

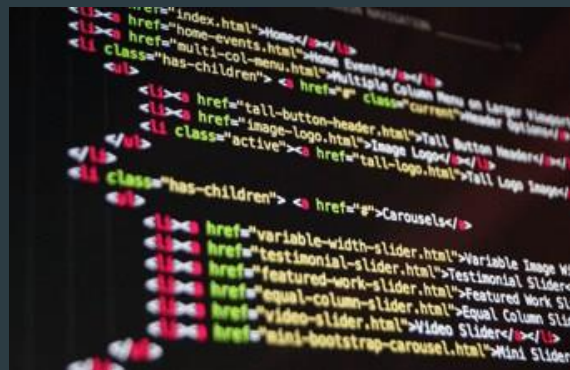
Web scraping para principiantes

Sara Rodríguez López



minsa1t
by Indra

NUESTRO OBJETIVO ES EXTRAER INFORMACIÓN
DE UNA PÁGINA WEB Y TRANSFORMARLA EN
INFORMACIÓN ESTRUCTURADA QUE PODEMOS
ALMACENAR Y ANALIZAR



Las webs están implementadas en HTML, un lenguaje basado en etiquetas

- ▶ Todos los elementos de una web tienen una etiqueta de comienzo `<element>` y otra de cierre `</element>`

```
<div>
  <h1>Esto es un título</h1>
  <p>Y esto un párrafo. Y ambos están dentro de un div</p>
</div>
```

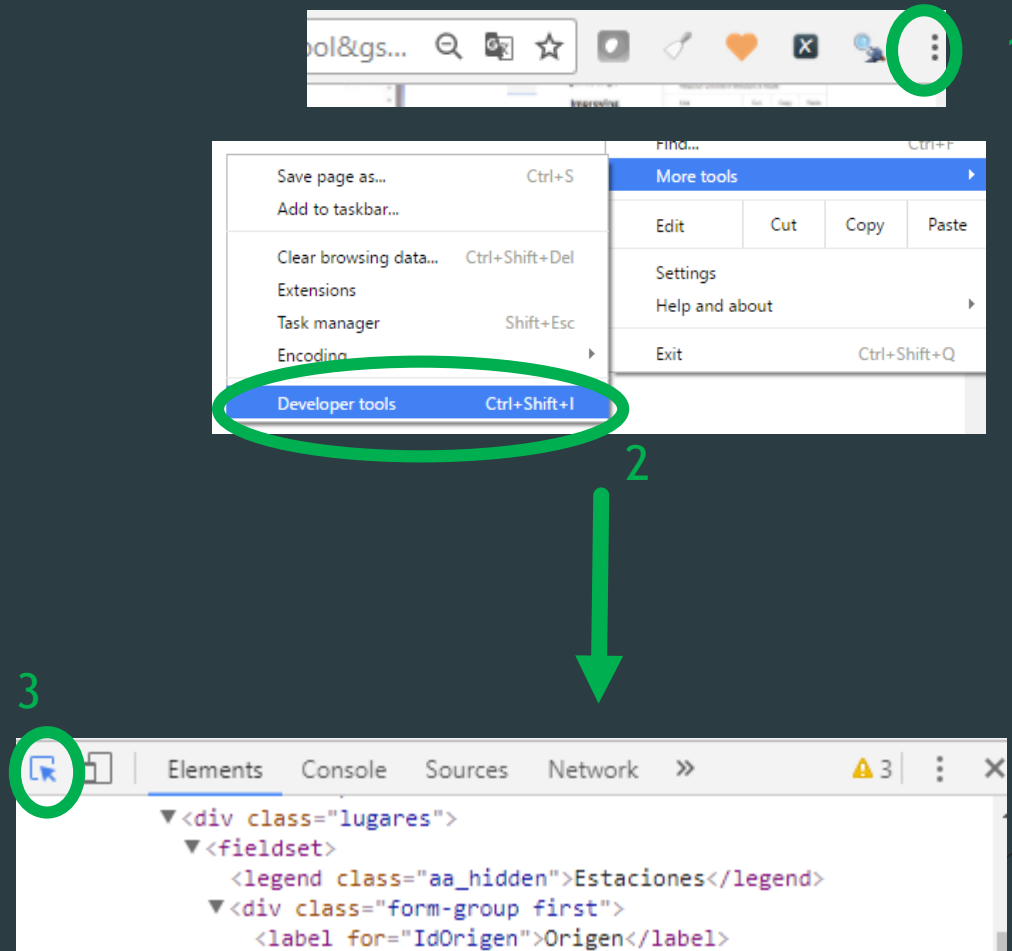
- ▶ Los *div* son elementos que me permiten agrupar varios elementos distintos en un mismo bloque.
- ▶ Los elementos de HTML pueden tener:
 - ▶ **Ids**: Diferencia a un element del resto de la página.

```
<p id="comentario_1">Primer comentario</p>
```

- ▶ **Classes**: Usado para categorizar los elementos, elementos similares tienen la misma clase.

```
<h1 class="titulo_grande">Un título</h1>
```

Utilizaremos las herramientas para desarrolladores de tu explorador para ver el código HTML



Vamos a revisar los conceptos de web scraping y web crawling

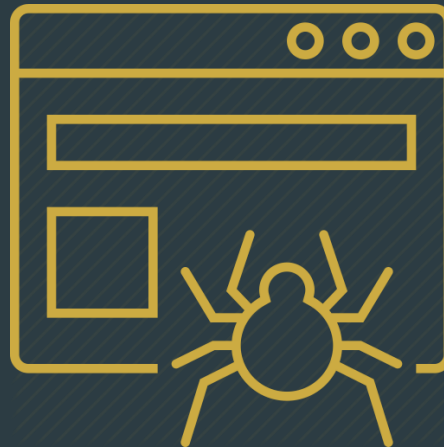
WEB SCRAPING

- ▶ ¿Qué es?
- ▶ ¿Qué librerías podría usar?
- ▶ Ejemplo con BeautifulSoup



WEB CRAWLING

- ▶ ¿Qué es?
- ▶ ¿Qué librerías podría usar?
- ▶ Ejemplo con Selenium

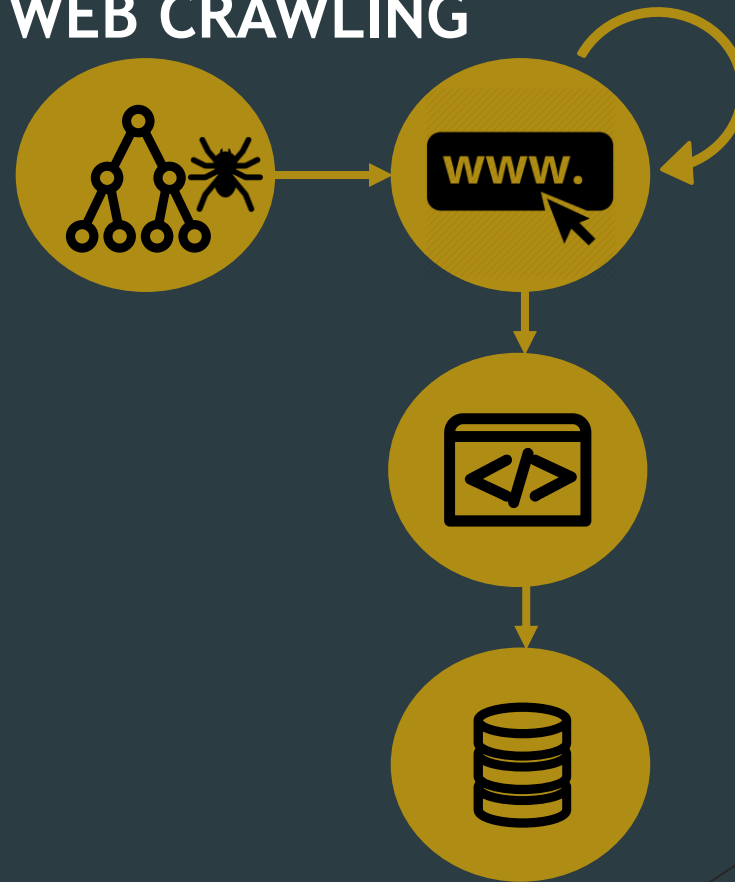


Diferencia entre web scraping y web crawling

WEB SCRAPING



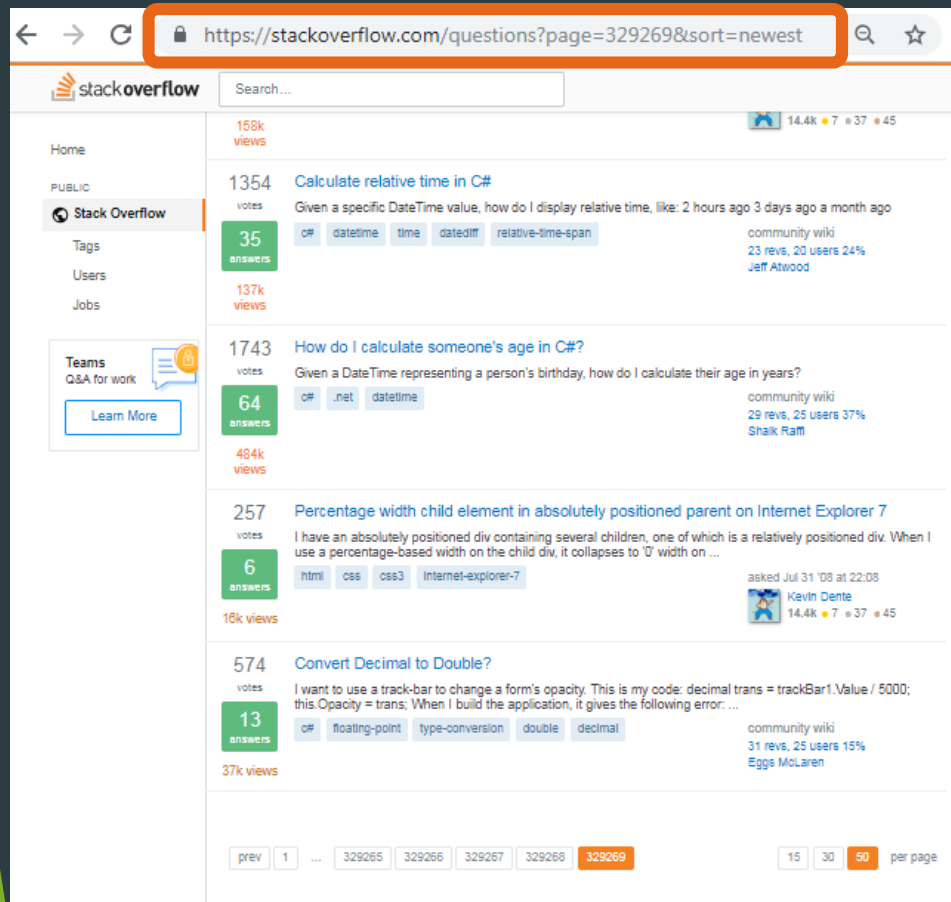
WEB CRAWLING



Ejemplos

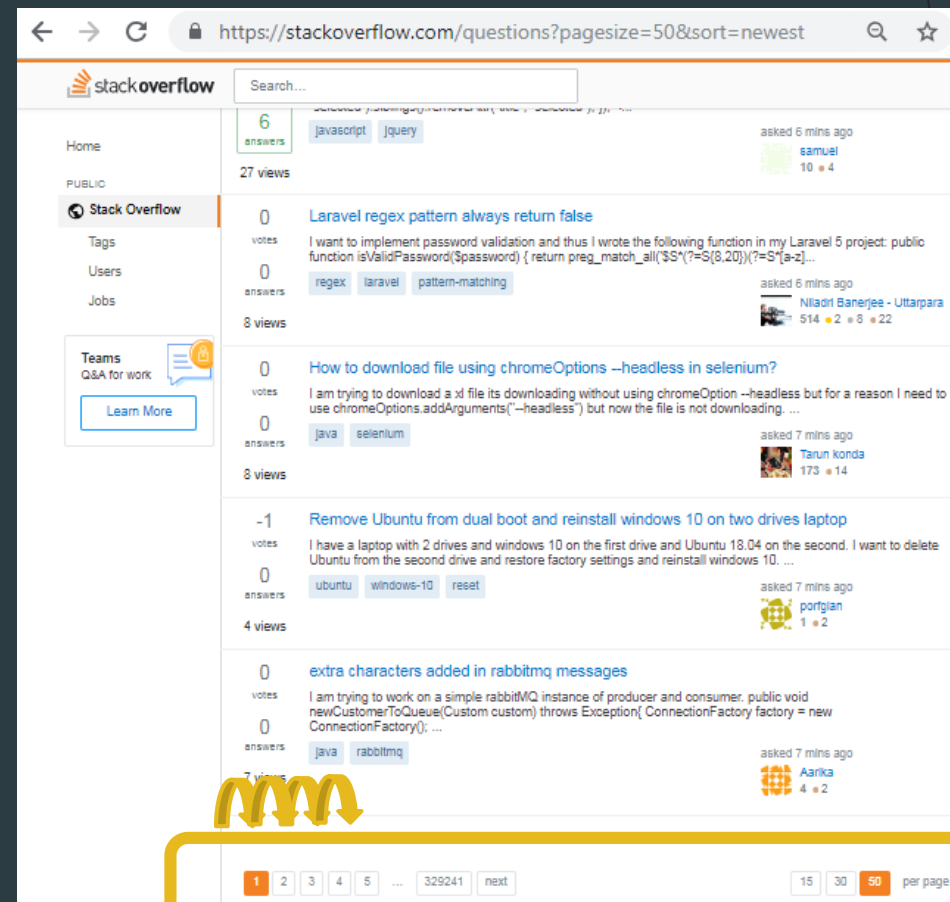
WEB SCRAPING

“Extraer las 50 primeras preguntas hechas en stackoverflow.”



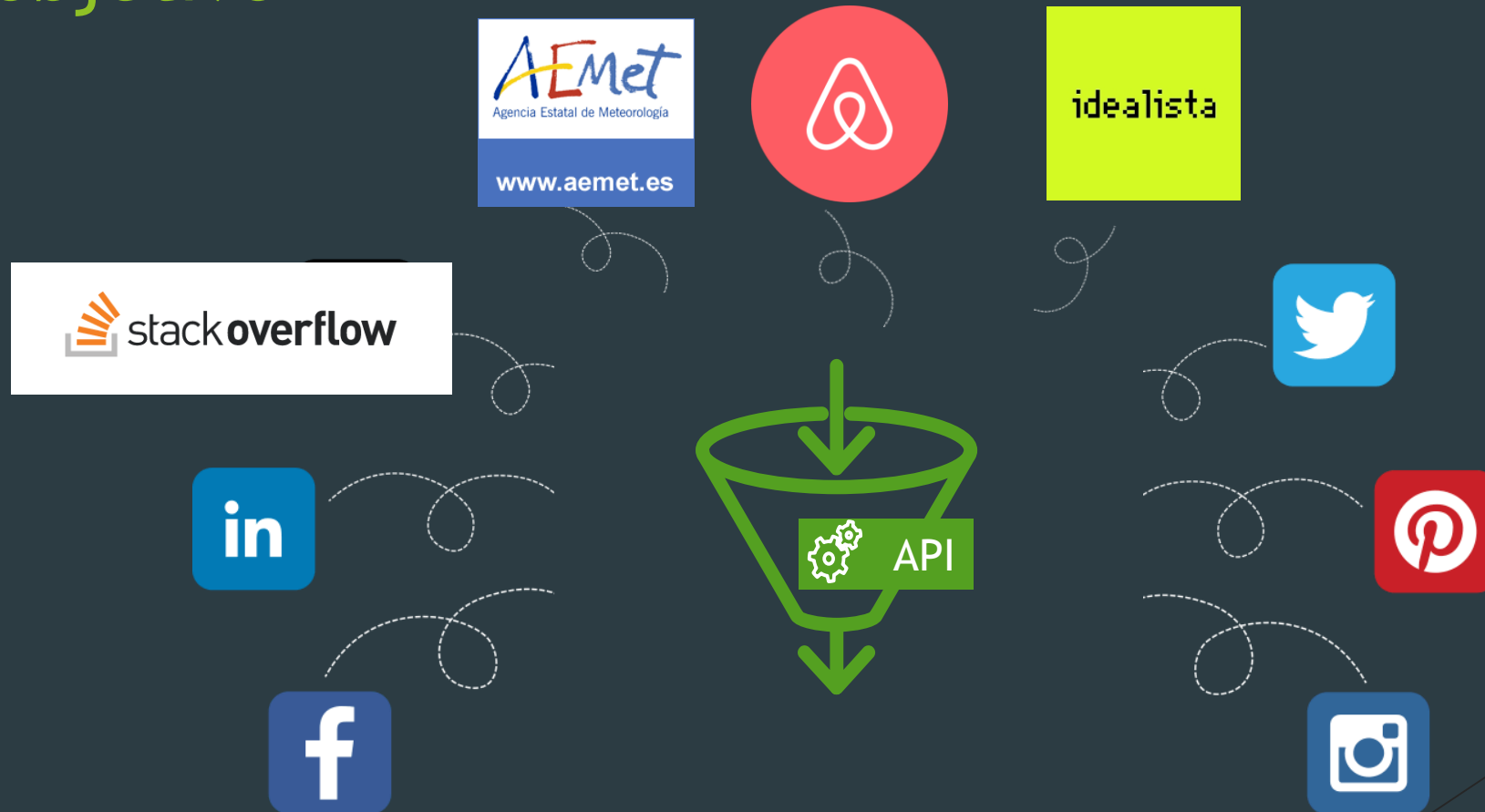
WEB CRAWLING

“Extraer todas las preguntas hechas en stackoverflow desde el comienzo hasta hoy.”





Comprueba antes si existe una API, valora las limitaciones de la misma y balancea con tu objetivo



WEB SCRAPING



¿Qué librerías existen para hacer web scraping?

- ▶ requests
- ▶ urllib/urllib2
- ▶ httplib/httpplib2
- ▶ BeautifulSoup (bs4)
- ▶ lxml
- ▶ re
- ▶ scrapy

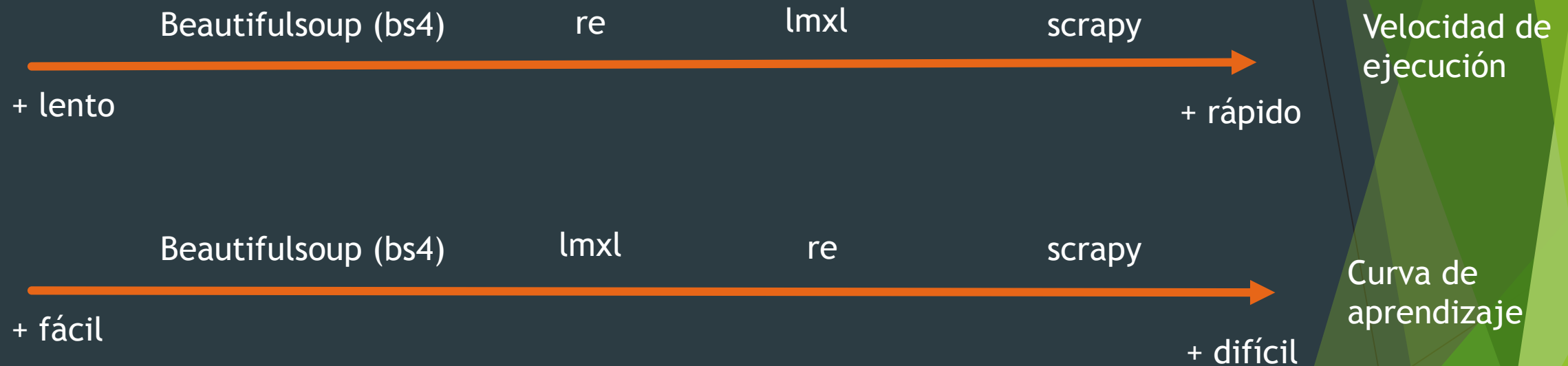
CONECTAR CON LA
WEB Y EXTRAER
CÓDIGO

PARSEAR
CÓDIGO HTML

TRATAR Y
GUARDAR INFO
DE INTERÉS

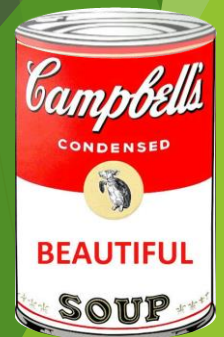


¿Qué librerías selecciono para hacer web scraping?



Para crear un script de manera rápida, con conocimientos no muy avanzados de python y con el que extraer poca información sin importar la velocidad de ejecución:


beautifulsoup.



Vamos a extraer de tripadvisor los 30 mejores restaurantes de Málaga ciudad, presentes en la primera página de la búsqueda

TripAdvisor LLC [US] | https://www.tripadvisor.es/Restaurants-g187438-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html


Reserva: 17/9/2018 20:00 2 clientes Encuentra un restaurante Borrar búsqueda



Luxalad
★★★★★ 127 opiniones
1 de 1.972 Restaurantes en Málaga
€, Comida rápida, Europea, Saludable, Opciones vegetarianas, Opciones veganas, (

"La belleza de lo simple" 11/09/2018
"Recomendable 100%" 05/09/2018

Hacer pedido online




Trattoria Mamma Franca
★★★★★ 343 opiniones
2 de 1.972 Restaurantes en Málaga
€€ - €€€, Italiana, Pizza, Mediterránea, Europea, Opciones vegetarianas, Opcione...

"Buena comida y gran atencion" 15/09/2018
"El sabor de la auténtica Italia" 15/09/2018

¡Mesas disponibles para esta noche!

20:30
21:00
21:30
[Ver otras horas](#)



La Barra de Zapata
★★★★★ 488 opiniones
3 de 1.972 Restaurantes en Málaga
€€ - €€€, Internacional, Mediterránea, Europea, Española, Opciones vegetarianas, C

Primero comprobamos si la web nos ofrece una API y lo fácil que es trabajar con ella.

TripAdvisor LLC [US] | <https://www.tripadvisor.com/APIAccessSupport>

tripadvisor

Hotels Vacation Rentals Flights Restaurants Things to do ...

Request TripAdvisor API Access

TripAdvisor grants a limited number of API keys to official tourism organizations and select other websites. It does not grant access for purposes of data analysis, research, testing, or similar uses.

Due to the high volume of requests, we are unable to respond to all applications. Please make sure that all fields are filled out completely and accurately, that your request includes a working URL, and that your intended use of TripAdvisor ratings is fully explained.

First Name

Last Name

Company

Company Location
Select one

Business Type
Select one

E-mail address

Phone number (if outside US & Canada)

On which website or app do you propose to use the API?
Name of website or app

URL (must be a working URL specific to your website and/or app, not your company)

Platform Check all that apply.
☐ Desktop ☐ Mobile

How many monthly unique visitors come to your website and/or app?
Select one

Description of how TripAdvisor content will be used/displayed
(if content is being requested for more than one URL, please list all applicable URLs in this field)

Determine Content API Eligibility

Thank you for your interest in the TripAdvisor Content API. Please note that this product is for the use of official tourism organizations and a very limited number of other partners only. Due to the volume of requests we cannot respond to all applications; if your request is approved you will be notified by email. Please select the boxes that apply to your proposed use of the Content API. You must select one, and may select multiple boxes.

☐ Official Tourism Organization/DMO/CVB

☐ Consumer-facing (B2C)

☐ Primarily used for data analysis and/or research

☐ Primarily used for B2B purpose

OK

....tardo menos en
hacerme un script....



```
import requests
url = "https://www.tripadvisor.es/Restaurants-g187438-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html"
peticion_restaurantes = requests.get(url)
textourl = peticion_restaurantes.text
```

```
import urllib
url = "https://www.tripadvisor.es/Restaurants-g187438-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html"
peticion_restaurantes = urllib.request.urlopen(url)
textourl = peticion_restaurantes.read()
```

```
print(texturl)
```

```
<!DOCTYPE html>\n<html xmlns:fb="http://www.facebook.com/2008/fbml" lang="es">\n<head>\n<meta http-equiv="content-type" content="text/html; charset=utf-8"/>\n<link rel="stylesheet" type="text/css" href="https://static.tacdn.com/css2/long_lived_global_legacy-v2318119761b.css" data-rup=\n'long_lived_global_legacy/'>\n<link rel="icon" id="favicon" href="https://static.tacdn.com/favicon.ico" type="image/x-icon"/>\n<link rel="preload" href="https://static.tacdn.com/css2/webfonts/TripAdvisorTripAdvisor-Regular.woff2v003.400.ass=font" type="font/woff2" crossorigin>\n<link rel="mask-icon" sizes="any" href="https://static.tacdn.com/img2/2008/tasquare.svg" color="#00a680"/>\n<script type="text/javascript" data-rup=\n'global_error'>\n!function(){function e(e,t,n,o,i,a){var d={error_script:t,line:n,column:o,ready_state:document.readyState};return s?((require.defined("@ta/util/Error"))&&require("@ta/util/Error")).record(i,"error post load: "+e,a,d,"ERROR",{isglobal:!0}),void r(e,i,"ErrorGlobal")):(f.push([msg:e,l:"","error:i,evt:a,data:d"],window.IS_DEBUG)?function n(e,r,t){if(require.defined("@ta/platform.sentry"))){var n=require("@ta/platform.sentry")["default"];if(n){if(e&&!r){var o=new Error("Unknown jQuery Error Event"),i=JSON.stringify(e);i.length>200&&(i=i.substring(0),Math.min(i.length,200));"..."},n.captureException(o,{logger:t,extra:{jQueryEvent:i}})}else n.captureException(r,{logger:t},function t(){require("@ta/util/Error"),function(e){for(i:f.length){var t=f.sift(i);t.msg.match(/^(^[\^\\w.])ta.*defin/)||e.record(t.error,"window.onerror: "+t.msg,t.evt,t.data,"ERROR",{isglobal:!0})}r(t.msg,t.error,"PageLoad")}}s=!0)}function n(){c=null,E=1,d=!null?function o(r,t,o,i,a,w){var s=w&&w.target;if(E){if(![d|&a&a.stack]&&(d=a,l))try{u=arguments.callee}catch(e){c=s,[s]=null&s=window}&&(s=1),e(r,t,o,i,d,{target:s,callee:u}),n()}else{d=a,E=!0,l=s;try{u=arguments.callee}catch(e){}}function i(e){e=e||window.event,o=e.message,e=e.filename,e.lineno,e.colno,e.error||[e,e]}function a(e){e=e||window.event,c=e.target||e.srcElement,w&&clearTimeout(w),w=setTimeout(function(){w=0,c=null,l=1})var d,l,c,u,w,s=a[e],E=!1;window.onerror=function(e,r,t,n,i){return o(e,r,t,n,i),window.event,function(IS_DEBUG)ow.addEventlistener(window).addEventlistener("error",i,1),window.addEventlistener("click",a,!0),window.addEventlistener("load
```

PARSEAR
CÓDIGO HTML



Creamos el objeto soup para ello le pasamos la variable string que contine la info de la web previamente capturada y el parser a emplear.

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(textourl,"lxml")
# Selecciona aquí el parseador https://www.crummy.com/software/BeautifulSoup/bs4/doc/#installing-a-parser (html.parser, lxml,...)
```


PARSEAR
CÓDIGO HTML



Cualquier objeto de `BeautifulSoup` tiene el método `.find` que permite buscar cualquier tipo de elemento HTML, y se queda con la primera ocurrencia.

Este método toma como argumentos interesantes:

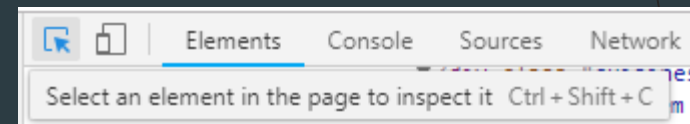
1. Un **string con el nombre del elemento a buscar** (por ejemplo `"h1"`, `"p"` o `"div"`).
2. Un argumento opcional llamado `id` que permite buscar solo elementos con el `id` especificado como string.
3. Un argumento opcional llamado `class_` que permite limitar la búsqueda a los elementos del HTML que tengan dicha `class`.

PARSEAR
CÓDIGO HTML



WEB SCRAPING

Vamos a buscar extraer el nombre del primer restaurante



a.property_title | 70.05 x 22

Luxalad

136 opiniones

1 de 2.050 Restaurantes en Málaga

€, Comida rápida, Europea, Saludable,...

"Reataurante" 17/09/2018

"La belleza de lo simple" 11/09/2018

```
<a target="_blank" href="/Restaurant Review-  
g187438-d13864986-Reviews-Luxalad-  
Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia  
.html" class="property_title" onclick=  
"ta.restaurant_list_tracking.clickDetailTitle('/  
Restaurant_Review-g187438-d13864986-Reviews-  
Luxalad-  
Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia  
.html', 'tags_category_tag_restaurants', '13864986'  
, '1', '1');">  
Luxalad  
</a> == $0
```

```
soup.find(class_="property_title")
```

```
<a class="property_title" href="/Restaurant_Review-g187438-d13864986-Reviews-Luxalad-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html" onclick="ta.restaurant_list_tracking.clickDetailTitle('/Restaurant_Review-g187438-d13864986-Reviews-Luxalad-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html','tags_category_tag_restaurants','13864986','1','1');" target="_blank">
```

Luxalad

```
</a>
```

```
soup.find(class_="property_title")  
<a class="property_title" href="/Restaurant_Review-g187438-d13864986-Reviews-Luxalad-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html" onclick="ta.restaurant_list_tracking.clickDetailTitle('/Restaurant_Review-g187438-d13864986-Reviews-Luxalad-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html','tags_category_tag_restaurants','13864986','1','1');" target="_blank">  
Luxalad  
</a>
```

Ahora podemos utilizar el método `.get_text` para quedarnos con el texto en bruto

```
soup.find(class_="property_title").get_text()  
'\nLuxalad\n'
```

Para eliminar los saltos de línea al principio y al final utilizamos el parámetro *strip*

```
soup.find(class_="property_title").get_text(strip=True)  
'Luxalad'
```

PARSEAR
CÓDIGO HTML



Si ahora queremos extraer el nombre de todos los restaurantes...

Cualquier objeto de `BeautifulSoup` tiene también el método `.find_all()`.

Este método funciona exactamente igual que `.find()`, pero con la diferencia de que busca todos los elementos que satisfacen la búsqueda (en lugar de solo el primero); y los devuelve en una lista de Python.

```
soup.find_all(class_="property_title")
```

```
[<a class="property_title" href="/Restaurant_Review-g187438-d13864986-Reviews-Luxalad-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html" onclick="ta.restaurant_list_tracking.clickDetailTitle('/Restaurant_Review-g187438-d13864986-Reviews-Luxalad-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html','tags_category_tag_restaurants','13864986','1','1');" target="_blank">
  Luxalad
</a>
<a class="property_title" href="/Restaurant_Review-g187438-d12699400-Reviews-Trattoria_Mamma_Franca-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html" onclick="ta.restaurant_list_tracking.clickDetailTitle('/Restaurant_Review-g187438-d12699400-Reviews-Trattoria_Mamma_Franca-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html','tags_category_tag_restaurants','12699400','2','2');" target="_blank">
  Trattoria Mamma Franca
</a>
<a class="property_title" href="/Restaurant_Review-g187438-d5804270-Reviews-La_Alacena_de_Francis-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html" onclick="ta.restaurant_list_tracking.clickDetailTitle('/Restaurant_Review-g187438-d5804270-Reviews-La_Alacena_de_Francis-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html','tags_category_tag_restaurants','5804270','3','3');" target="_blank">
  La Alacena de Francis
</a>
<a class="property_title" href="/Restaurant_Review-g187438-d14158262-Reviews-Marisqueria_La_Mayor-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html" onclick="ta.restaurant_list_tracking.clickDetailTitle('/Restaurant_Review-g187438-d14158262-Reviews-Marisqueria_La_Mayor-Malaga_Costa_del_Sol_Province_of_Malaga_Andalucia.html','tags_category_tag_restaurants','14158262','4','4');" target="_blank">
```

```
[i.get_text(strip=True) for i in soup.find_all(class_="property_title")]
```

```
['Luxalad',
 'Trattoria Mamma Franca',
 'La Alacena de Francis',
 'Marisquería La Mayor',
 'La Barra de Zapata',
 'Café Tramezzino',
 'Byoko',
 'Kortxo',
 'Da Saveria comida italiana casera',
 'Spago's - Fresh Pasta',
 'Brutus',
 'Mura Mura Osteria Cafe',
 'La Récréation',
 'Restaurante Cávea',
 'La Luz de Candela',
 'La Tranca',
```

Inspeccionamos el código HTML buscando extraer información adicional cómo la valoración, el número de opiniones, el tipo de comida,...

Luxalad span.reviewCount | 87.17 x 16

127 opiniones

1 de 1.972 Restaurantes en Málaga

€, Comida rápida, Europea, Saludable, Opci...

"La belleza de lo simple" 11/09/2018

"Recomendable 100%" 05/09/2018

Luxalad div.popIndex.rebrand.popIndexDefault | 284.25 x 16

1 de 1.972 Restaurantes en Málaga

€, Comida rápida, Europea, Saludable, Opci...

"La belleza de lo simple" 11/09/2018

"Recomendable 100%" 05/09/2018

span.ui_bubble_rating.bubble_50::after | 90 x 20

127 opiniones

1 de 1.972 Restaurantes en Málaga

€, Comida rápida, Europea, Saludable, Opci...

"La belleza de lo simple" 11/09/2018

"Recomendable 100%" 05/09/2018

Repetimos la lógica empleada con la extracción del primer elemento y combinamos el uso de las funciones: *find*, *find_all*, *get* y *get_text*.

```
import requests
from bs4 import BeautifulSoup
import pandas as pd

def get_tripadvisor(URL):
    # Nombre ciudad
    nombreCiudad = URL.split("-")[2].split("_")[0]
    # Conexión
    petition_restaurantes = requests.get(URL)
    textourl = petition_restaurantes.text # is a string that contains the web page source
    # Compruebas conexión ok
    if petition_restaurantes.status_code == 200:

        # Objeto beautiful soup
        soup = BeautifulSoup(textourl, "lxml")

        # creas listas que irás relenando con datos
        name = []
        position = []
        rating = []
        numReview = []
        euros = []
        food = []

        # Parseas el código html extrayendo la información de interés
        for sec in soup.find_all(class_="shortSellDetails"):
            # posicion
            pos = sec.find(class_="popIndex rebrand popIndexDefault").get_text(strip=True) # extraccion
            pos = int(pos.split(" de ")[0]) # tratamiento
            position.append(pos) # append

            # restaurant name
            nam = sec.find(class_="property_title").get_text(strip=True) # extraccion
            name.append(nam)

            # rating
            rate = sec.find(class_="ui_bubble_rating").get("alt") # extraccion
            rate = float(rate.split(" de ")[0].replace(",",".")) # tratamiento
            rating.append(rate)

            # number or reviews
            reviews = sec.find(class_="reviewCount").get_text(strip=True) # extraccion
            reviews = int(reviews.replace("opiniones", "").replace(".", "")) # tratamiento
            numReview.append(reviews)
```

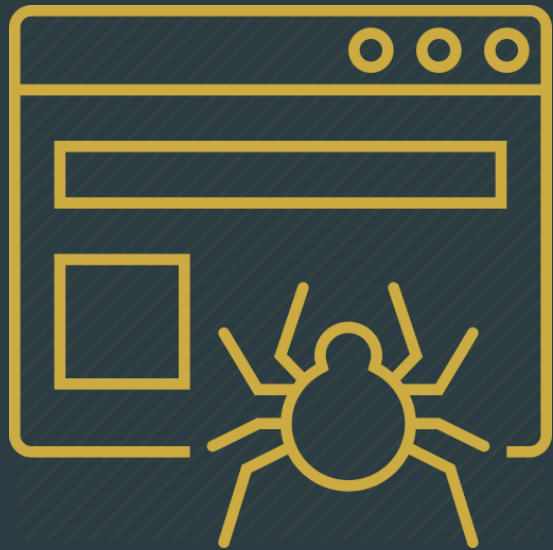
Código explicado paso a paso y completo disponible en:
<https://sararl.github.io/>

Y voilà! Obtenemos una tabla resumen con toda la información de los 30 primeros restaurantes.

	restaurants	ratings	number_reviews	price_category	food_category
Malaga					
1	Luxalad	5.0	139	€	Comida rápida,Europea,Saludable,Opciones veget...
2	Trattoria Mamma Franca	5.0	352	€€ - €€€	Italiana,Pizza,Mediterránea,Europea,Opciones v...
3	La Alacena de Francis	5.0	327	€€ - €€€	Mediterránea,Europea,Española,Rusa,Opciones ve...
4	La Barra de Zapata	5.0	507	€€ - €€€	Internacional,Mediterránea,Europea,Española,Op...
5	Marisquería La Mayor	5.0	138	€€ - €€€	Marisco,Mediterránea,Española
6	Café Tramezzino	5.0	134	€	Café,Europea,Saludable,Opciones veganas,Opcion...
7	Byoko	5.0	278	€€ - €€€	Francesa,Café,Mediterránea,Española,Saludable,...
8	Kortxo	5.0	417	€€ - €€€	Española,Internacional,Opciones vegetarianas,O...
9	Da Saveria comida italiana casera	4.5	632	€	Italiana,Mediterránea,Europea,Opciones vegetar...
10	Spago's - Fresh Pasta	5.0	337	€	Italiana,Comida rápida,Mediterránea,Europea,Op...
11	Brutus	4.5	203	€€ - €€€	Americana,Latina,Pizza,Saludable,Opciones vege...
12	Mura Mura Osteria Cafe	4.5	708	€€ - €€€	Italiana,Fusión,Mediterránea,Europea,Tienda go...
13	La Luz de Candela	4.5	807	€€ - €€€	Francesa,Mediterránea,Europea,Española,Opcione...
14	Restaurante Cávea	4.5	119	€€ - €€€	Española,Fusión,Mediterránea,Europea,Opciones ...
15	La Récréation	4.5	270	€€ - €€€	Francesa,Mediterránea,Fusión,Pub restaurante,O...
16	La Tranca	4.5	825	€	Bar,Mediterránea,Española,Opciones vegetarianas
17	Riosol	4.5	345	€€ - €€€	Mediterránea,Española,Opciones vegetarianas,Op...
18	Asador Ovidio	4.5	626	€€ - €€€	Asador,Mediterránea,Barbacoa,Europea,Española,...
19	García Taberna	4.5	458	€€ - €€€	Internacional,Mediterránea,Europea,Española,Co...
20	Pampa Grill Restaurante Argentino	4.5	336	€€ - €€€	Asador,Barbacoa,Argentina,Opciones sin gluten
21	La Recova	4.5	789	€	Española
22	Pizzeria Italiana Vittoria	4.5	384	€€ - €€€	Italiana,Pizza,Mediterránea,Europea,Opciones v...
23	Prohibitox	4.5	198	€€ - €€€	Europea,Asiática,Fusión,Internacional,Opciones...
24	Buenavista Gastrobar & Tapas	4.5	295	€€ - €€€	Mediterránea,Europea,Española,Fusión,Opciones ...
25	Siete Semillas	5.0	100	€€ - €€€	Mediterránea,Saludable,Española,Opciones veget...
26	La Proa de Teatinos	4.5	704	€€ - €€€	Marisco,Mediterránea,Europea,Española,Opciones...
27	Restaurante Alexso	4.5	303	€€ - €€€	Mediterránea,Europea,Española,Contemporánea,Fu...
28	Las Golondrinas	4.5	231	€€ - €€€	Mediterránea,Europea,Española,Asador,Opciones ...
29	El Meson de Cervantes	4.5	4992	€€ - €€€	Mediterránea,Europea,Española,Fusión,Opciones ...
30	Mesón Alberto	4.5	224	€€ - €€€	Española

Código explicado paso a paso y completo disponible en:
<https://sararl.github.io/>

WEB CRAWLING



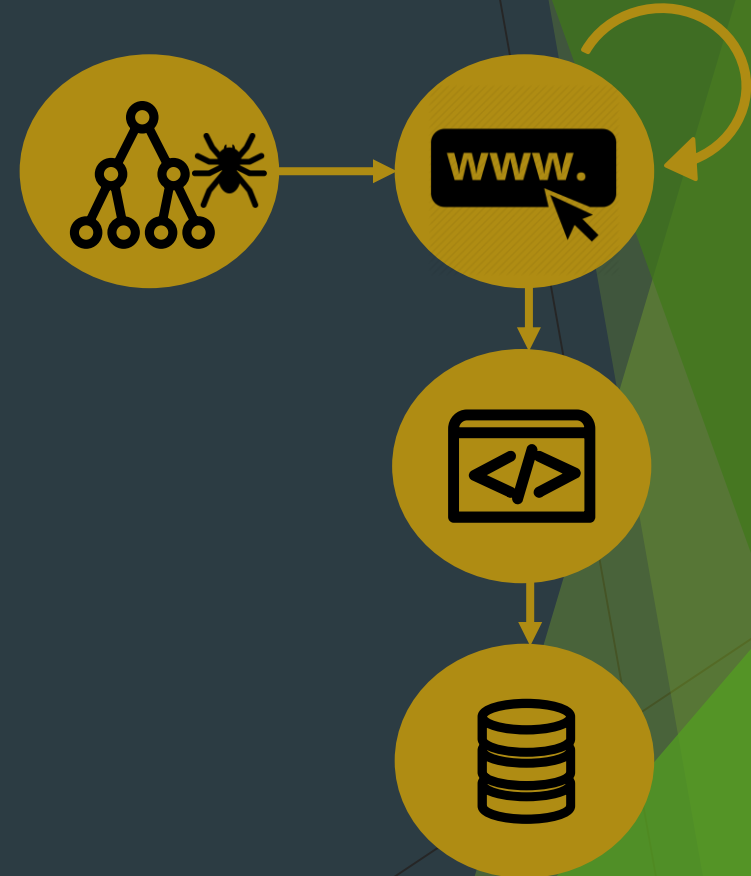
¿Qué librerías existen para hacer web crawling?

- ▶ ▶ Selenium
- ▶ ▶ scrapy
- ▶ BeautifulSoup (bs4)
- ▶ lxml
- ▶ re

CONECTAR CON LA
WEB, EXTRAER
CÓDIGO Y
NAVEGAR POR LA
WEB

PARSEAR
CÓDIGO HTML

TRATAR Y
GUARDAR INFO
DE INTERÉS



¿Qué librerías selecciono para hacer web crawling?

Para crear un script de manera rápida, con el que extraer poca información sin importar la velocidad de ejecución:

Selenium, o si ya tienes conocimientos de BeautifulSoup puedes combinar ambas.



Veremos como funciona Selenium extrayendo todos los restaurantes de Málaga en Tripadvisor

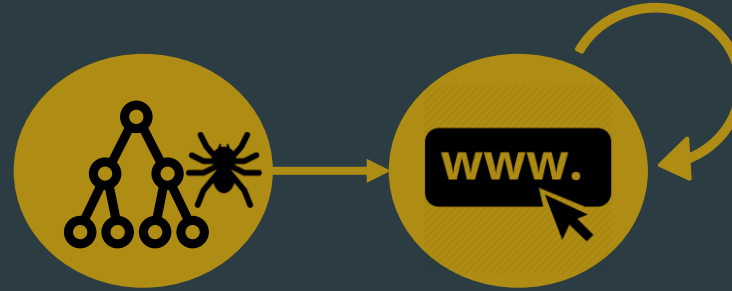
	position	restaurants	ratings	number_reviews	price_category	food_category
0	1	Luxalad	5.0	141	€	Comida rápida, Europea, Saludable, Opciones veget...
1	2	Trattoria Mamma Franca	5.0	380	€€ - €€€	Italiana, Pizza, Mediterránea, Europea, Opciones v...
2	3	La Alacena de Francis	5.0	328	€€ - €€€	Mediterránea, Europea, Española, Rusa, Opciones ve...
3	4	La Barra de Zapata	5.0	516	€€ - €€€	Internacional, Mediterránea, Europea, Española, Op...
4	5	Café Tramezzino	5.0	139	€	Café, Europea, Saludable, Opciones vegetarianas, O...
5	6	Marisquería La Mayor	5.0	158	€€ - €€€	Marisco, Mediterránea, Española, Opciones vegetar...
6	7	Byoko	5.0	285	€€ - €€€	Francesa, Café, Mediterránea, Española, Saludable, ...
7	8	Kortxo	5.0	423	€€ - €€€	Española, Internacional, Opciones vegetarianas, O...
8	9	Da Saveria comida italiana casera	4.5	634	€	Italiana, Mediterránea, Europea, Opciones vegetar...
9	10	Spago's - Fresh Pasta	5.0	337	€	Italiana, Comida rápida, Mediterránea, Europea, Op...
10	11	Brutus	4.5	203	€€ - €€€	Americana, Latina, Pizza, Saludable, Opciones vege...
11	12	Mura Mura Osteria Cafe	4.5	709	€€ - €€€	Italiana, Fusión, Mediterránea, Europea, Tienda go...
12	13	Restaurante Cávea	5.0	123	€€ - €€€	Española, Fusión, Mediterránea, Europea, Opciones ...
13	14	La Luz de Candela	4.5	808	€€ - €€€	Francesa, Mediterránea, Europea, Española, Opcione...
14	15	La Récréation	4.5	272	€€ - €€€	Francesa, Mediterránea, Fusión, Pub restaurante, O...
15	16	La Tranca	4.5	826	€	Bar, Mediterránea, Española, Opciones vegetarianas
16	17	Risolì	4.5	345	€€ - €€€	Mediterránea, Española, Opciones vegetarianas, Op...
17	18	García Taberna	4.5	481	€€ - €€€	Internacional, Mediterránea, Europea, Española, Co...

2042	El Comedor	0.0	0	€€ - €€€	Internacional, Mediterránea, Española, Fusión, Sal...
2043	Icaro Gastrobar	0.0	0		
2044	Bar Sabor A Monte	0.0	0	€€ - €€€	Española
2045	Indian Internacional Restaurant El perchel	0.0	0		Italiana, De Oriente Medio
2046	Rincón Latino	0.0	0		
2047	Casa Félix	0.0	0	€	Española
2048	The Black Bull	0.0	0	€	Bar, Café, Pub
2049	Ikni	0.0	0		
2050	Molkamo bar	0.0	0		Bar, Pizza
2051	lindaraja gril	0.0	0	€€€€	Española
2052	Taperia/Bocateria Nam Nam Poti Poti	0.0	0		Mediterránea
2053	El Lorro	0.0	0		Española
2054	Hípico del limonar	0.0	0		Española
2055	Cafetería Andersen	0.0	0		Internacional, Española, Fusión
2056	Cafe Pasarela	0.0	0		

2057 rows x 6 columns

Código en:
<https://github.com/sararl/python-playground/tree/master/ExtractTripAdvisorInfo>

CONECTAR CON LA
WEB, EXTRAER
CÓDIGO Y NAVEGAR
POR LA WEB



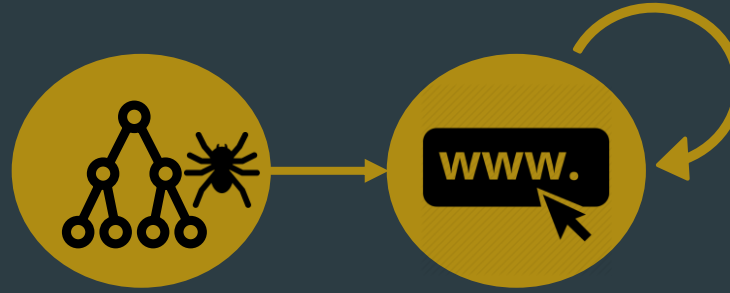
WEB CRAWLING

CARGAS LAS LIBRERÍAS E INICIALIZAS EL DRIVER

```
from selenium import webdriver  
# we are initializing "Firefox" by making an object of it  
driver = webdriver.Firefox()  
# we connect to the url  
driver.get("https://www.tripadvisor.es/Restaurants")
```



CONECTAR CON LA
WEB, EXTRAER
CÓDIGO Y NAVEGAR
POR LA WEB



WEB CRAWLING

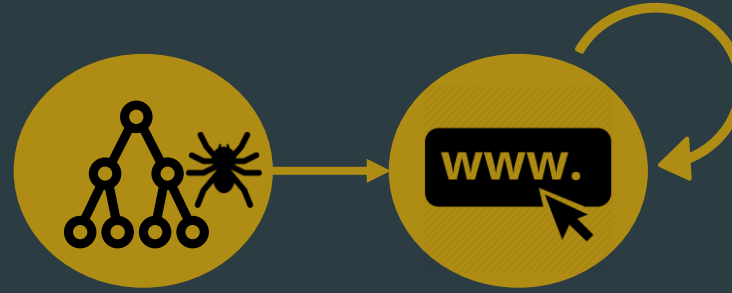
BUSCAMOS ELEMENTOS EN EL CÓDIGO HTML

Los objetos webdriver de Selenium tienen varios métodos que te permiten buscar cualquier tipo de elemento HTML, por su nombre, etiqueta id, class, xpath,....

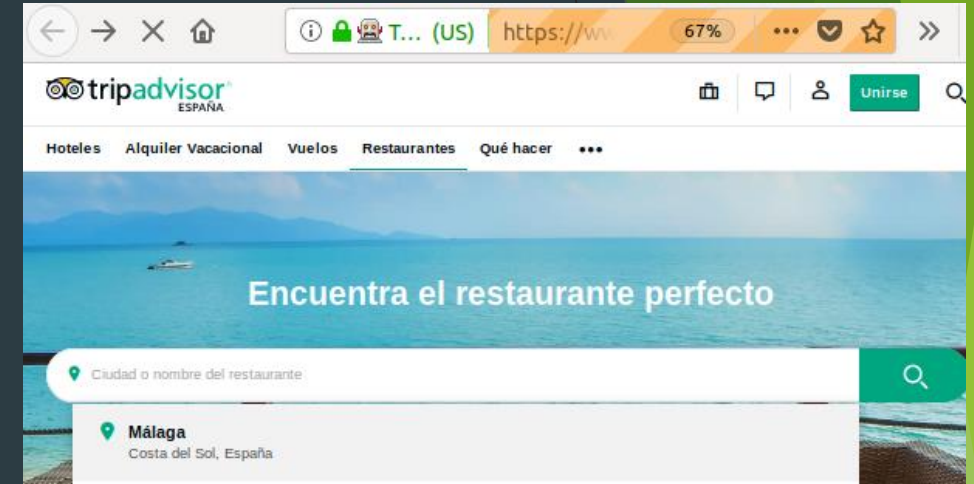
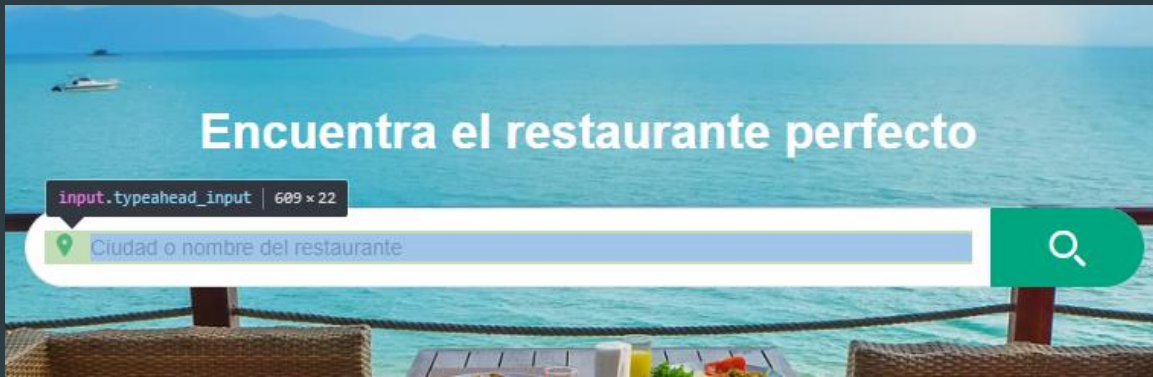
```
find_element_by_id  
find_element_by_name  
find_element_by_xpath  
find_element_by_link_text
```

```
find_element_by_partial_link_text  
find_element_by_tag_name  
find_element_by_class_name  
find_element_by_css_selector
```


CONECTAR CON LA
WEB, EXTRAER
CÓDIGO Y NAVEGAR
POR LA WEB



WEB CRAWLING



Seleccionas el elemento
Borras el campo
Escribes en el campo
Pulsas enter




```
# Fill box with city name
nameRestaurant = driver.find_element_by_class_name("typeahead_input")
nameRestaurant.clear()
nameRestaurant.send_keys("Málaga")
nameRestaurant.send_keys(Keys.RETURN)
```


PARSEAR CÓDIGO HTML

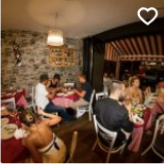


WEB CRAWLING

Reserva: 17/9/2018 20:00 2 clientes Encuentra un restaurante Borrar búsqueda




Luxalad
127 opiniones
1 de 1.972 Restaurantes en Málaga
€, Comida rápida, Europea, Saludable, Opciones vegetarianas, Opciones veganas, ()
"La belleza de lo simple" 11/09/2018
"Recomendable 100%" 05/09/2018
[Hacer pedido online](#)



Trattoria Mamma Franca
343 opiniones
2 de 1.972 Restaurantes en Málaga
€€ - €€€, Italiana, Pizza, Mediterránea, Europea, Opciones vegetarianas, Opcione...
"Buena comida y gran atencion" 15/09/2018
"El sabor de la auténtica Italia" 15/09/2018
[Ver otras horas](#)

¡Mesas disponibles para esta noche!

20:30
21:00
21:30



La Barra de Zapata
488 opiniones
3 de 1.972 Restaurantes en Málaga
€€ - €€€, Internacional, Mediterránea, Europea, Española, Opciones vegetarianas, C



```
find_element_by_id  
find_element_by_name  
find_element_by_xpath  
find_element_by_link_text  
find_element_by_partial_link_text  
find_element_by_tag_name  
find_element_by_class_name  
find_element_by_css_selector
```

o bien aplico lo que ya he aprendido de BeautifulSoup...



```
textourl = driver.page_source  
soup = BeautifulSoup(textourl, "lxml")
```

NAVEGAR POR LA WEB



```
<div class="deckTools btm">  
  <div class="unified pagination js_pagelinks">  
    <span class="nav previous disabled">  
      Anterior  
    </span>  
    <a data-page-number="2" data-offset="30" href="/  
      Restaurants-g187438-0a30-  
      Malaga Costa del Sol Province of Malaga Andalusia.html  
      #EATERY LIST CONTENTS" class="nav next rndBtn  
      ui_button primary taLnk" onclick="  
      ta.restaurant_filter.paginate(this.getAttribute('data-  
      offset')); ta.trackEventOnPage('STANDARD_PAGINATION',  
      'next', '2', 0); return false;  
    ">  
      Siguiete  
    </a> == $0
```

```
# click next page  
nextPage = driver.find_element_by_class_name('nav.next.rndBtn.ui_button.primary.taLnk')  
nextPage.click()
```

Repetimos la lógica empleada: nuestro driver va navegando por las distintas páginas y vamos extrayendo la información de los restaurantes de cada página

	position	restaurants	ratings	number_reviews	price_category	food_category
0	1	Luxalad	5.0	141	€	Comida rápida, Europea, Saludable, Opciones veget...
1	2	Trattoria Mamma Franca	5.0	380	€€ - €€€	Italiana, Pizza, Mediterránea, Europea, Opciones v...
2	3	La Alacena de Francis	5.0	328	€€ - €€€	Mediterránea, Europea, Española, Rusa, Opciones ve...
3	4	La Barra de Zapata	5.0	516	€€ - €€€	Internacional, Mediterránea, Europea, Española, Op...
4	5	Café Tramezzino	5.0	139	€	Café, Europea, Saludable, Opciones vegetarianas, O...
5	6	Marisquería La Mayor	5.0	158	€€ - €€€	Marisco, Mediterránea, Española, Opciones vegetar...
6	7	Byoko	5.0	285	€€ - €€€	Francesa, Café, Mediterránea, Española, Saludable, ...
7	8	Kortxo	5.0	423	€€ - €€€	Española, Internacional, Opciones vegetarianas, O...
8	9	Da Saveria comida italiana casera	4.5	634	€	Italiana, Mediterránea, Europea, Opciones vegetar...
9	10	Spago's - Fresh Pasta	5.0	337	€	Italiana, Comida rápida, Mediterránea, Europea, Op...
10	11	Brutus	4.5	203	€€ - €€€	Americana, Latina, Pizza, Saludable, Opciones vege...
11	12	Mura Mura Osteria Cafe	4.5	709	€€ - €€€	Italiana, Fusión, Mediterránea, Europea, Tienda go...
12	13	Restaurante Cávea	5.0	123	€€ - €€€	Española, Fusión, Mediterránea, Europea, Opciones ...
13	14	La Luz de Candela	4.5	808	€€ - €€€	Francesa, Mediterránea, Europea, Española, Opcione...
14	15	La Récréation	4.5	272	€€ - €€€	Francesa, Mediterránea, Fusión, Pub restaurante, O...
15	16	La Tranca	4.5	826	€	Bar, Mediterránea, Española, Opciones vegetarianas
16	17	Risolì	4.5	345	€€ - €€€	Mediterránea, Española, Opciones vegetarianas, Op...
17	18	García Taberna	4.5	481	€€ - €€€	Internacional, Mediterránea, Europea, Española, Co...

2042	El Comedor	0.0	0	€€ - €€€	Internacional, Mediterránea, Española, Fusión, Sal...
2043	Icaro Gastrobar	0.0	0		
2044	Bar Sabor A Monte	0.0	0	€€ - €€€	Española
2045	Indian Internacional Restaurant El perchel	0.0	0		Italiana, De Oriente Medio
2046	Rincón Latino	0.0	0		
2047	Casa Félix	0.0	0	€	Española
2048	The Black Bull	0.0	0	€	Bar, Café, Pub
2049	Ikni	0.0	0		
2050	Molkamo bar	0.0	0		Bar, Pizza
2051	lindaraja gril	0.0	0	€€€€	Española
2052	Taperia/Bocateria Nam Nam Poti Poti	0.0	0		Mediterránea
2053	El Lorro	0.0	0		Española
2054	Hípico del limonar	0.0	0		Española
2055	Cafetería Andersen	0.0	0		Internacional, Española, Fusión
2056	Cafe Pasarela	0.0	0		

2057 rows × 6 columns

Código en:
<https://github.com/sararl/python-playground/tree/master/ExtractTripAdvisorInfo>

MUCHAS GRACIAS



sara@sararl.com

[sararodriguezlopez](#)

<https://sararl.github.io/>