TEC | Tecnológico de Costa Rica

# Introducción al Desarrollo de Páginas Web

Tarea Investigación 2

**Estudiante:**
Marco Soto Morera
2014046514

**Profesor:**
Efrén Jiménez Delgado

Grupo 51

I Semestre - 2022

## Resume:

Web scraping es una forma de minería de datos no estructurada, que permite extraer información de páginas web, escanear su código HTML y generar patrones de extracción de datos. El documento muestra las técnicas que existen, situaciones legales de ciertas páginas, herramientas existentes y con que lenguajes de programación se puede utilizar.

Se tomó el parseo de la página de Amazon, específicamente en la sección de tecnología y computadoras. Es desde aquí, que se realiza el scraping.

# Tabla de contenidos

## Introducción:

Realizar un web scraping a cualquier página es posible. Hay que tener algunas consideraciones y más que todo van con respecto a la página que se le quiere hacer el scraping.

En la sección de tecnología de la página de Amazon se puede realizar el procedimiento sin ningún problema, utilizando las herramientas de Python y la librería scrapy.

## Desarrollo:

En la página de Amazon se encuentra una serie de listas de artículos en donde las personas pueden comprar el producto que deseen. Para realizar la tarea de investigación se escogió las computadoras que se encuentran en la categoría de tecnología.

RESULTADOS

Computadora HP Stream pantalla 14" HD (1366 x 768), procesador Intel Celeron N4000 de doble núcleo, 4 GB de RAM, 32 GB eMMC, HDMI, WiFi, cámara web, Bluetooth, Win10 S, rosa...

★★★★½ 2,812

US$207$^{00}$

Con envíos a Costa Rica

Más opciones de compra
US$206.00 (18 ofertas nuevas)

HP Elite ordenador de sobremesa Intel Core i5 de 3,1 GHz, 8 gb Ram, 1 TB de disco duro, DVDRW, monitor LCD de 19 pulgadas, teclado, ratón, WiFi inalámbrico, Windows 10 (renovado)

★★★★☆ 4,775

US$206$^{89}$ ~~US$223.97~~

Con envíos a Costa Rica

Más opciones de compra
US$203.97 (9 ofertas nuevas)

Chromebook HP MediaTek MT8183 - MT8183 - 4 GB de RAM - 32 GB de almacenamiento eMMC - Pantalla HD de 11.6 pulgadas - con sistema operativo Chrome - (11a-na0010nr, modelo 2020)

★★★★½ 2,428


Acer Aspire 5 A515-56-36UT - Laptop delgada con procesador Intel Core i3-1115G4 de 11.ª generación, pantalla Full HD de 15.6 pulgadas, memoria DDR4 de 4 GB, unidad de estado sólido SSD NVMe de 128 GB, WiFi 6, Amazon Alexa, Windows 10 Home (modo S)

Visita la tienda de Acer

4.5 ★ 1,015 calificaciones | 122 preguntas respondidas

Amazon's Choice en Computadoras Laptops Tradicion...

Precio recomendado: ~~US$389.99~~

Precio: US$360.47

Ahorras: US$29.52 (8%)

US$ 88.34 de envío y depósito de derechos de importación a Costa Rica Detalles

Disponible a un precio menor de otros vendedores que podrían no ofrecer envío Prime gratis.

CPU: **i3-1115G4**

| | |
|---|---|
| **Series** | A515-56-36UT |
| **Marca** | Acer |
| **Usos específicos del producto** | Multimedia, Personal, Empresa |
| **Tamaño de pantalla** | 15.6 Pulgadas |
| **Sistema operativo** | Windows 11 S |
| **Entrada de interfaz humana** | Teclado |
| **Fabricante de CPU** | Intel |
| **Descripción de la tarjeta** | Integrated |
| **Características especiales** | Gesto multitáctil |
| **Color** | Plateado |

¿Qué contiene la caja?

- Portátil
- Adaptador de CA
- Cable de alimentación

Descripción del producto

Todo está en la innovación. Los productos Acer están diseñados para satisfacer tus necesidades y accesibilidad con potentes funciones que se adaptan a tu estilo de vida. El Aspire 5 incluye mucha potencia en el diseño portátil para adaptarse a las necesidades multi-tarea tuyas y de tu familia. El potente procesador Intel Core i3 de 11.ª generación es ideal para juegos ligeros, productividad y tareas de trabajo. Disfruta de la edición de fotos y video en la pantalla Full HD de 15.6 pulgadas con gráficos Intel UHD. Puedes convertir esta Acer Aspire 5 en una sala de reuniones virtual con una cámara web de alta definición y micrófonos dobles integrados mediante la tecnología de voz purificada de Acer para llevar tu conversación con claridad. Ya sea un chat de video, entretenimiento en transmisión o trabajo desde el hogar, permanecerás conectado a tu red con Wi-Fi 6 de doble banda que funciona de forma inteligente con tu router para aumentar la eficiencia de la red. Un teclado ergonómico levantado hace que trabajar en esta laptop Acer Aspire 5 sea más cómodo. (NX.AASAA.001).

**Título:** Atributo que se utiliza para conocer el atributo del nombre del producto

XPATH: class="a-size-large a-spacing-none" h1 span

**Precio:** Atributo que permite conocer el precio del articulo

XPATH: class="a-section a-spacing-micro" div span

**Serie:** Atributo que permite conocer la serie de la computadora

XPATH: class="a-normal a-spacing-micro" table tr span(1)

**Marca:** Atributo que permite conocer la marca de la computadora

XPATH: class="a-normal a-spacing-micro” table tr span (2)

**Usos específicos del producto:** Atributo que permite conocer el tipo de uso del artículo

XPATH: class="a-normal a-spacing-micro" table tr span(3)

**Tamaño de pantalla:** Atributo que permite conocer el tamaño de la pantalla de la computadora

XPATH: class="a-normal a-spacing-micro" table tr span(4)

**Sistema operativo:** Atributo que permite conocer el tipo de sistema operativo con el que viene la computadora

XPATH: class="a-normal a-spacing-micro" table tr span(5)

**Entrada de interfaz:** Atributo que permite conocer cuál es el medio de entrada para utilizar el producto
XPATH: class="a-normal a-spacing-micro" table tr span(6)

**Fabricante de CPU:** Atributo que permite conocer quien fue el fabricante de CPU
XPATH: class="a-normal a-spacing-micro" table tr span(7)

**Descripción de la tarjeta:** Atributo que permite conocer si la tarjeta está integrada o no
XPATH: class="a-normal a-spacing-micro" table tr span(8)

**Características especiales:** Atributo que permite conocer las características especiales del producto
XPATH: class="a-normal a-spacing-micro" table tr span(9)

**Color:** Atributo que permite conocer el color de la computadora
XPATH: class="a-normal a-spacing-micro" table tr span(10)

**Contenido de la caja:** Atributo que permite conocer lo que viene dentro de la caja además del producto
XPATH: class="a-section a-spacing-medium a-spacing-top-small" div span

**Descripción:** Atributo que permite conocer detalles más específicos del producto
XPATH: id="productDescription" div p

**Imagen:** Atributo que permite conocer la imagen asociada al producto
XPATH: class="imgTagWrapper"] div img

## Conclusiones y recomendaciones:

Utilizar la librería Scrapy, fue la mejor opción porque favorece el cumplimiento de las normas. Incluso es importante recalcar que la librería es un marco colaborativo y de código abierto para extraer los datos que necesita de los sitios web, de una manera rápida, simple, pero extensible.

Como recomendación es importante indicar que no a todas las páginas web se les puede hacer Web scrapping. Una de ella es Facebook, esto debido a la cantidad de robots antiparseo que existen en el sitio.

## Bibliografía y repositorio:

https://www.python.org/

https://github.com/marsep27/Tarea-de-Investigaci-n-2