JÚLIA
RODRÍGUEZ

Proyecto Web Scraping

10 de abril, 2021

Índice

01 Objetivos de negocio

Obtener datos para analizar el posicionamiento orgánico de los productos de Revlon Professional en Amazon España

Objetivo 1: analizar el posicionamiento de los productos cuando alguien busca el nombre de la marca.

Objetivo 2: analizar el posicionamiento de los productos de Revlon Professional y de los principales competidores cuando alguien busca productos de peluquería profesional sin especificar marca.

REVLON PROFESSIONAL™

WEB SCRAPING PROJECT

02

Datos que necesitamos

- Nombre del producto
- Marca
- Precio final
- Precio a granel
- Precio sin descuento
- Rating
- Número de reviews
- Fecha de entrega

REVLON PROFESSIONAL™

1-48 de 782 resultados para **"revlon professional"**

Ordenar por: Destacados

Con derecho a envío gratis

- [ ] Envío gratis

Envío GRATIS en pedidos superiores a 19€ en Libros o a 29€ en las demás categorías de productos

Día de entrega

- [ ] Recíbelo mañana

Productos frescos

- [ ] Amazon Fresh
- [ ] Supermercado DIA

Departamento

‹ Cualquier departamento
‹ Belleza
‹ Cuidado del cabello
Productos para el cuidado del cabello
Mascarillas de pelo
Champús
Acondicionadores de pelo
Máscaras de tinte de pelo

Valoración media de los clientes

★★★★☆ o más
★★★☆☆ o más
★★☆☆☆ o más
★☆☆☆☆ o más

Marca

- [ ] REVLON PROFESSIONAL
- [ ] REVLON
- [ ] EQUAVE
- [ ] L'Oréal Professionnel
- [ ] TIGI
- [ ] TIGI Bed Head
- [ ] Ckeyin
- [ ] BED HEAD by TIGI

REVLON

Descubre los productos de Revlon

Descúbrelo REVLON ›

Revlon Wonder Woman Super Lustrous Pintalabios (Amazon)
★★★★½ 308

Revlon Wonder Woman Gift Pack
★★★★½ 8

Revlon Wonder Woman Ultra HD Vinyl Mouse Labial (So Shady)
★★★★☆ 72

Patrocinado

Patrocinado

UniqOne Revlon Professional Classico Tratamiento en Spray para Cabello 150 ml y Super10R Mascarilla 300 ml - Pack 300 ml
★★★★½ 71

Revlon Professional UniqOne Champú y Acondicionador 1000 ml
★★★★½ 926
16,99€ (1,70 €/100 ml)
Envío GRATIS

UniqOne Revlon Professional - Tratamiento para el cabello, Coco, 150 ml
★★★★½ 8.810
6,95€ (4,63 €/100 ml) ~~11,66€~~
6,60 € con el descuento de Compra

Revlon Professional ProYou Textura de Peinado para Cabello Fino 350 ml
★★★★☆ 71
12,99€ (12,99 €/100 g)
12,34 € con el descuento de Compra recurrente

1-48 de más de 5.000 resultados para **"profesional"**

Ordenar por: Destacados

Con derecho a envío gratis

☐ Envío gratis

Envío GRATIS en pedidos superiores a 19€ en Libros o a 29€ en las demás categorías de productos

Día de entrega

☐ Recíbelo mañana

Productos frescos

☐ Amazon Fresh

☐ Supermercado DIA

Departamento

‹ Cualquier departamento

Productos para el cuidado del cabello

Valoración media de los clientes

★★★★☆ o más

★★★☆☆ o más

★★☆☆☆ o más

★☆☆☆☆ o más

Marca

☐ RATEL

☐ Philips

Precio

0 - 5 EUR

5 - 10 EUR

10 - 20 EUR

20 - 50 EUR

Más de 50 EUR

EUR Mín. | EUR Máx | Ir

Ofertas

☐ Ofertas

Timoteí

Descubre nuestra gama de productos

Descúbrelo Timotei ›

Champús

Mascarillas

Bio

Patrocinado

Patrocinado

Hatteker Barbero Electrico Cortapelos Profesional Hombre Cortador Pelo Cortadora de Pelo Recargable Máquina de Afeitar...

★★★★☆ 15.163

Patrocinado

abril et nature - Pack Regalo Tratamiento Profesional NATURE-PLEX Cabello dañado - Incluye Mascarilla Pelo, Aceite Pelo y...

★★★★☆ 65

Patrocinado

abril et nature | Nature Frizz Tratamiento Profesional para Cabello Encrespado (Protector Cabello - 100ml)

17,50€ (17,50 €/100 ml)

Patrocinado

HAIRMED - N8 Acondicionador Profesional para el Cabello - Acondicionador Desenredante con Colágeno - 200 ml

★★★★☆ 17

03 Métodos y resultados

1. BeautifulSoup
2. Scrapy
3. Selenium (descartado)

BeautifulSoup

Para el objetivo 1, necesitamos obtener 796 artículos distribuidos en 17 páginas

1. Obtenemos los XPaths de los productos y de todos los campos que nos interesan.
2. Generamos las URLs de todas las páginas con una f string y un for loop.
3. Parseamos todas las páginas con un for loop convertiendo cada producto en un diccionario.
4. Lo convertimos todo a DataFrame y lo exportamos como csv.

Resultado:

```
# Checking there are no errors
df2.head()
```

	prod_name	price	old_price	price_ml	stars	reviews	delivery_date
0	Revlon Professional UniqOne Champú y Acondicio...	NaN	NaN	NaN	4.5	3003.0	None
1	UniqOne Revlon Professional - Tratamiento para...	6.95	11.66	4.63	4.5	8705.0	6 de abril
2	UniqOne Revlon Professional Classico Tratamien...	14.95	15.60	NaN	4.5	70.0	None
3	Revlon Professional ProYou Textura de Peinado ...	12.99	NaN	NaN	4.0	71.0	6 de abril
4	REVLON PROFESSIONAL Nutri Color Filters #400 T...	11.71	15.00	NaN	4.6	80.0	6 de abril

```
# Checking there are as many rows as products in the amazon section
df2.shape
```

```
(796, 7)
```

Scrapy

Para el objetivo 2, necesitamos obtener más de 5000 artículos distribuidos en 117 páginas

1. Creamos un proyecto de scrapy y una spider.
2. Con Scrapy shell comprobamos que los XPaths de los campos que nos interesan funcionan correctamente.
3. Definimos la función parse de la spider para que haga yield de un diccionario para cada producto, con todos los campos necesarios. Ponemos None por defecto si el campo está vacío para evitar errores. Dentro de la misma función, incluimos la request a la siguiente página y un callback a la función parse para que empiece de nuevo.
4. Agregamos un User Agent en Settings.
5. Hacemos crawl y lo guardamos todo en un fichero json.
6. Comprobamos que está correcto con pandas y lo convertimos a csv.

Resultado:

```
products.shape
```

```
(5601, 7)
```

```
products.head()
```

	prod_name	price_whole	price_bulk	price_old	stars	reviews	delivery_date
0	Válquer Profesional Pack Tratamiento Capilar A...	24,14	(24,14 €/unidad)	None	3,2 de 5 estrellas	27.00	mañana, 7 de abril
1	Válquer Profesional Ice Hair Mask. Mascarilla ...	13,99	(4,66 €/100 ml)	None	4,4 de 5 estrellas	24.00	mañana, 7 de abril
2	KOKEN - Champú Neutro Profesional 5L- Fórmula ...	14,90	(0,30 €/100 ml)	None	None	NaN	None
3	Pack Champu 0% Sulfatos 1000ml + Mascarilla 0%...	22,48	(22,48 €/unidad)	27,00 €	4,4 de 5 estrellas	10.00	None
4	L'Oréal Professionnel Champú Absolut Repair, 5...	11,86	(2,37 €/100 ml)	None	4,4 de 5 estrellas	4.79	domingo, 11 de abril

FOLDERS

- amazon
  - __pycache__
  - spiders
    - __pycache__
    - __init__.py
    - productos
  - __init__.py
  - items.py
  - middlewares
  - pipelines.py
  - settings.py

productos.py

```python
import scrapy


class ProductosSpider(scrapy.Spider):
    name = 'productos'
    allowed_domains = ['amazon.es']
    # Url of the HairCare Products section of Amazon Spain, with "professional" typed into the search box
    start_urls = ['https://www.amazon.es/s?k=profesional&rh=n%3A4347699031&__mk_es_ES=%C3%85M%C3%85%C5%BD%C3%95%C3%

    def parse(self, response):
        # Obtaining a list of products by selecting them by class with xpath
        products = response.xpath("//div[@class='sg-col-4-of-12 s-result-item s-asin sg-col-4-of-16 sg-col sg-col-4
        # Iterating over the list to create a dictionary for every product
        for product in products:
            yield {
                'prod_name':product.xpath(".//span[@class='a-size-base-plus a-color-base a-text-normal']/text()").g
                'price_whole':product.xpath(".//span[@class='a-price-whole']/text()").get(default=None),
                'price_bulk':product.xpath(".//span[@class='a-size-base a-color-secondary']/text()").get(default=No
                'price_old':product.xpath(".//span[@class='a-price a-text-price']/span/text()").get(default=None),
                'stars':product.xpath(".//span[@class='a-icon-alt']/text()").get(default=None),
                'reviews':product.xpath(".//span[@class='a-size-base']/text()").get(default=None),
                'delivery_date':product.xpath(".//span[@class='a-text-bold']/text()").get(default=None)
            }

        # Changing to the next page when there are no more products to scrape
        next_page = response.xpath("//li[@class='a-last']/a/@href").get()
        if next_page is not None:
            next_page = response.urljoin(next_page)
            yield scrapy.Request(next_page, callback=self.parse)
```

04 Obstáculos

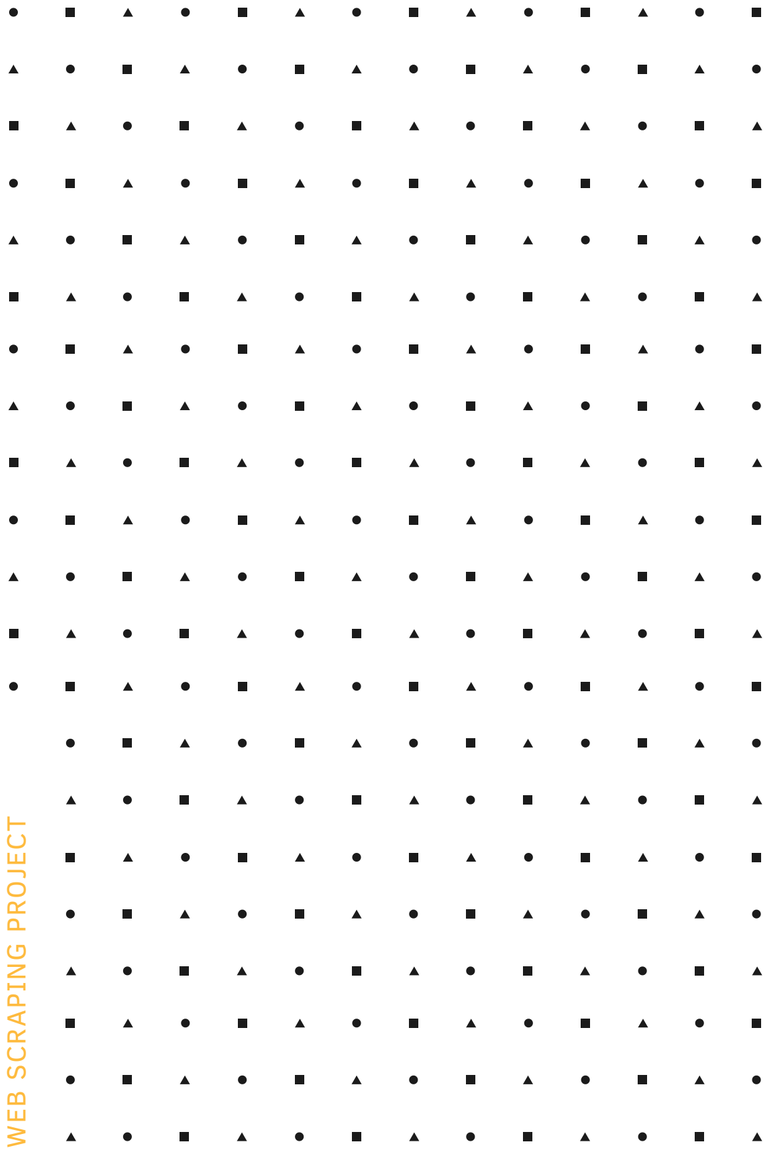
- Complejidad del código HTML de Amazon
- No existe el campo "Marca"
- En algunos productos hay campos vacíos
- Conocimiento superficial de XPath
- User Agent en Scrapy -> Settings
- Identation error -> Convert to spaces

05 Aprendizajes

- HTML y XPath
- BeautifulSoup, Scrapy y Selenium
- Sublime Text, resolver identation errors

06 Siguientes pasos

- Limpiar y analizar los datos.
- Hacer una spider más compleja que entre en la página de cada producto.
- Con más conocimiento de XPath, encontrar la marca dentro de la página del producto a partir del texto "Marca" o "Fabricante", que no siempre se encuantra dentro del mismo elemento HTML.



WEB SCRAPING PROJECT

Gracias!