

Scraping javascript sin javascript



dir(self)

__name__: Xavier Petit

__mail__: nuxion@gmail.com

__bio__: Tecnico en computacion, sysadmin en una primer época, programador desde los últimos 5 años. Actualmente trabajando con datos y machine learning en Algorinfo (<https://www.algorinfo.com>).

__twitter__: @xpetitde

TOC

Caso 1

El comienzo de una búsqueda

Query params + a href links
+ beautiful soup

Caso 2

Siguiendo el precio del dólar

Inspector + json

Caso 3

No todo es json o html

robots.txt + sitemap.xml

Caso 4

Javascript lo mira por youtube

beautiful soup (de nuevo)
+ Regex (o no..)

Caso 1

El comienzo de una búsqueda

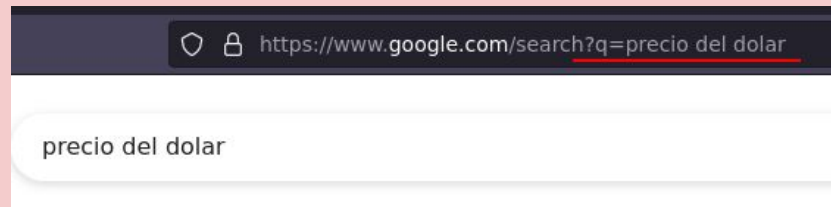
Query params + a href links + beautiful soup



Caso 1

El primer paso para un crawler es identificar los links de un sitio.

No sólo levantamos links sino que además verificamos que utilizando **query params en un URL podemos hacer búsquedas de forma programática.**



```
<style>.bmr/nv1{padding:20px}</style>  
▼ <form action="/search" method="GET" role="search">  
  ▼ <div jsmode="vWNDde" jsdata="MUIEvd;_;BgKlKk">  
    ▼ <div class="A8SBwf" jscontroller="W5mj0c" jsmode
```

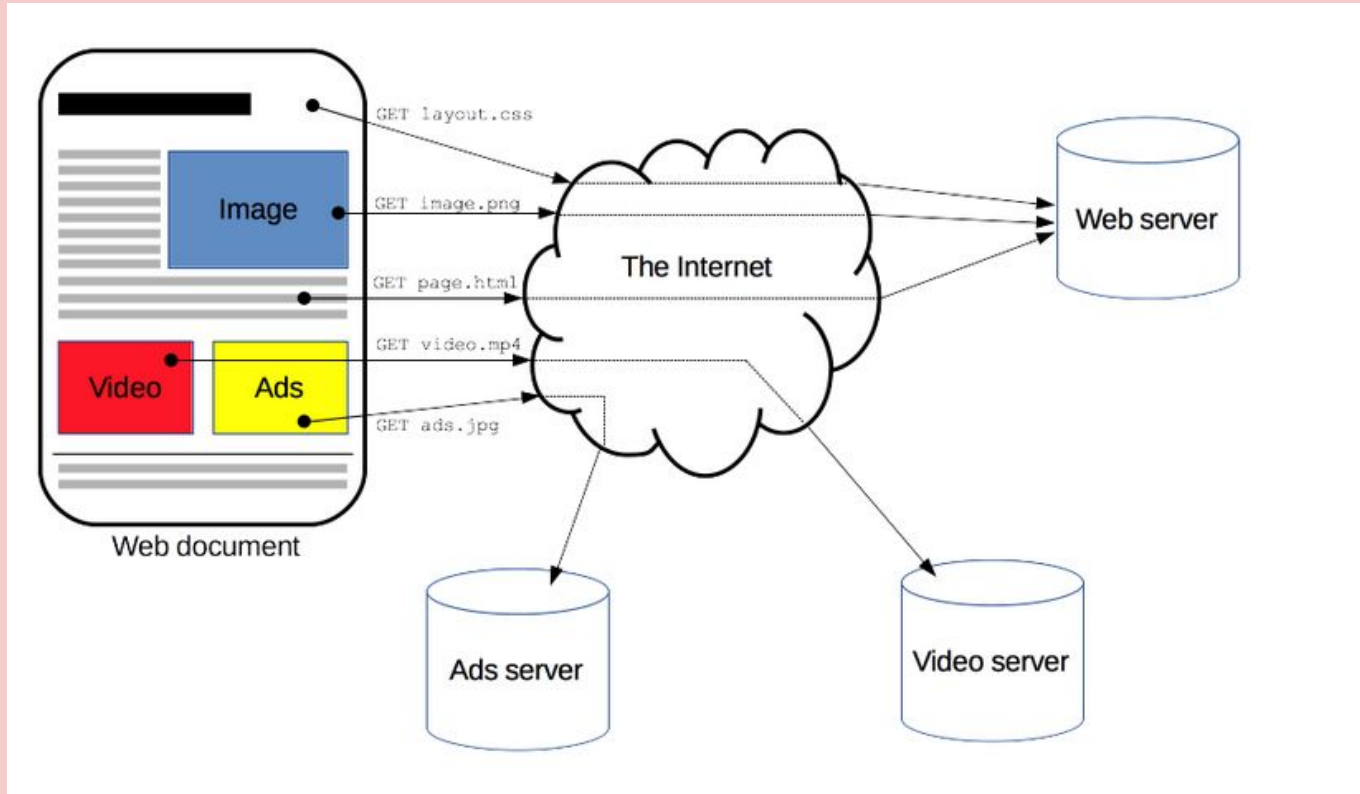
Caso 2

Siguiendo el precio del dólar

Inspector + JSON

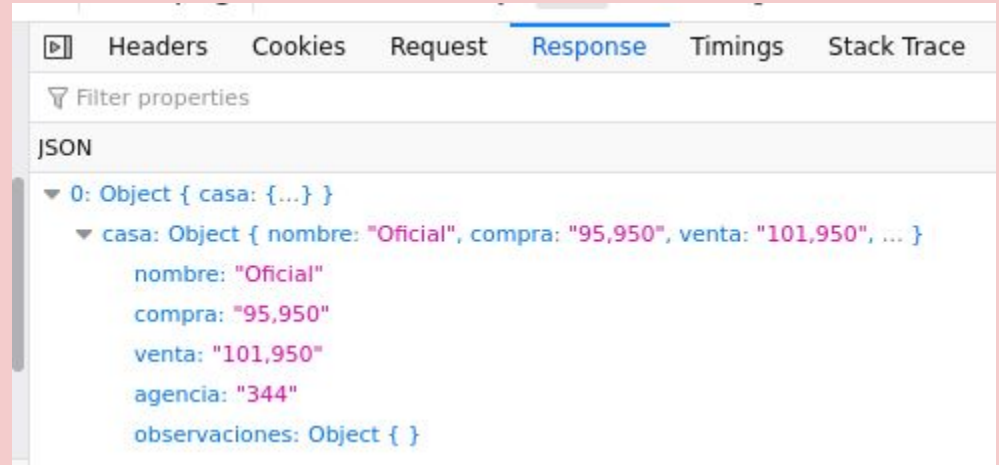


Caso 2



Caso 2

- Pestaña de network en el navegador
- XHR Requests
- Parsing JSON



Caso 3

No todo es JSON o HTML

`Robots.txt + sitemap.xml`



Caso 3

- En este caso el sitio es totalmente una SPA, si desactivo javascript del browser, no se ve nada.

- Alternativas:

- Que es el protocolo robots.txt?
- Para que se usan los sitemaps ?

```
# Disallow all crawlers access to certain pages.
User-agent: *

Disallow: /img/*
Disallow: /account/*
Disallow: /login/*
Disallow: /checkout/*
Disallow: /busca/*
Disallow: /quick-view/*
Disallow: /espiar/*
Noindex: /buscapagina/*

Sitemap: https://supermercado.carrefour.com.ar/sitemap.xml
```

```
-<urlset>
-<url>
  -<loc>
    https://www.carrefour.com.ar/palillos-ensobrados-iberia-80-u/p
  </loc>
  <lastmod>2021-07-23T06:04:16.245Z</lastmod>
</url>
-<url>
  -<loc>
    https://www.carrefour.com.ar/fuente-rectangular-pyrex-con-tapa-18-x-13-x-4-cm-1-u/p
  </loc>
  <lastmod>2021-07-23T06:04:16.245Z</lastmod>
</url>
-<url>
  -<loc>
    https://www.carrefour.com.ar/fuente-cuadrada-pyrex-con-tapa-21-cm/p
  </loc>
  <lastmod>2021-07-23T06:04:16.245Z</lastmod>
</url>
```

Caso 3

because this is the first one they have seen. Almost daily, we receive an email something like, "Wow, you looked at a lot of pages from my web site. How did you like it?" There are also some people who do not know about the robots exclusion protocol, and think their page should be protected from indexing by a statement like, "This page is copyrighted and should not be indexed", which needless to say is difficult for web crawlers to understand. Also,

Caso 4

JavaScript lo mira por youtube

beautiful soup (de nuevo) + Regex (o no..)



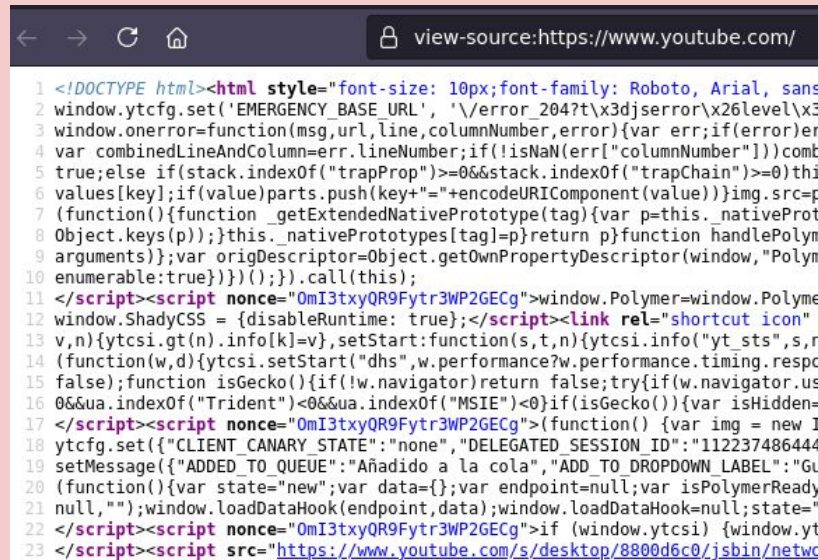
Caso 4

Youtube no es una SPA, pero sin embargo envía mucho código javascript directamente en el html de la página

JSON significa JavaScript Object Notation

Por ende la misma estructura para definir datos en javascript es la que muchas veces

Un backend manda al front.



```
1 <!DOCTYPE html><html style="font-size: 10px;font-family: Roboto, Arial, sans
2 window.ytcfg.set('EMERGENCY_BASE_URL', '\error_2047t\x3djserror\x26level\x3
3 window.onerror=function(msg,url,line,columnNumber,error){var err;if(error)er
4 var combinedLineAndColumn=err.lineNumber;if(!isNaN(err["columnNumber"]))comb
5 true;else if(stack.indexOf("trapProp")>=0&&stack.indexOf("trapChain")>=0)thi
6 values[key];if(value)parts.push(key+"="+encodeURIComponent(value))}img.src=p
7 (function(){function _getExtendedNativePrototype(tag){var p=this._nativeProt
8 Object.keys(p));this._nativePrototypes[tag]=p}return p}function handlePolym
9 arguments));var origDescriptor=Object.getOwnPropertyDescriptor(window,"Polym
10 enumerable:true)}}})();).call(this);
11 </script><script nonce="0mI3txyQR9Fytr3WP2GECg">window.Polymer=window.Polyme
12 window.ShadyCSS = {disableRuntime: true};</script><link rel="shortcut icon"
13 v,n){ytcsi.gt(n).info[k]=v},setStart:function(s,t,n){ytcsi.info("yt_sts",s,r
14 (function(w,d){ytcsi.setStart("dhs",w.performance.timing.respo
15 false);function isGecko(){if(!w.navigator)return false;try{if(w.navigator.us
16 0&&ua.indexOf("Trident")<0&&ua.indexOf("MSIE")<0){if(isGecko()){var isHidden=
17 </script><script nonce="0mI3txyQR9Fytr3WP2GECg">(function() {var img = new I
18 ytcfg.set({"CLIENT_CANARY_STATE":"none","DELEGATED_SESSION_ID":"112237486444
19 setMessage({"ADDED_TO_QUEUE":"Añadido a la cola","ADD_TO_DROPDOWN_LABEL":"Gl
20 (function(){var state="new";var data={};var endpoint=null;var isPolymerReady
21 null,"");window.loadDataHook(endpoint,data);window.loadDataHook=null;state="
22 </script><script nonce="0mI3txyQR9Fytr3WP2GECg">if (window.ytcsi) {window.yt
23 </script><script src="https://www.youtube.com/s/desktop/8800d6c0/jsbin/netw
```

Caso 4

```
re.findall(  
r"{.+[:],].+}|\\[.+[,:].+\]  
",  
s)
```



Conclusiones

Hacer scraping, más allá de una necesidad concreta de conseguir información, puede ser útil como vía de aprendizaje.

Las técnicas de scraping desafían los conocimientos de web development (front y back), del protocolo HTTP (get, post, queryparams, etc), standards de la web: sitemap, robots.txt, rss, xml.

Y por que no, también otros lenguajes como javascript.

Referencias

- <https://developer.mozilla.org/en-US/docs/Web/HTTP/Methods>
- <https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview>
- <https://github.com/mraible/history-of-web-frameworks-timeline>
- <https://sitemaps.org/>
- <https://developers.google.com/search/docs/advanced/robots/intro>
- <http://infolab.stanford.edu/~backrub/google.html>
- <https://www.facebook.com/notes/10158791368532200/>
- <https://web.dev/>