**Web Scraping**

**Extracción automática de información de una página web**. Esta información suele ser el contenido de una página, pero también puede incluir información en los encabezados de la página, enlaces presentes en la página o cualquier otra información incrustada en el HTML de la página.

1º **import requests, import bs4**

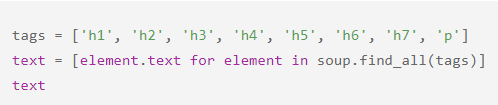
2ºrequests**.get(** url, {headers} **)**

3º **.content()**

4º bs4**.BeautifulSoup(**3º, "html.parser"**)**

5º **.find\_all()**

* **todos los elementos** html indicados:

****

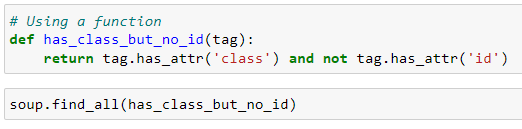
* **algunos elementos** html (los que contienen cierta class, id, …):



→ “table” es un elemento html → [contenido html](https://my.ironhack.com/lms/courses/course-v1:IRONHACK+DAPT+202110_MAD/modules/125b92448525402a887a1f5c7455737c/units/f48258aa73964be69fd62590aac59fd3)

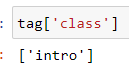
→ {class} es un identificador

* **usar funciones**:

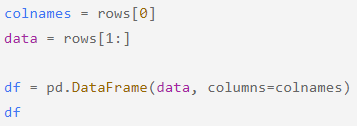
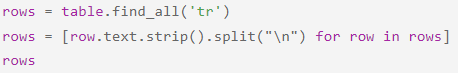


5º - **.name** , **.attrs** , **.string** (texto del elemento), **.html**, **.headers** → si queremos obtener algo muy específico

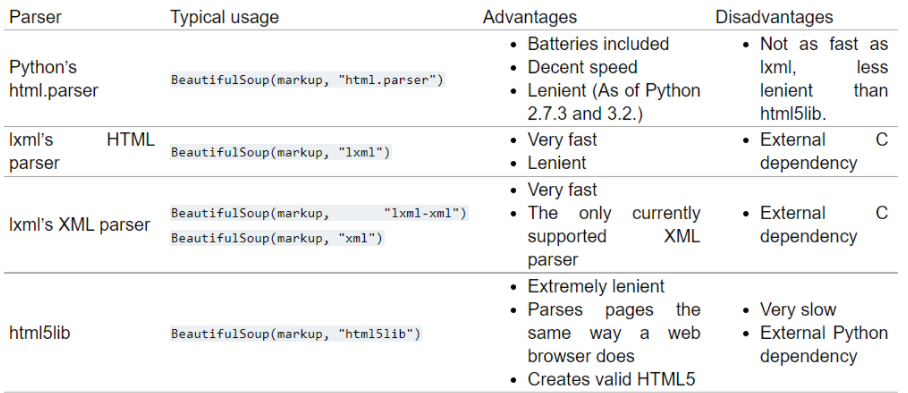
- var**[“class”]**, var**[“id”]** ,… → si queremos obtener las clases, id, etc



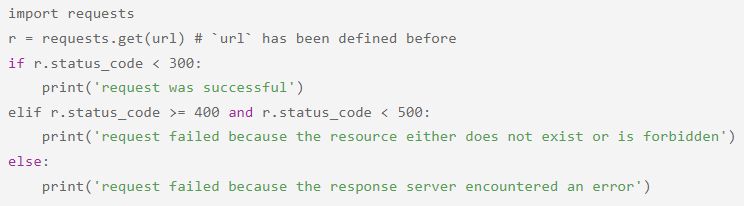
6º **tratar los datos**: es útil crear listas y pasarlas a DataFrames



**TIPOS DE “bs4.BeautifulSoup” PARSERS**



**Añadir en el código excepciones por si las response dan fallo:**



*import requests*

*r = requests.get(url) # `url` has been defined before*

*if r.status\_code < 300:*

*print('request was successful')*

*elif r.status\_code >= 400 and r.status\_code < 500:*

*print('request failed because the resource either does not exist or is forbidden')*

*else:*

*print('request failed because the response server encountered an error')*

**Handling Request Errors**

* **Exceder Max Redirects**: hay un número máximo de redirecciones por defecto (30) que puede anular max\_redirects. Si el número de redireccionamientos excede el límite, requests lanzará una TooManyRedirects excepción

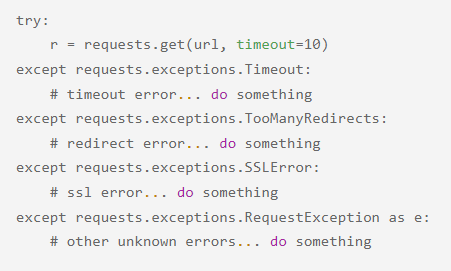




* **Se acaba el tiempo de respuesta del servidor**: requests normalmente esperará un largo período de tiempo hasta que el servidor remoto cierre la conexión y luego lanzará una ConnectTimeout excepción. Sería bueno proporcionar un timeout para limitar la cantidad de tiempo de espera.



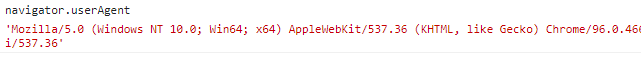
Sería bueno también utilizar try, except para que el código siga corriendo y si hay error te diga que error es

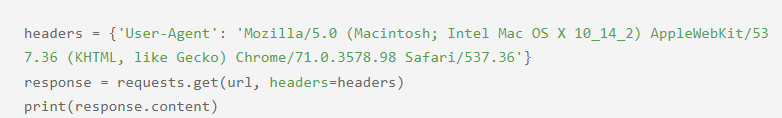


**User Agent**

Identifica el navegador y el sistema operativo desde el cual se realiza la solicitud.

Podemos “falsear” el user agent, por ejemplo, para cuando utilicemos un bot:





**Requests ASÍNCRONAS**

**import asyncio, requests**

****

**NO exceder LÍMITE de requests**

Para controlar la tasa de solicitud de sus scripts, puede esperar un período de tiempo después de realizar cada solicitud llamando time.sleep().



**NO exceder LÍMITE de requests**

Eche un vistazo a Github's robots.txten <https://github.com/robots.txt>

### **ETIQUETAS INICIALES O DE RAÍZ**

* **<!DOCTYPE html>** Indica al navegador que el documento está basado en el estándar HTML5
* **<html> </html>** Representa la raíz de un documento HTML. Todos los demás elementos de la estructura HTML deben ser recogidos dentro de estas etiquetas.

### **METADATOS DEL DOCUMENTO**

* **<head> </head>** Representa una colección de metadatos acerca del documento, incluyendo enlaces a, o definiciones de, scripts y hojas de estilo. El resto de etiquetas de metadatos, **irán recogidas dentro de las etiquetas de apertura y cierre del head**. Importante explicar que estos metadatos del documento, es información para el navegador y no contenido que será visible en la página web. A excepción de la etiqueta <title> que veremos a continuación.
* **<title> </title>** Etiqueta usada para definir el título de la página web.
* **<link>** Se usa para enlazar recursos externos al documento HTML. El ejemplo más común son **las hojas de estilos CSS.**
* **<meta>** Etiqueta usada para definir otros metadatos que no se pueden definir con una etiqueta HTML especifica. Por ejemplo para definir el autor del sitio, o la descripción del mismo.
* **<style> </style>** Etiquetas usadas para introducir código CSS en línea, es decir, en el propio documento HTML.

### **ETIQUETAS DE SECCIONES O PARA ESTRUCTURAR EL HTML**

* **<body> </body>** Al contrario que la etiqueta de metadatos <head>, todo lo que quieras mostrar en la página web debe ir recogido dentro de las etiquetas de apertura y cierre de <body>. **Este contenido será el que se muestre en la web.**
* **<nav> </nav>** Usadas para definir el contenido que será la sección de navegación de la web.
* **<main> </main>** Se usa para definir el contenido principal del documento. Solamente puede existir uno por documento.
* **<section> </section>** Define una sección del documento
* **<article> </article>** Define contenido independiente de la web.
* **<aside> </aside>** Dentro de estas etiquetas suele alojarse el contenido adicional de la web. Suele ser contenido relacionado con la web pero de poca importancia
* **<h1>,<h2>,<h3>,<h4>,<h5>,<h6>** Son etiquetas HTML muy importantes, ya que son usadas para jerarquizar el contenido de la web. Las etiquetas se usan para explicar brevemente el contenido que irá a continuación.
* **<header> </header>** Se usan para definir la cabecera de la página web. Suele contener el logotipo, menú de navegación, etc.
* **<footer> </footer>** Usadas para definir el pie de página.

### **ETIQUETAS PARA LA AGRUPACIÓN DE CONTENIDO**

* **<p> </p>** Etiqueta usada para escribir párrafos de texto.
* **<hr>** Etiqueta utilizada para «romper» entre dos secciones de una web. Usada comúnmente como separador.
* **<pre> </pre>** Usada para pegar texto manteniendo el pre formato propio del texto.
* **<blockquote> </blockquote>** Se usan para indicar que el contenido es texto citado.
* **<ol> </ol>** Etiquetas para crear una lista ordenada
* **<ul> </ul>** Etiquetas para crear una lista des-ordenada
* **<li> </li>** Etiquetas que recogen el contenido de un elemento de una lista, sea ordenada o no.
* **<dl> </dl>** Usada para crear una lista de definiciones.
* **<dt> </dt>** Representa un término definido por la siguiente etiqueta <dd>
* **<dd> </dd>** Se usa para definir los términos listados antes que él.
* **<figure> </figure>** Indica una figura ilustrada como parte del documento HTML5.
* **<figcaption> </figcaption>** Utilizada para definir la leyenda de una figura.
* **<div> </div>** Etiqueta común utilizada para crear un contenedor genérico.

### **ETIQUETAS SEMÁNTICAS PARA TEXTO**

* **<a> </a>** Etiqueta utilizada para crear hiperenlaces en el documento HTML
* **<strong> </strong>** Etiqueta para definir una palabra o conjunto de ellas como importantes. **Tiene una fuerte importancia en el SEO de la página.**
* **<small> </small>** Utilizada para dejar un comentario aparte, del tipo una nota de derechos de autoría, u otros textos que no son esenciales para la comprensión del documento.
* **<cite> </cite>** Para indicar el título de una obra
* **<sub> </sub>** y **<sup> </sup>** Etiquetas utilizadas para representar un subíndice o superíndice.
* **<mark> </mark>** Usada para resaltar texto
* **<span> </span>** Etiqueta HTML sin ningún significado específico. Se usa conjuntamente con los atributos «class» o «id» para atribuirle ciertas características.
* **<br>** Etiqueta utilizada para crear un salto de línea

### **ETIQUETAS PARA INCRUSTAR CONTENIDO**

* **<img>** Etiqueta para «pintar» una imagen en la página web.
* **<iframe> </iframe>** Es una etiqueta que sirve para anidar otro documento HTML dentro del documento principal.
* **<embed>** Usada para integrar una aplicación o contenido interactivo externo que no suele ser HTML.
* **<object> </object>** Utilizada llamar a un recurso externo de la web. Este recurso será tratado como una imagen, o un recurso externo para ser procesado por un plugin.
* **<video> </video>** Se usa para reproducir video en la página web junto a sus archivos de audio y capciones asociadas.
* **<audio> </audio>** Usada para cargar un archivo de audio o stream de audio.
* **<source>** Permite a autores especificar recursos multimedia alternativos para las etiquetas de <video> o <audio>
* **<svg> </svg>** Se usa para llamar a una imagen vectorizada.

### **ETIQUETAS PARA LA CREACIÓN DE TABLAS**

* **<table> </table>** Etiquetas de apertura y cierre de una tabla. El resto de etiquetas de la tabla han de ir siempre recogidas entre estas dos etiquetas.
* **<caption> </caption>** Usada para indicar el título de la tabla.
* **<colgroup> </colgroup>** Etiqueta utilizada para agrupar dos o más columnas de una tabla.
* **<tbody> </tbody>** Usada para describir los datos concretos de una tabla.
* **<thead> </thead>** Indica el bloque de filas que describen las etiquetas de las columnas de la tabla.
* **<tfoot> </tfoot>** Indica los bloques de filas que describen los resúmenes, o datos totales de una columna de una tabla.
* **<tr> </tr>** Se usa para indicar una fila de celdas de una tabla.
* **<td> </td>** Usada para definir una celda de una tabla.
* **<th> </th>** Etiqueta que se usa para definir el encabezado de una celda

### **ETIQUETAS PARA LA CREACIÓN DE FORMULARIOS**

* **<form> </form>** Etiqueta de apertura y cierre de un formulario de una página web. El resto de etiquetas de formulario deben ir siempre recogidas entre estas etiquetas de apertura y cierre de formulario.
* **<fieldset> </fieldset>** Etiqueta que representa un conjunto o agrupación de elementos de un formulario. «Pinta» un recuadro alrededor de las etiquetas que estén contenidas dentro del <fieldset>
* **<legend> </legend>** Etiqueta ligada a <fieldset>. Indica el título del <fieldset>
* **<label> </label>** Se usa para definir el nombre o título de un control del formulario.
* **<input>** Pinta un campo de introducción de datos para el usuario. Es de las principales etiquetas de un formulario.
* **<button> </button>** Etiqueta utilizada para representar un botón en el formulario.
* **<select> </select>** Input que permite una selección entre un conjunto de opciones.
* **<option> </option>** Etiqueta ligada a <select>. Permite añadir diferentes opciones al <select>
* **<textarea> </textarea>** Añade un campo al usuario para que pueda introducir texto en unas líneas máximas de texto que el desarrollador puede definir.

**LISTADO DE ATRIBUTOS HTML5**

| **Atributo** | **Etiqueta HTML5** | **Descripción** |
| --- | --- | --- |
| accept | **<form>, <input>** | Tipos de archivos aceptados por el servidor. Puedes definirlos con valores separados por comas, por ejemplo: .doc, .ods, .txt, etc. O también puedes indicarle un conjunto de tipos de archivos, como por ejemplo: «video/\*» o «audio/\*» |
| accept-charset | **<form>** | Define una lista de caracteres soportados en el formulario. Por ejemplo: accept-charset=»ISO-8859-1″ |
| accesskey | **Atributo Global** | Podrás configurar un botón de tu teclado para que se ejecute acción en la página web. Por ejemplo, un enlace hacia la página de contacto o la home. |
| action | **<form>** | El atributo especifica la URL dónde se enviarán los datos recogidos en el formulario, a menudo un archivo .php dentro del mismo servidor. |
| align | **<applet>, <caption>, <col>, <colgroup>, <hr>, <iframe>, <img>, <table>, <tbody>, <td>, <tfoot> , <th>, <thead>, <tr>** | Especifica el alineamiento horizontal del elemento.  Este atributo sigue funcionando pero no es aconsejable usarlo ya que está en desuso. Lo recomendable es usar reglas CSS en su lugar. |
| alt | **<applet>, <area>, <img>, <input>** | Texto alternativo que se mostrará en el navegador en caso de que la imagen no se pueda mostrar. Muy importante de cara a accesibilidad. |
| async | **<script>** | Indica que el script debe ejecutarse de forma asíncrona. |
| autocomplete | **<form>, <input>** | Indica si los controles en este formulario pueden tener por defecto sus valores completados automáticamente por el navegador. |
| autofocus | **<button>, <input>, <keygen>, <select>, <textarea>** | El elemento debe tener el foco automáticamente una vez se cargue la página web. |
| autoplay | **<audio>, <video>** | Hace que el audio o video comience a reproducirse una vez se cargue la página web. |
| bgcolor | **<body>, <col>, <colgroup>, <marquee>, <table>, <tbody>, <tfoot>, <td>, <th>, <tr>** | Te permite añadir un color de fondo al elemento HTML. El color debe ser añadido en el modo de color RGB, por ejemplo #FF00AA.  Este atributo sigue funcionando pero no es aconsejable usarlo ya que está en desuso. Lo recomendable es usar reglas CSS en su lugar. |
| border | **<img>, <object>, <table>** | Permite definir un borde al elemento HTML.  Este atributo sigue funcionando pero no es aconsejable usarlo ya que está en desuso. Lo recomendable es usar reglas CSS en su lugar. |
| buffered | **<audio>, <video>** | Con este atributo podrás definir un rango de tiempo del elemento multimedia que será usado como buffer. |
| charset | **<meta>, <script>** | Atributo necesario par definir la codificación de caracteres de la página o del script. |
| checked | **<command>, <input>** | Usado para definir que un elemento HTML debe estar marcado por defecto al cargar la página web. |
| cite | **<blockquote>, <del>, <ins>, <q>** | Contiene una URL que apunta a la fuente de la cita o cambio. Muy importante de cara a una correcta optimización SEO de la página. |
| class | **Atributo Global** | Otorga una clase creada en CSS a un elemento HTML. Se usa principalmente para aplicar estilos a elementos con propiedades en común. |
| code | **<applet>** | Especifica la URL del archivo de tipo applet que será cargado y ejecutado. |
| color | **<basefont>, <font>, <hr>** | Atributo usado para indicar el color del texto de un elemento.  Este atributo sigue funcionando pero no es aconsejable usarlo ya que está en desuso. Lo recomendable es usar reglas CSS en su lugar. |
| cols | **<textarea>** | Dentro de una etiqueta <textarea>, se usa para definir el número de columnas que tendrá el campo de texto. |
| colspan | **<td>, <th>** | Dentro de una tabla, define el número de columnas que una celda debe ocupar. |
| content | **<meta>** | Un atributo que permite definir el contenido de la etiqueta HTML <meta> que acompaña. |
| contenteditable | **Atributo Global** | Define si el contenido del elemento HTML podrá ser editable o no. |
| contextmenu | **Atributo Global** | Define el ID del elemento <menu> que servira como menú de contexto para otro elemento. |
| controls | **<audio>, <video>** | Indica si el navegador debe mostrar los controles multimedia del reproductor integrado al usuario. |
| coords | **<area>** | Un conjunto de valores que especifican las coordenadas dentro de la etiqueta area. |
| data | **<object>** | Especifica la URL, absoluta o relativa del objeto o recurso. |
| data-\* | **Atributo Global** | Permite asociar atributos presonalizados a un elemento HTML. |
| datetime | **<del>, <ins>, <time>** | Permite definir la fecha y hora asociadas con el elemento. |
| default | **<track>** | Con este atributo podrás definir que por defecto, una pista por ejemplo de subtítulos esté habilitada. Siempre y cuando las preferencias del usuario no indiquen algo diferente. |
| defer | **<script>** | Indica que el código debe ser ejecutado despues de que la página este cargada. |
| dir | **Atributo Global** | Con este atributo puedes definir la dirección del texto en un elemento HTML. Los valores permitidos son ltr (Izquierda-a-Derecha) o rtl (Derecha-a-Izquierda). |
| dirname | **<input>, <textarea>** | Similar a la anterior. Indica la dirección del texto asociada con el elemento al momento de enviar el formulario. |
| disabled | **<button>, <command>, <fieldset>, <input>, <keygen>, <optgroup>, <option>, <select>, <textarea>** | Permite que un elemento HTML esté deshabilitado y el usuario no pueda interactuar con el. |
| download | **<a>, <area>** | Si este atributo se añade al elemento HTML, el recurso indicado podrá descargarse al dispositivo del usuario. |
| draggable | **Atributo Global** | Define si el elemento puede ser arrastrado. |
| enctype | **<form>** | Define el tipo de contenido del formulario cuando el método del mismo es POST. |
| for | **<label>, <output>** | Atributo usado para asociar o describir elementos que pertenecen a éste. |
| form | **<button>, <fieldset>, <input>, <keygen>, <label>, <meter>, <object>, <output>, <progress>, <select>, <textarea>** | Atributo para indicar el formulario al que pertenece el elemento. |
| headers | **<td>, <th>** | A través de «ids» indicados en los encabezados de una tabla <th> podrás asociarlos a diferentes elementos de la tabla <td>. Muy útil de cara a la accesibilidad y por lo tanto para el SEO. |
| height | **<canvas>, <embed>, <iframe>, <img>, <input>, <object>, <video>** | Permite especificar la altura de los elementos HTML indicados.  Este atributo sigue funcionando pero no es aconsejable usarlo ya que está en desuso. Lo recomendable es usar reglas CSS en su lugar. |
| hidden | **Atributo Global** | Este atributo permite indicar que el elemento HTML no debe ser visible. Es habitual usarlo por ejemplo, para ocultar elementos que no deben ser visibles inicialmente pero que después de una acción si deben verse. |
| high | **<meter>** | El atributo especifica el intervalo donde el atributo «value» tomara el valor como alto. |
| href | **<a>, <area>, <base>, <link>** | La URL de un recurso asociado. |
| hreflang | **<a>, <area>, <link>** | Especifica el lenguaje del recurso asociado. |
| http-equiv | **<meta>** | Indica una directiva que puede alterar el comportamiento de la página y como es leída por los exploradores. |
| id | **Atributo Global** | Usado para asignar un identificador único a un elemento HTML. Usado generalmente para acceder al mismo desde CSS o Javascript. No pueden existir dos «ids» iguales en una misma página web. |
| ismap | **<img>** | Indica que la imagen es parte de un mapa de imagen del servidor. |
| itemprop | **Atributo Global** | Atributo que permite dar más información a Google acerca de imágenes o datos dentro de cualquier tipo de etiquetas. Por ejemplo se puede indicar propiedades como: actores , clasificación, género, etc . |
| label | **<track>** | Especifica el título de la pista con un formato legible por el usuario. Usado conjuntamente con el atributo «for» |
| lang | **Atributo Global** | Define el lenguaje utilizado en el elemento. |
| language | **<script>** | Define el lenguaje (de programación) utilizado en el script. |
| list | **<input>** | Identifica una serie de valores predefinidos para ser sugeridos al usuario en un menú desplegable. |
| loop | **<audio>, <bgsound>, <marquee>, <video>** | Atributo que habilita al medio a volver a reproducirse desde el principio cuando llega al final. |
| low | **<meter>** | Similar al atributo high, solo que en este caso, permite definir cuando un valor estará en el rango bajo. |
| max | **<input>, <meter>, <progress>** | Define el máximo valor aceptado. |
| maxlength | **<input>, <textarea>** | Define el mayor número de caracteres aceptados. |
| media | **<a>, <area>, <link>, <source>, <style>** | Indica el tipo de medio para el cual el recurso vinculado fue creado. |
| method | **<form>** | Define el método HTTP a emplear para enviar el formulario. Puede ser GET (predeterminado) o POST. |
| min | **<input>, <meter>** | Define el valor mínimo aceptado. |
| multiple | **<input>, <select>** | Atributo que define si se puede introducir múltiples valores. |
| muted | **<video>** | Al reproducir un contenido audiovisual, este atributo indica si el audio será silenciado inicialmente al cargar la página. |
| name | **<button>, <form>, <fieldset>, <iframe>, <input>, <keygen>, <object>, <output>, <select>, <textarea>, <map>, <meta>, <param>** | Nombre del elemento. Utilizado a menudo por el servidor para identificar los campos en el envío de formularios. |
| novalidate | **<form>** | Con este atributo indicamos al formulario que no valide el contenido del mismo. |
| open | **<details>** | Indica si los detalles de la página serán mostrados cuando cargue la misma. |
| optimum | **<meter>** | Con este atributo se define el valor numérico óptimo. |
| pattern | **<input>** | Este atributo define una expresión regular con la cual el valor del elemento debe ser validado antes de aceptar el contenido del mismo. |
| ping | **<a>, <area>** | Indica si se enviara un ping o notificacion a la URL del recurso cuando el usuario utlize el vinculo. |
| placeholder | **<input>, <textarea>** | Proporciona una sugerencia al usuario de lo que puede ser introducido en el campo de un formulario por ejemplo. Muy util de cara a la accesibilidad. |
| poster | **<video>** | Atributo que permite añadir una imagen que se mostrará mientras se carga un video. En caso de que la URL del video este «rota», el poster seguirá cargándose consiguiendo con esto no ver un hueco vacio en la web. |
| preload | **<audio>, <video>** | Indica si el recurso o partes del mismo deben ser precargadas. |
| readonly | **<input>, <textarea>** | Con este atributo se define si el elemento puede o no ser editado por el usuario. |
| rel | **<a>, <area>, <link>** | Especifica la relación entre el objeto destino y el objeto enlace. Importante de cara una correcta optimización SEO. |
| required | **<input>, <select>, <textarea>** | Con este atributo hacemos que la etiqueta Html sea, o no, de obligada cumplimentación. |
| reversed | **<ol>** | Con este atributo podemos modificar el comportamiento estándar de una lista ordenada haciendo que se vea listada en orden descendente en lugar de ascendente. |
| rows | **<textarea>** | Define el número de filas en un área de texto. |
| rowspan | **<td>, <th>** | Define el número de filas que una celda de una tabla debe abarcar. |
| sandbox | **<iframe>** | Lista de restricciones a ser desactivadas en el iframe. |
| scope | **<th>** | En una tabla permite definir las celdas que la cabecera deberá ocupar en este elemento. |
| seamless | **<iframe>** | Indica si el iframe debe ser cargado como parte del mismo documento. |
| selected | **<option>** | Define un valor que aparecerá por defecto seleccionado al cargar la página. |
| size | **<input>, <select>** | Define el ancho del elemento (en píxeles). En este caso es mejor por tanto usar reglas CSS. En cambio. si el atributo del elemento es del tipo «text» o «password», entonces es el número de caracteres que se podrán introducir. |
| span | **<col>, <colgroup>** | Indica el número de columnas que se agrupan en una tabla. |
| spellcheck | **Atributo Global** | Con este atributo se permite que un texto pueda ser autocorregido ortográficamente, o no. |
| src | **<audio>, <embed>, <iframe>, <img>, <input>, <script>, <source>, <track>, <video>** | La URL del contenido integrable en la etiqueta HTML que lo acompaña. |
| srcdoc | **<iframe>** | Especifica el contenido de la página incluida. |
| srclang | **<track>** | Especifica el lenguaje del contenido de la pista. |
| srcset | **<img>** | Este atributo define diferentes versiones de la imagen para mostrar una o otra según el tipo de pantalla. |
| start | **<ol>** | Define el número inicial de la lista ordenada (si es diferente a 1). |
| style | **Atributo Global** | Define los estilos CSS que anulan los estilos establecidos previamente. En este caso son llamados estilos en línea ya que se introducen directamente en la etiqueta HTML en cuestión. |
| target | **<a>, <area>, <base>, <form>** | Con este atributo se determina el marco destino en un vínculo.  Valores especiales: «\_blank» indica una nueva ventana, «\_parent» indica el marco padre del marco que contiene el código fuente, «\_self» indica el marco donde está el código fuente y «\_top» indica la ventana completa del navegador. |
| title | **Atributo Global** | Texto a ser mostrado cuando el cursor esté sobre el elemento. Muy útil de cara a la usabilidad y para una correcto posicionamiento SEO. |
| type | **<button>, <input>, <command>, <embed>, <object>, <script>, <source>, <style>, <menu>** | Define el tipo de elemento HTML. |
| usemap | **<img>, <input>, <object>** | Indica la URL parcial de un image map asociado con el elemento.  Note: Este atributo no puede ser usado si el elemento <img> desciende de un <button> o <a> |
| value | **<button>, <option>, <input>, <li>, <meter>, <progress>, <param>** | Define el valor predeterminado a ser mostrado al cargar la página. |
| width | **<canvas>, <embed>, <iframe>, <img>, <input>, <object>, <video>** | Para los elementos enumerados aquí, esto establece el ancho del elemento. Aunque como en caso anteriores, lo correcto sería usar reglas CSS. |
| wrap | **<textarea>** | Indica la forma en la cual el texto debe ser envuelto. |