

## ESQUEMA MENTAL PARA SCRAPY y BEAUTIFULSOUP

### 0. ESTRATEGIA GENERAL

- a. En este curso, elegir HTML estático. Botón derecho ver código fuente de la página y ver que ya aparece la información que busco, si no aparece entonces hay JS y haría falta herramientas adicionales.
  - b. Usar el Inspector y localizar el contenedor repetible
  - c. Identifica subelemento internos
  - d. Probar selectores, no siempre va a la primera
  - e. Iterar sobre el contenedor padre para extraer subelementos
  - f. Limpiar
  - g. Si es necesario Pagar
- 

### 1. Inspeccionar la página y ver que **datos quieres extraer**:

- a. Nombre
- b. Precio
- c. URL del producto
- d. Stock
- e. Descripción...

Empieza con 1 o 2, y luego lo complicas si quieres.

### 2. Identifica el **contenedor REPETIBLE**

- a. Puede ser un div, article, li...
- b. Buscamos una clase o atributo que se repita:

`<div class="producto-name">....</div>`

`<article class="product"> ...</article>`

`<li class="item">....</li>`

Y en el spider iteramos:

`for producto in response.css("div.product-name")`

### 3. Encontrar **subelementos dentro del contenedor (son ejemplos)**

DATO	POSIBLE ELEMENTO HTML	SELECTOR CSS
Titulo	<h2> , <a> , <span>	h2 a::text
Precio	<span>	span.price::text
URL	<a href="...">	a::attr(ref)
Imagen	<img>	img::attr(src)
Descripcion	<div> , <p> , <span>	div.description::text

- Usa **::text** para el **contenido visible**
- Usa **::attr(atributo)** para **enlaces, imágenes o atributos**
- **.get()** devuelve primer valor (string, número), **.getall()** devuelve todos (hace falta .join)

### 4. Iteración y **extracción de datos**

- Itera sobre el contenedor padre**
- Extrae **subelementos de cada contenedor**
- Maneja datos vacíos y evitar errores (generalmente con un if ... else None)

### 5. **Paginación**

*# Paginación*

```
siguiente = response.css("ul.pagination a.next::attr(href)").get()
```

```
if siguiente:
```

```
yield response.follow(siguiente, self.parse)
```

### 6. Limpieza de Datos

- Con **.strip()** o **.join()**
- Controlar valores vacíos