Breve introducción a HTML para hacer extracción automática

Por: Silvia Gutiérrez [CC BY 2.0]

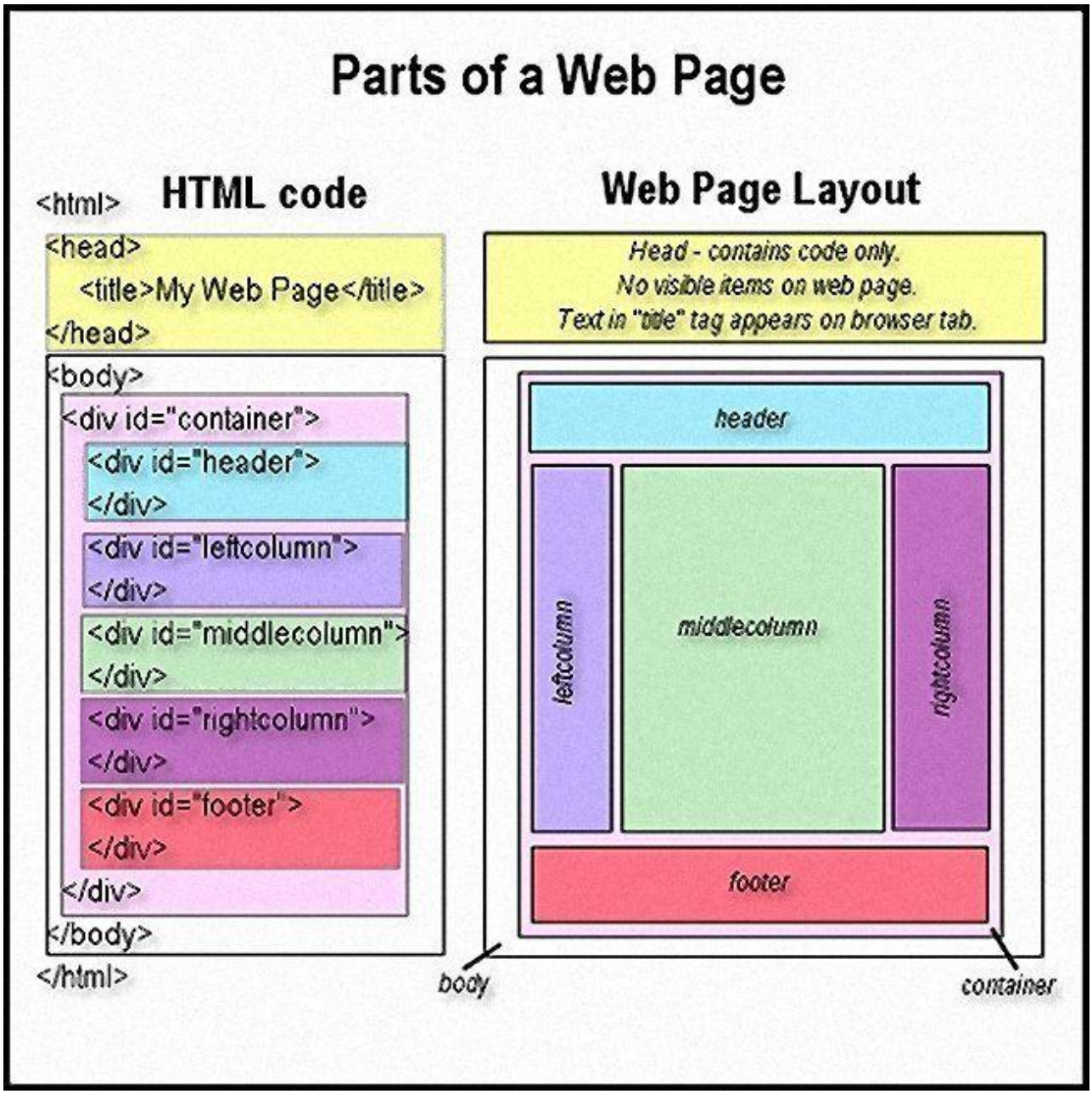
HTML es el acrónimo para Hypertext Markup Language.

Es el **lenguaje utilizado para describir ("marcar") la estructura del contenido** de una página web.

HTML en sí mismo **no determina cómo se *ve* la página** (eso se hace con otro lenguaje: CSS).

Ésta será una breve introducción para entender cómo funcionan estas estructuras para que después puedas extraer contenido específico de las secciones de una página.

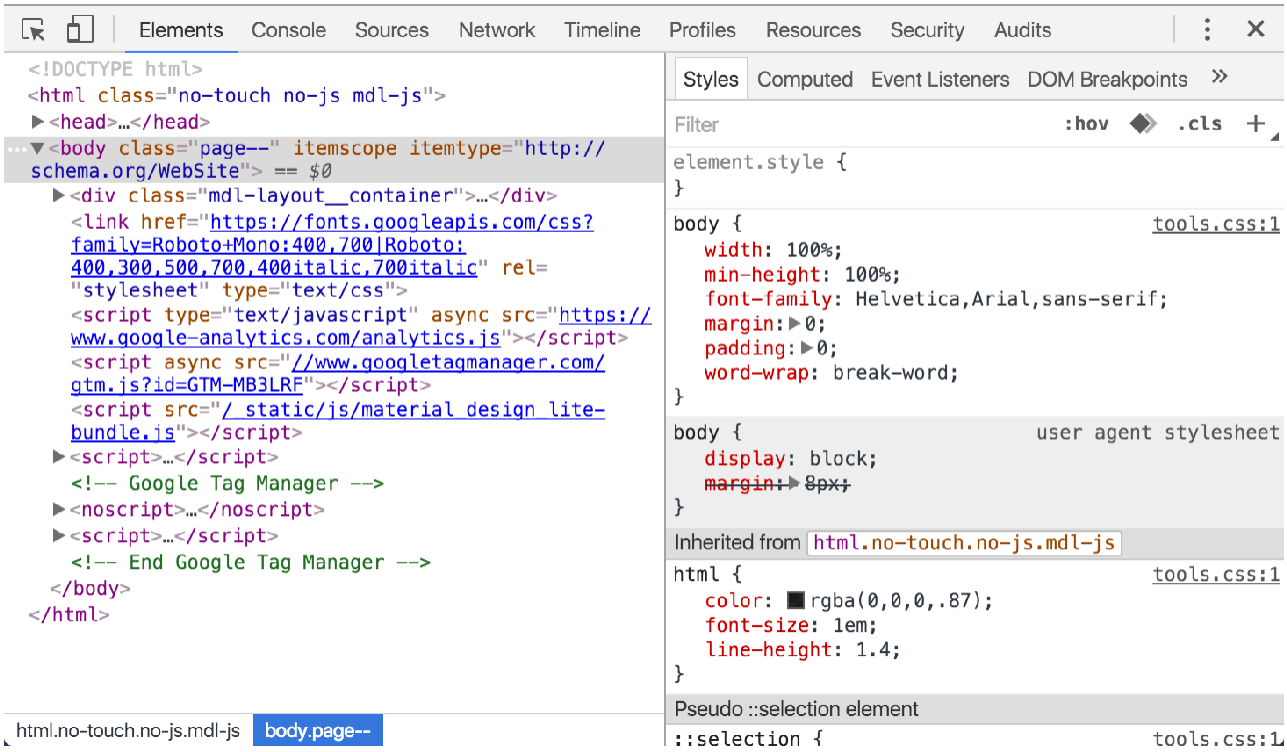
Estructura básica de HTML



La estructura de HTML tiene la **forma de un árbol genealógico** y en cada rama pueden haber anidadas una o más sub-divisiones.

👉 **Ejercicio:**
a) Observa la imagen superior. Ahora abre una página web en Google Chrome y presiona Ctrl+Shift+I, y encuentra las etiquetas descritas arriba en la ventana de "Elementos"

b) Contrae y expande estas etiquetas dando clic sobre los triangulitos, ¿puedes observar la estructura de árbol?

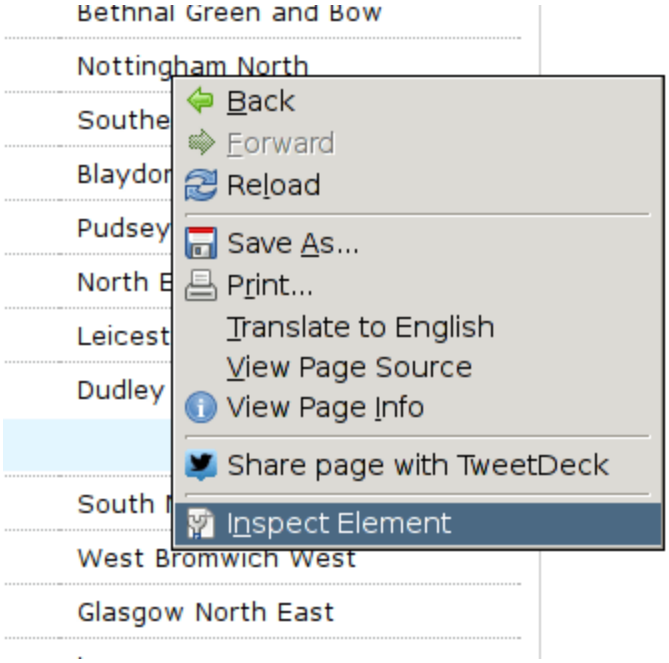


👉 **Ejercicio:** Ahora veremos qué significan estos elementos haciendo pruebas en https://www.w3schools.com/html/html_basic.asp

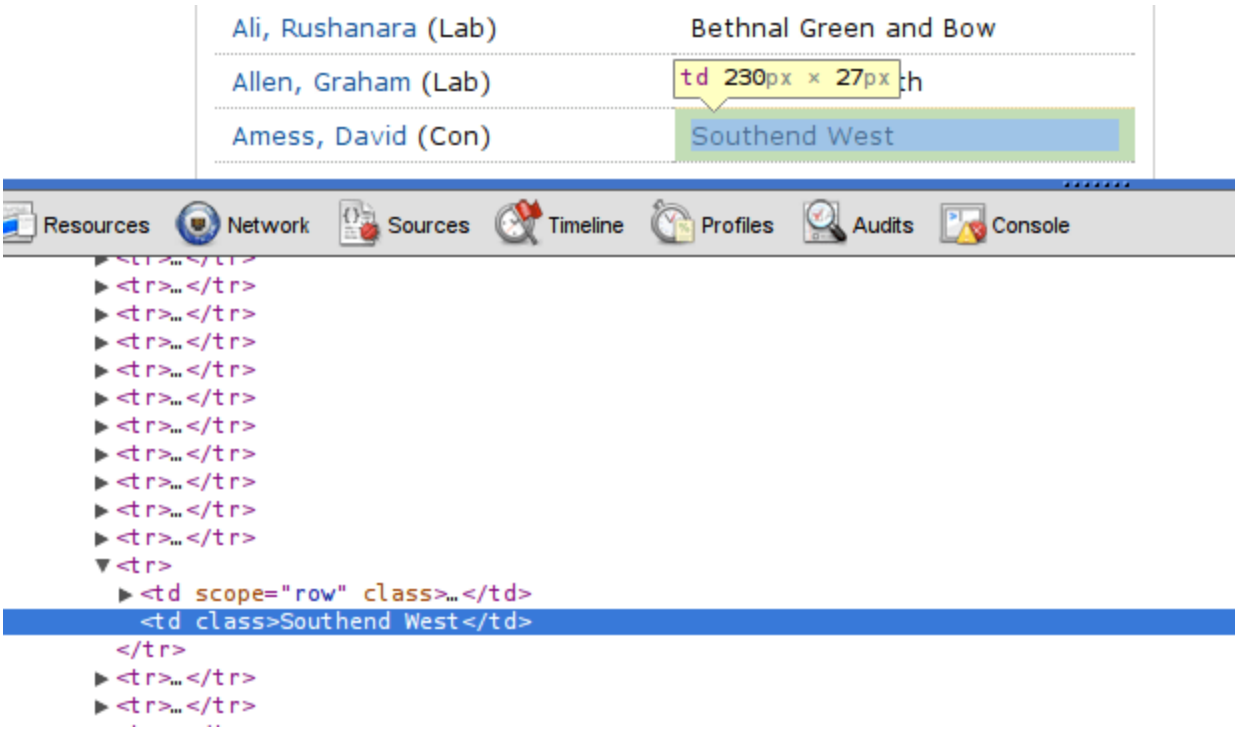
1. Document: Cambia el texto dentro de las etiquetas <h1> y <p> y observa qué ocurre
2. Heading: inserta etiquetas de encabezados con otros números, ¿qué sucede?
3. Paragraphs: inserta otra línea de texto entre etiquetas <p> y deduce para qué sirve esta etiqueta
4. Links: observa esta etiqueta y determina: ¿cómo es su estructura?
5. Images:
   a. observa la estructura de esta etiqueta
   b. cambia los tamaños de altura y ancho (width, height)
   c. cambia la ruta (w3schools.jpg) por una inexistente y observa qué pasa
   d. sustituye la url incorrecta por un la url de un .gif que encuentres en la web

Explorando HTML con Google Chrome

1. Abre la lista de los miembros del parlamento del Reino Unido en Chrome: http://www.parliament.uk/mps-lords-and-offices/mps/
2. En una de las entradas de la lista da clic con tu botón derecho y selecciona "Inspeccionar Elemento"



3. Chrome abrirá una nueva página debajo de la página web donde se mostrará el pedazo de código de HTML de la sección seleccionada



4. Como sabemos, las cuñas (<>) representan las etiquetas (*tags*) de HTML.
5. Mueve tu cursor hacia arriba y hacia abajo para ver el código de otras secciones
6. Puedes expandir o colapsar ciertas secciones haciendo clic sobre las cuñas
7. ¿Notaste que cada una de las filas de la lista de miembros del parlamento está señalada con una etiqueta específica?
8. ¿Qué crees que significa?
9. En un nivel más arriba hay una etiqueta <td></td>, ¿de qué crees que es acrónimo?
10. ¿Para qué se usa?
11. Para tenerlo aún más claro, ve a: https://www.w3schools.com/tags/tag_td.asp

👉Si quieres aprender más de HTML, no te pierdas:
https://www.codecademy.com/es/learn/learn-html