

ACÀMICA

TEMA DEL DÍA

Scraping

El arte de conseguir los datos de la web.



Agenda

Daily

Explicación: Scraping.

Break

Proyecto

Cierre



Daily



Daily



Sincronizando...

Bitácora



¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

Challenge



¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

Repaso de la bitácora HTML



¿Qué es HTML?

HTML es un lenguaje de marcado que se utiliza para el desarrollo de páginas de Internet.

Los navegadores son los encargados de interpretar el código HTML y mostrarlo en la forma que lo hayamos desarrollado.



HTML

¿Por qué utilizar HTML?

Porque es el lenguaje que interpretan los navegadores cómo Edge, Chrome, Firefox, Safari y algunos otros más.

Estos navegadores se rigen con los estándares de la W3C, lo cual nos permite realizar un desarrollo y estar seguros podrá ser interpretado en cualquier navegador.



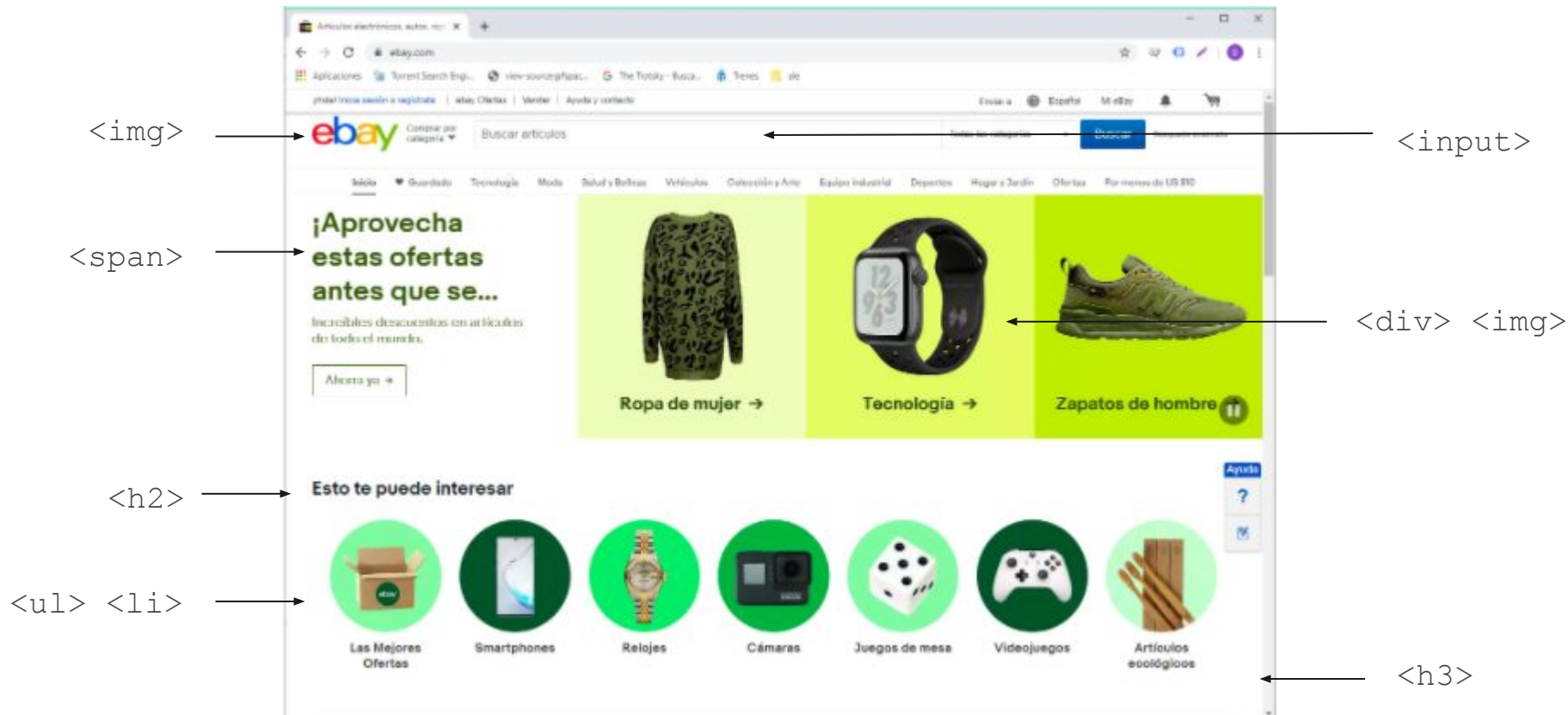
HTML

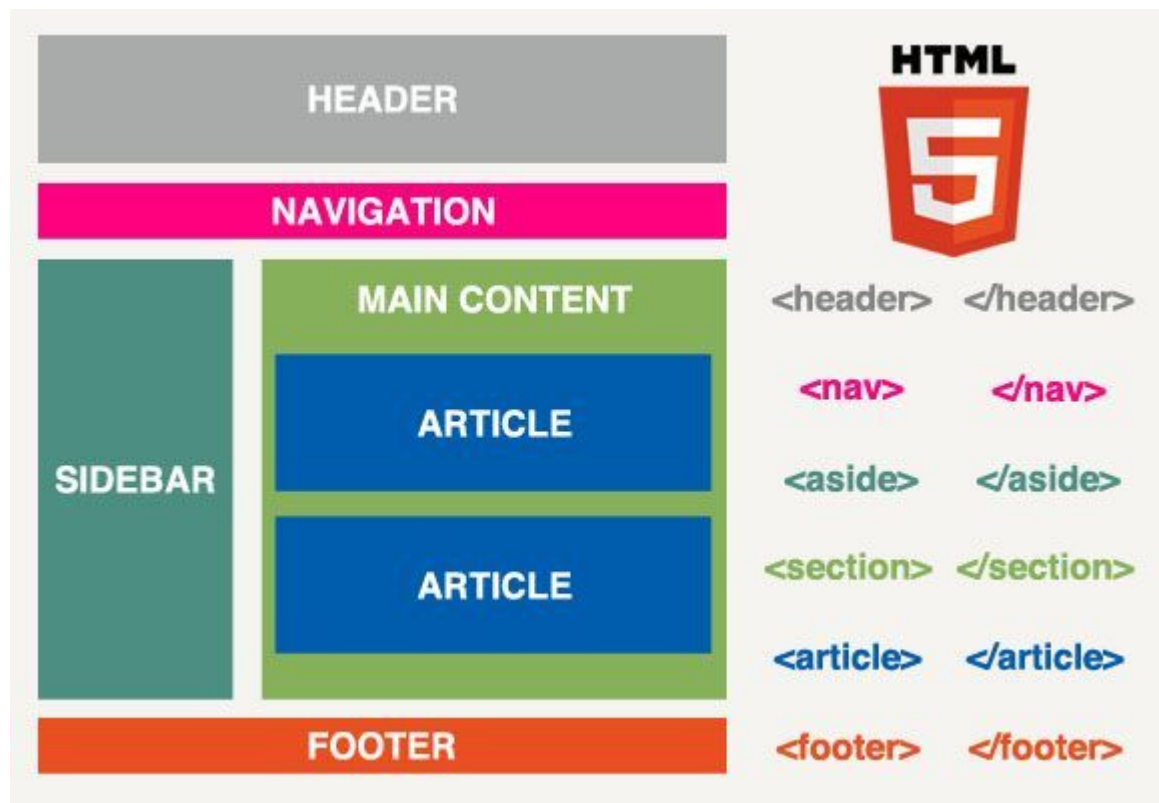
¿Donde encontramos HTML?

HTML lo encontramos en cada una de las webs que visitamos asiduamente.

A través de tags damos estructura a nuestro contenidos.







[Fuente](#)

Scraping



El arte de sacar datos de la web.

El arte de sacar datos de la web.

Para esto, usamos las librerías:

- Selenium
- BeautifulSoup

El arte de sacar datos de la web.

Para esto, usamos las librerías:

- Selenium para navegar por la web y obtener el HTML
- BeautifulSoup

El arte de sacar datos de la web.

Para esto, usamos las librerías:

- Selenium para navegar por la web y obtener el HTML
- BeautifulSoup para analizar ese HTML

Hands-on training



Hands-on training

DS_Bitácora_44_Scraping.ipynb

*¡Tienen **15 minutos** para consultar o probar algo!*



A close-up photograph of a white ceramic cup filled with a latte. The coffee has a thick layer of white foam with intricate brown latte art designs, including a central heart shape and radiating lines. The cup sits on a matching white saucer. In the background, a white napkin with a silver fork is visible, all set against a dark, textured surface.

¡BREAK!

Ph. Credit: Drew Coffmann



Workshop Proyecto 03



¡Continúa trabajando en tu proyecto!

Júntate con compañeros/as que estén trabajando en el mismo proyecto (máximo 3 personas).

1. Pongan un objetivo *alcanzable* para lo que queda del encuentro. Debe ser el mismo para todos/as. Por ejemplo, hacer una primera versión del Análisis Exploratorio de Datos, explorar una pregunta, implementar un modelo sencillo, etc. ¿Cuánto tiempo creen que les llevará?
2. 15 minutos antes del final del encuentro, hagan una puesta en común. ¿Llegaron a donde esperaban llegar? ¿Qué se les dificultó?

Presta atención a lo siguiente: ¿Puedes transmitir información de forma clara y concisa a una persona que está trabajando en tu mismo proyecto?

¡No te olvides de subir tu progreso al repositorio!

Recursos



Acceso a datos

- El capítulo 3 del libro de Steven Skiena, *The Data Science Design Manual*, tiene es sumamente recomendable para aprender más sobre acceso y manejo de datos.
- En [este video de youtube](#) encontrarás un tutorial de un proyecto simple de Scraping. Te recomendamos que lo mires antes de ver el notebook. ¡Dura 12 minutos y está en español!
- Como siempre, la documentación de [BeautifulSoup](#) y de [Selenium Web Driver](#) son visitas obligadas si quieres trabajar con estas librerías.



Para la próxima

- Avanza con el proyecto.
- Lee la bitácora 45 y carga las dudas que tengas al Trello.

ACÀMICA