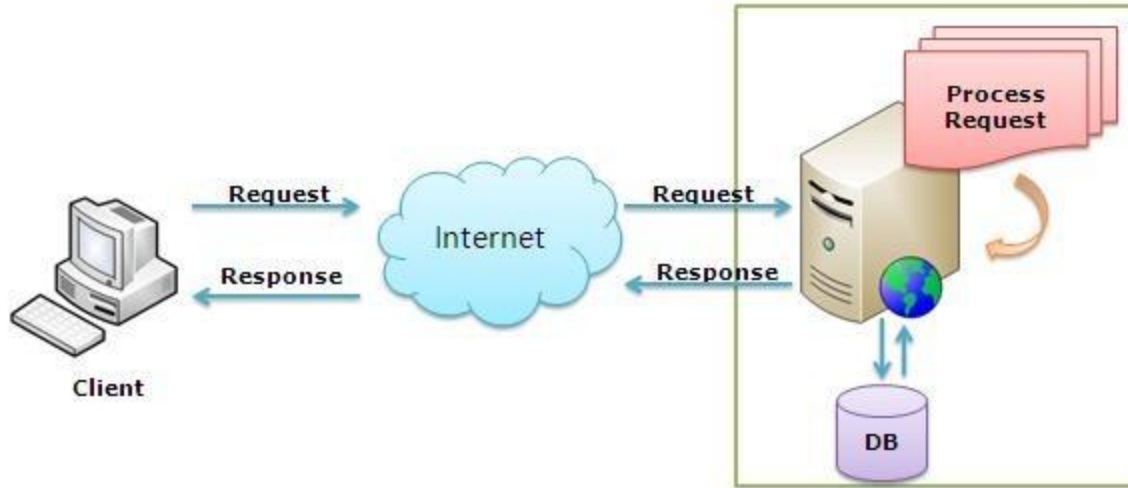
```
HTTP/1.1 200 OK
Date: Mon, 27 Jul 2009 12:28:53 GMT
Server: Apache/2.2.14 (Win32)
Last-Modified: Wed, 22 Jul 2009 19:15:56 GMT
Content-Length: 88
Content-Type: text/html
Connection: Closed
```

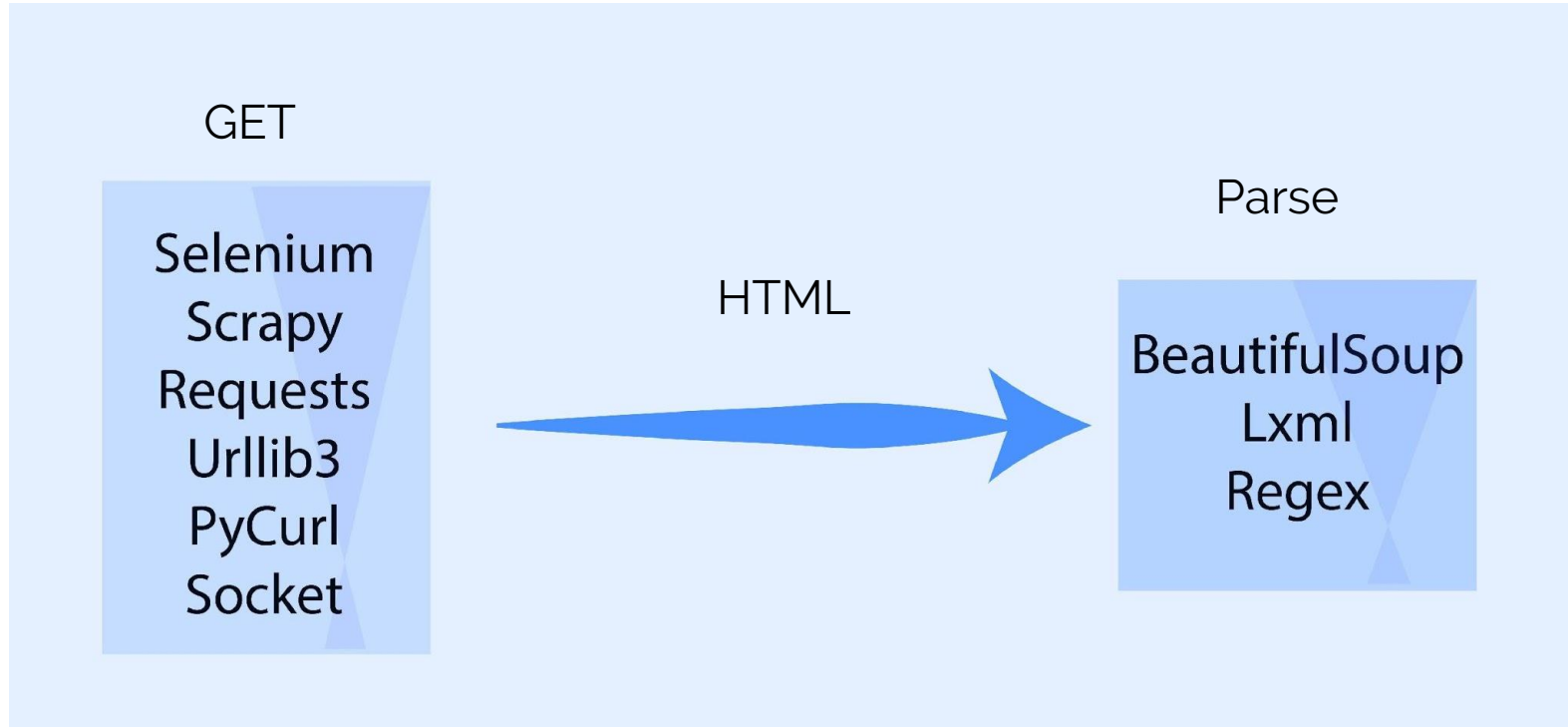
content

```
<html>
  <body>

    <h1>Hello, World!</h1>

  </body>
</html>
```





- El *HyperText Markup Language* es el lenguaje de marcado (*markup language*) estándar en la web.
- Permite armar textos estructurados enriquecidos con archivos multimedia. Junto con CSS y JavaScript forman la tríada de tecnologías centrales de la Web.

```
<!DOCTYPE html>
<html>

  <head>
    <title> Title here </title>
  </head>

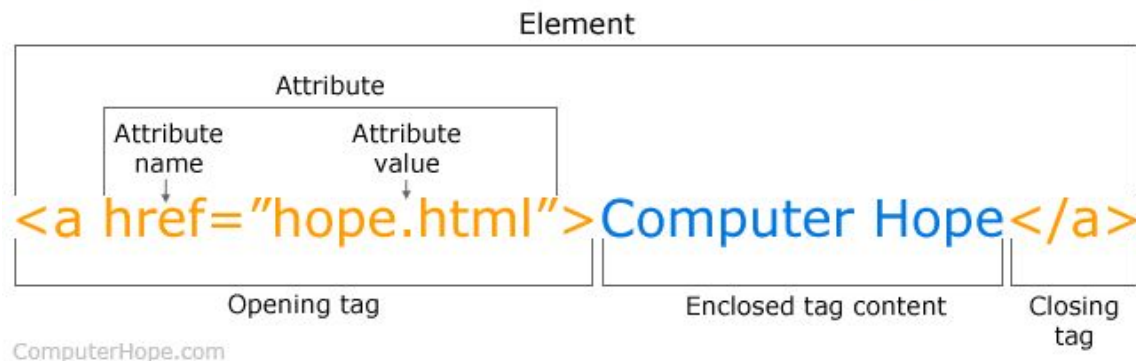
  <body>
    Web page content goes here.
  </body>

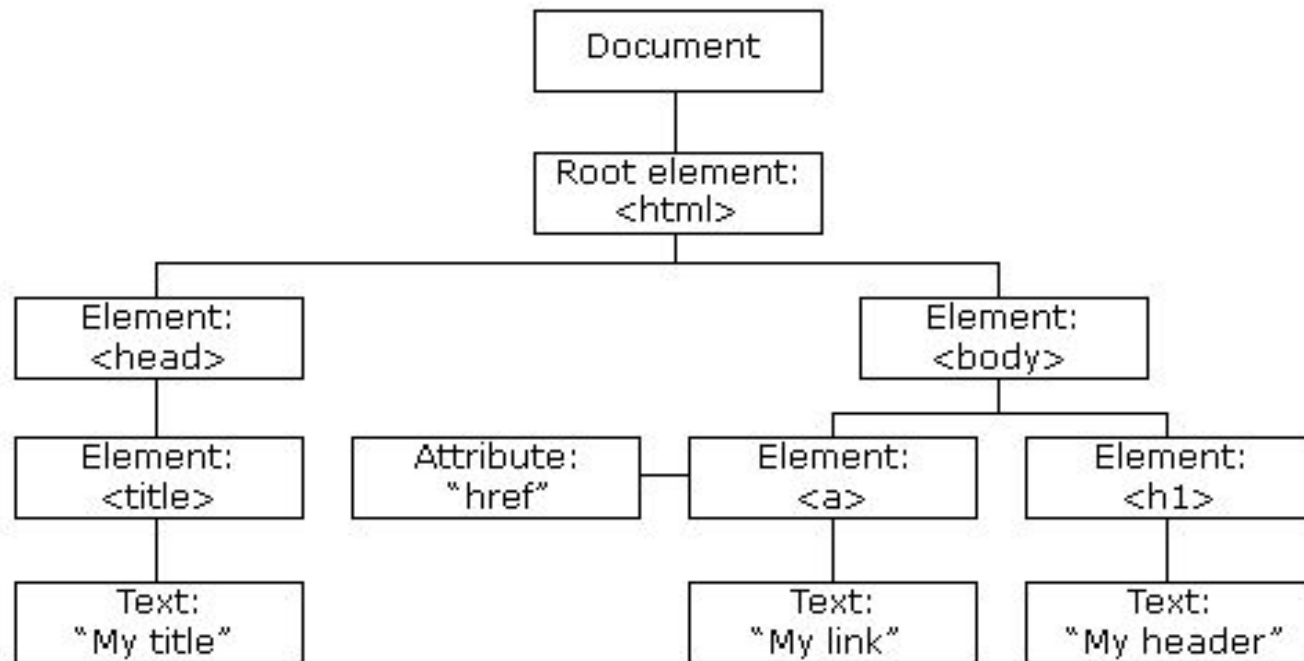
</html>
```

Se compone por elementos semánticos delimitados por etiquetas o *tags*, que se encapsulan en ángulos <>.

Los mismos pueden poseer atributos separados con espacio y con el valor entre comillas

Breakdown of an HTML Tag





ComputerHope.com

Gibson Les Paul - Guitarras Eléctricas Gibson en Mercado Libre Argentina - Google Chrome

Seguro | [https://listado.mercadolibre.com.ar/gibson-les-paul#D\[A:gibson-les-paul,L:1](https://listado.mercadolibre.com.ar/gibson-les-paul#D[A:gibson-les-paul,L:1)

Aplicaciones GitHub - jakevdp GitHub - cs109/c GitHub - jdwitter GitHub - whitea GitHub - donnem A gallery of inter GitHub - hangtw GitHub - amuelle

mercado libre

Regístrate Ingresar Vender

Guitarras > Eléctricas > Gibson

Gibson les paul

568 resultados

Ordenar publicaciones

Más relevantes |

Condición

Nuevo (215)

Usado (351)

Ubicación

Capital Federal (240)

Bs.As. G.B.A. Sur (170)

Bs.As. G.B.A. Oeste (42)

Bs.As. G.B.A. Norte (37)

Santa Fe (22)

Córdoba (13)

Buenos Aires Interior (12)

Bs.As. Costa Atlántica (10)

\$ 27.707

Guitarra Gibson Les Paul 2017
Tribute Faded Honey Burst - JM

1 vendido

\$ 27.000

Hasta 6 cuotas sin interés

Abrir enlace en una pestaña nueva

Abrir enlace en una ventana nueva

Abrir el enlace en una ventana de navegación de incógnito

Guardar enlace como...

Copiar dirección de enlace

AdBlock

Inspeccionar Ctrl+Mayús+I

```
distance">...</div>
<div id="searchResults" class="section search-results list-view grid
search-results-core" imgtype="listingLightweight">
::before
<li class="results-item article grid item-info-height-139">
<div class="rowItem item item--grid new item-over" id=
"MLA647417868">
<div class="item_image item_image--grid">...</div>
<span class="item-loading_status-bar item-loading_hide">...</span>
<a href="https://articulo.mercadolibre.com.ar/MLA-647417868-
guitarra-gibson-les-paul-2017-tribute-faded-honey-burst-JM" class=
"item_info-link item_js-link">
<div class="item_info">
<div class="item_price">...</div>
<h2 class="item_title list-view-item-title">
<span class="main-title"> Guitarra Gibson Les Paul 2017
Tribute Faded Honey Burst </span> == $0
</h2>
<div class="item_status">...</div>
</div>
</a>
<form class="item_bookmark-form" action="/search/bookmarks/
MLA647417868/make" method="post" id="bookmarkForm">...</form>
::after
</div>
</li>
<li class="results-item article grid item-info-height-139">...</li>
<li class="results-item article grid item-info-height-139">...</li>
<li class="results-item article grid item-info-height-139">...</li>
<li class="results-item article grid item-info-height-139">...</li>
on #searchResults li #MLA647417868 a div h2 span.main-title
```

Entre las opciones para usar HTTP con Python, la librería recomendada oficialmente es Requests.

Está escrita sobre urllib3 y http.client, y ofrece un cómodo nivel de abstracción.

Resuelve fácilmente autenticaciones, HTTPS, persistencia de cookies...



- Una expresión regular es una secuencia de caracteres que determina un patrón de búsqueda
- Es un lenguaje muy flexible que sirve para identificar y extraer información de un cuerpo de caracteres no estructurado.

() capturing group

(?:) non-capturing group

\w carácter alfanumérico

. cualquier cosa menos \n

\d dígito

| operador "or"

\s whitespace

[m-z3-9] rangos

[] conjunto

Cuantificadores:

- + uno o más del elemento anterior
- * cero o más del elemento anterior
- {4,} cuatro o más del elemento anterior
- ? cambia el operador anterior de "greedy" a "lazy". Ejemplo:

Lazy
`(.*?)<`

`contacto@digitalhouse.com<
b>Digital House`

Greedy `(.)<` `contacto@digitalhouse.comDigital House`

Esta librería permite obtener la información contenida en HTML/XML y extraerla en un formato ordenado.

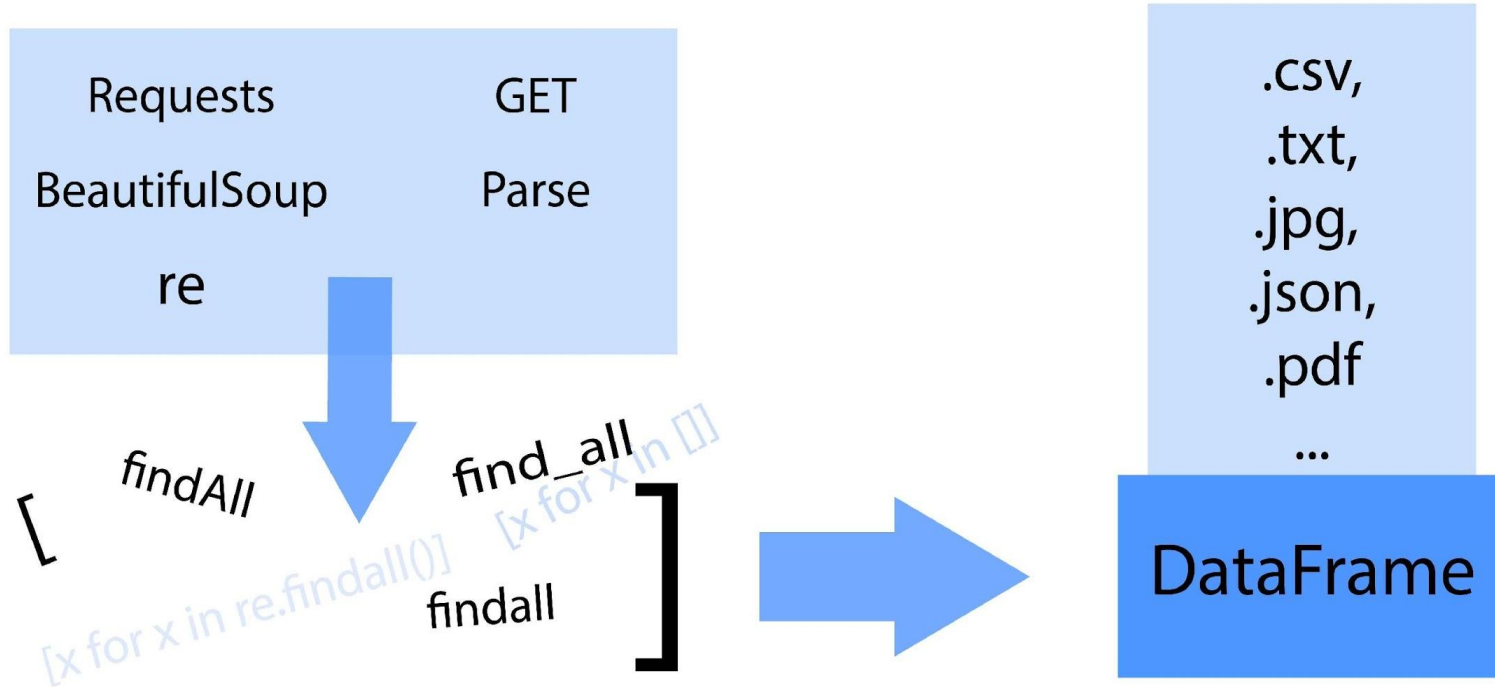
Tolera documentos HTML mal formados, ya que está implementado con Regular Expressions, y no parseando el HTML a una estructura DOM.

Permite distintos parsers, entre los que se incluye lxml.



Práctica Guiada I

Práctica Guiada II



Laboratorio

- Esta técnica permite obtener información no estructurada y normalmente poco accesible en un formato simple para un posterior análisis.
- El web scraping es de interés cuando se necesita acceder a información dinámica.
- Python nos facilita el acceso a la información de interés sin la necesidad de tener un conocimiento exhaustivo de HTML o XML, utilizando librerías de distinto nivel entre las cuales están Requests y BeautifulSoup.

Para persistencia de cookies:

requests.Session()

Cuando se necesita ejecutar JavaScript:

Selenium

Para mayor rendimiento temporal:

Lxml

Extras de infraestructura:

Threading

Queue