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 0

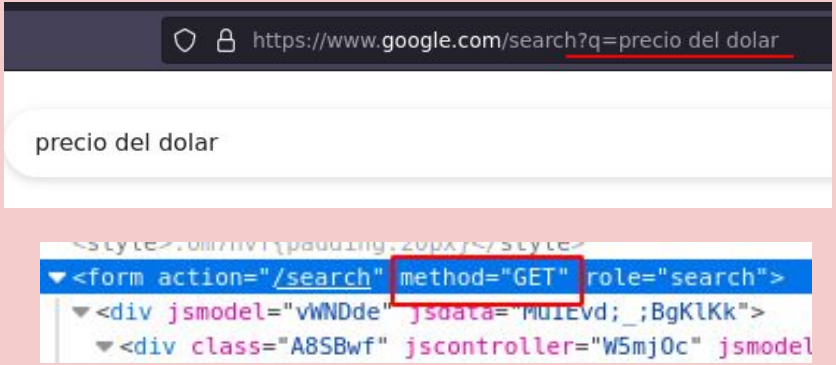
El comienzo de una búsqueda

Query params + ah ref links + beautiful soup

Caso 0

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.**



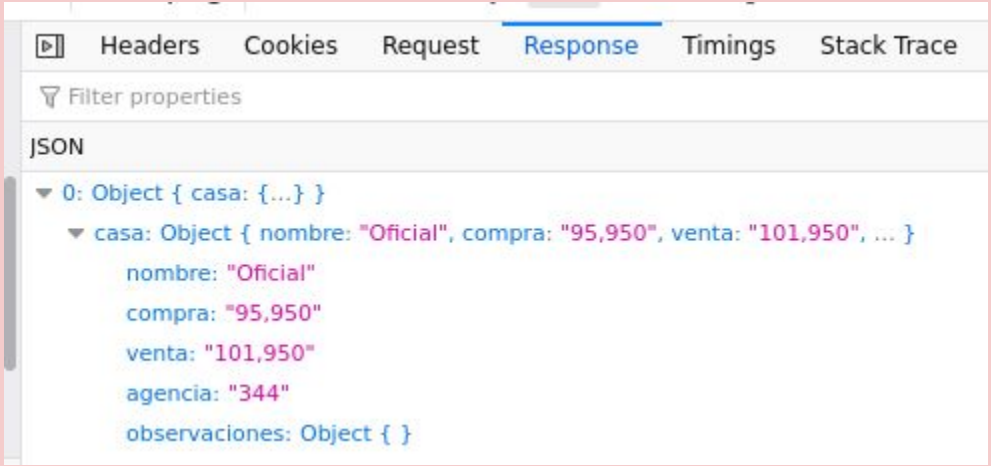
Caso 1

Siguiendo el precio del dólar

Inspector + JSON

Caso 1

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



Caso 2

No todo es JSON o HTML

Robots.txt + sitemap.xml

Caso 2

- 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

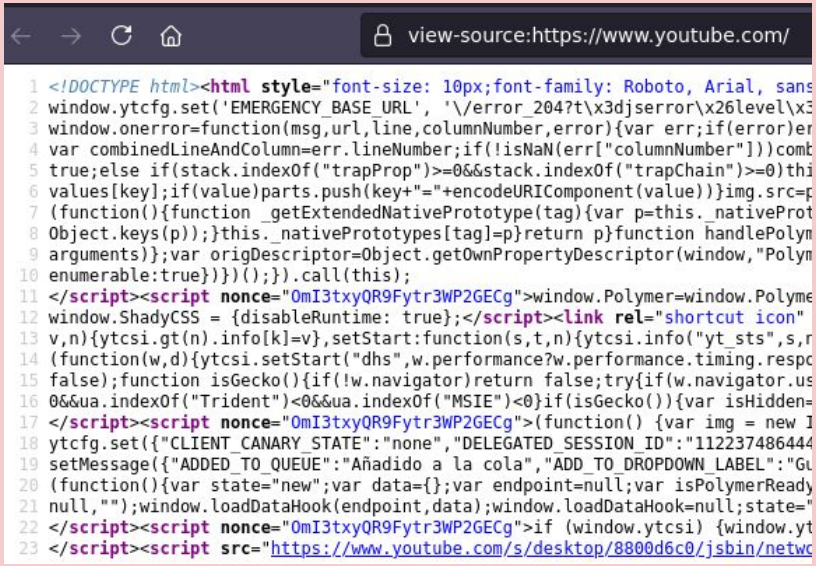
JavaScript lo mira por youtube

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

Caso 3

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.



Caso 3

```
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.