

INTRODUCCIÓN A WEB SCRAPING CON R

Expositores






Francisco Pautt

 [linkedin.com/in/francisco-pautt](https://www.linkedin.com/in/francisco-pautt)
 [franpautt](https://github.com/franpautt)
 [@franpautt](https://www.instagram.com/franpautt)






Juan David Ibáñez

 [linkedin.com/in/juan-ibáñez](https://www.linkedin.com/in/juan-ibáñez)
 [juandibanezc](https://github.com/juandibanezc)
 [@juandibanezc](https://www.instagram.com/juandibanezc)



Carlos Granadillo

 [linkedin.com/in/carlos-granadillo](https://www.linkedin.com/in/carlos-granadillo)
 [CarlosGranadillo](https://github.com/CarlosGranadillo)
 [@carlosgranadillo](https://www.instagram.com/carlosgranadillo)



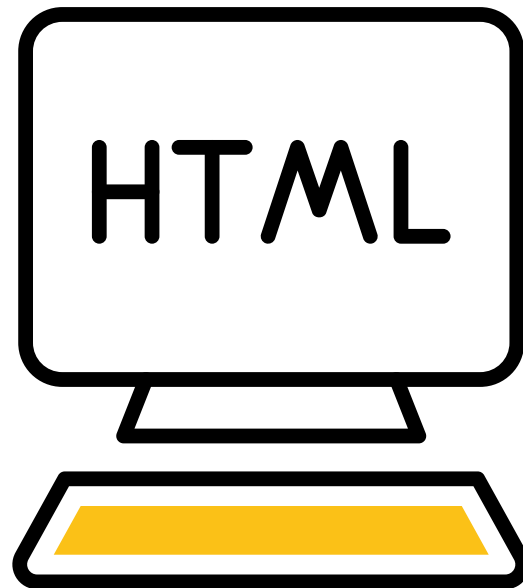
¿Qué es el Web Scraping?

El Web Scraping (“raspado” de páginas web) consiste en la extracción de datos significativos de una o varias páginas web determinadas, para una manipulación o análisis posterior.

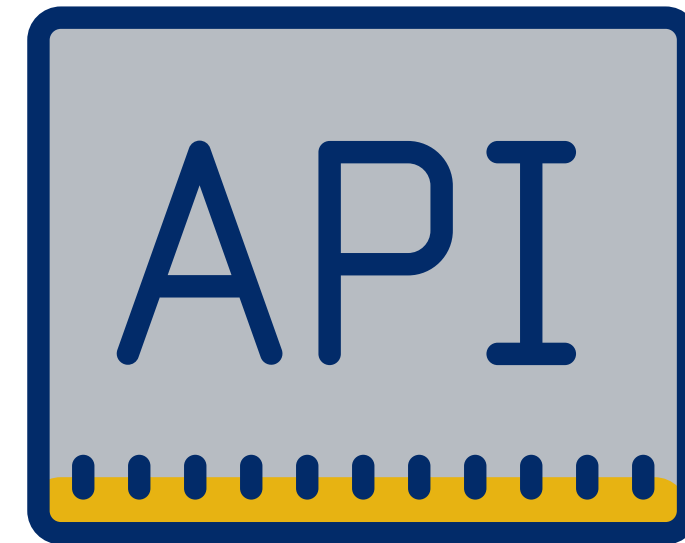
Tipos de Web Scraping

Dentro de la actividad del Web Scraping encontramos dos escenarios diferentes:

Screen scraping: extrae datos del código fuente del sitio web, con el analizador html (fácil) o la coincidencia de expresiones regulares (menos fácil).



API web (application programming interface): el sitio web ofrece un conjunto de solicitudes http estructuradas que devuelven archivos JSON o XML.



Importancia y aplicaciones

El Web Scraping es aplicable en distintos sectores:



Comercial y ventas

Añadir datos a nuestras bases de datos de clientes, prospectos, suscriptores, etc.



Monitoreo de precios de la competencia



Investigación de mercado

Investigar compradores, tendencias y monitorizar nuestra marca.



Detectar influencers

Planificar la campaña de marketing de tu empresa.



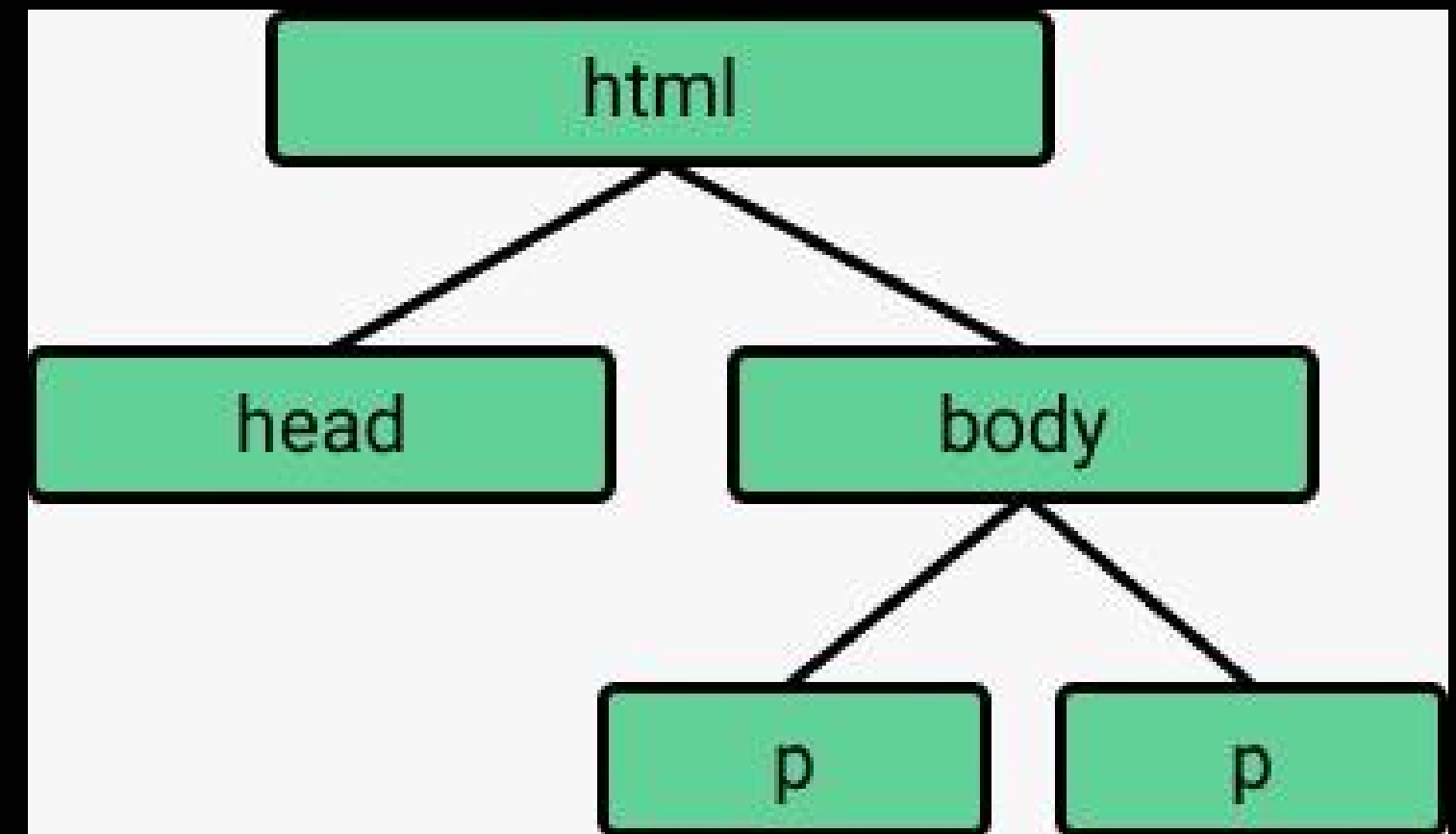
HTML

Breve Introducción



Las siglas HTML significan **Hyper Text Markup Language**, un lenguaje que describe la estructura de las páginas web. Algunas características:

- Hay una jerarquía. Página de HTML esta compuesta de elementos.
- Los elementos tienen tags.
- Tags son nombres para el contenido (título, tabla, etc.)
- `<algo>` abre una sección (algo es el tag de la sección).
- `</algo>` cierra esa sección.
- Entre apertura y cierre se encuentra el contenido.



Una sección puede y en general tiene más secciones adentro. De manera que podemos tener algo parecido a la figura, que se lee:

- `<html>`: abro sección html (página)
- `<head>`: abro encabezado
- `<title>`: abro título
- Page Title: el título de la página (**Esto es contenido**)
- `</head>`: cierro encabezado
- `<body>`: abro cuerpo
- `<h1>`: abro heading
- My First Heading: el heading (**Esto es contenido**)
- `</h1>`: cierro heading
- `<p>`: abro párrafo
- My First Paragraph: el párrafo (**Esto es contenido**)
- `</p>`: cierro párrafo
- `</body>`: cierro cuerpo
- `</html>`: cierro la página

```
<html>
<head>
<title>Page Title</title>
</head>
<body>

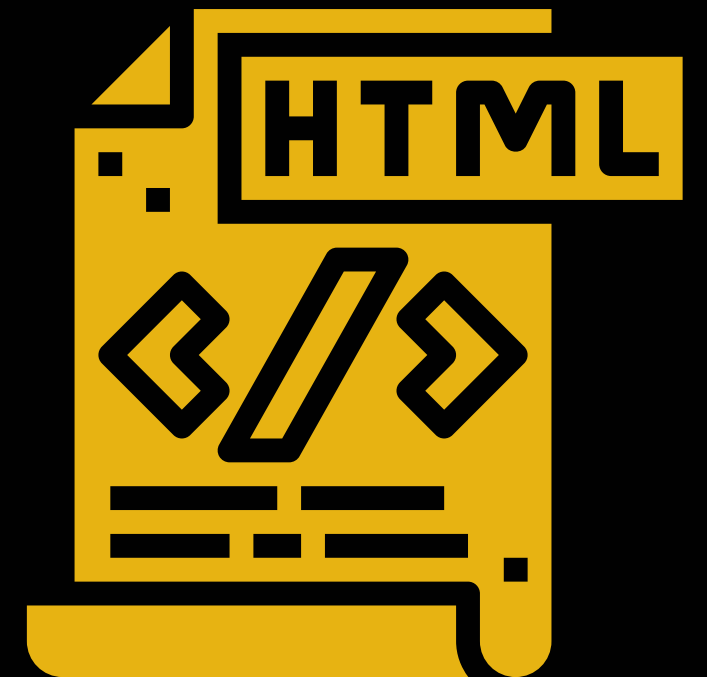
<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
</html>
```


Todos los elementos de HTML pueden tener atributos que nos dan información adicional para ese elemento (por ejemplo, su formato).

Algunos atributos vienen preespecificados:

- **href** significa que lo que sigue es un link
- **src** significa que lo que sigue es el nombre de archivo de una imagen
- **lang** permite especificar lenguaje
- **style** permite especificar estilo (color, tamaño de fuente, etc)
- **width** permite especificar el ancho de una imagen
- **height** permite especificar el alto de una imagen



Un atributo especificado por el usuario es una clase:

- Acá todas las ciudades que pertenecen a la Clase de “**cities**” van a tener fondo negro y color Blanco

```
<!DOCTYPE html>
<html>
<head>
<style>
.cities {
  background-color: black;
  color: white;
  margin: 20px;
  padding: 20px;
}
</style>
</head>
<body>

<div class="cities">
  <h2>London</h2>
  <p>London is the capital of England.</p>
</div>

<div class="cities">
  <h2>Paris</h2>
  <p>Paris is the capital of France.</p>
</div>

<div class="cities">
  <h2>Tokyo</h2>
  <p>Tokyo is the capital of Japan.</p>
</div>

</body>
</html>
```



Paquete Rvest

Rvest es un paquete que permite la realización del Web Scraping en el software, dicho paquete permite la extracción de datos de la web y transformarlos en información útil. Este paquete está diseñado para trabajar junto con **magittr** para facilitar labores en el web scraping.

A continuación, en **negrita**, se listan las funciones más importantes de **Rvest**. Entre comillas se describirán los parámetros más usados.

- **read_html(«url»)** con esta función se crea un objeto que contiene todo el código o etiquetas HTML.
- **html_nodes(«objeto html», «etiqueta css»)** es usada para seleccionar partes del objeto que contiene todo el código html. El segundo parámetro es la clase CSS que está relacionada con la sección que deseamos extraer.
- **html_name()** obtiene los atributos html
- **html_text()** extrae el texto html
- **html_attr()** regresa los atributos específicos html
- **html_attrs()** obtiene los atributos html
- **html_table()** convierte una tabla html en una estructura de datos en R

Ejemplo básico



Después de tanta teoría, pasemos a la práctica!



Caso de aplicación

IMDb

- 1. The Godfather** (1972)
R | 175 min | Crime, Drama
★ 9.2 ☆ Rate **100** Metascore
The aging patriarch of an organized crime dynasty transfers control of his clandestine empire to his reluctant son.
Director: [Francis Ford Coppola](#) | Stars: [Marlon Brando](#), [Al Pacino](#), [James Caan](#), [Diane Keaton](#)
Votes: 1,570,798 | Gross: \$134.97M
[Watch on Prime Video](#)
Included with Prime
- 2. Schindler's List** (1993)
R | 195 min | Biography, Drama, History
★ 8.9 ☆ Rate **94** Metascore
In German-occupied Poland during World War II, industrialist [Oskar Schindler](#) gradually becomes concerned for his Jewish workforce after witnessing their persecution by the Nazis.
Director: [Steven Spielberg](#) | Stars: [Liam Neeson](#), [Ralph Fiennes](#), [Ben Kingsley](#), [Caroline Goodall](#)
Votes: 1,182,082 | Gross: \$96.90M
- 3. 12 Angry Men** (1957)
Approved | 96 min | Crime, Drama
★ 8.9 ☆ Rate **96** Metascore
A jury holdout attempts to prevent a miscarriage of justice by forcing his colleagues to reconsider the evidence.
Director: [Sidney Lumet](#) | Stars: [Henry Fonda](#), [Lee J. Cobb](#), [Martin Balsam](#), [John Fiedler](#)
Votes: 667,640 | Gross: \$4.36M
- 4. Life Is Beautiful** (1997)
PG-13 | 116 min | Comedy, Drama, Romance
★ 8.6 ☆ Rate **89** Metascore
When an open-minded Jewish librarian and his son become victims of the Holocaust, he uses a perfect mixture of will, humor, and imagination to protect his son from the dangers around their camp.
Director: [Roberto Benigni](#) | Stars: [Roberto Benigni](#), [Nicoletta Braschi](#), [Giorgio Cantarini](#), [Giustino Durano](#)
Votes: 604,959 | Gross: \$57.60M

Video TARZAM es una empresa familiar que alquila juegos y películas. Dicha empresa, lo contrata a usted para que los asesore planteando e implementando nuevas estrategias para poder seguir en el mercado a pesar de la creciente competencia que hay en este sector.

Usted piensa en varias ideas, entre esas:

Hacer gráficos atractivos que muestren curiosidades de las mejores películas que ha habido, para que más personas se interesen en ver este tipo de filmes.

Aun así, se podrían interesar, pero seguir recurriendo a plataformas como Netflix, Amazon Prime, ect. y no a video TARZAM. Por esta razón usted complementa la estrategia dando este valor añadido:

“Todo el que alquile una película con video TARZAM se le enviará un video resumen con las mejores partes y puntos destacados”.

Hay un problema: Video TARZAM le dice que no tiene un registro de su información, pues ellos no conocían la importancia de ir guardando los datos.

Pero, para hacer los gráficos usted necesita datos, para analizar con que películas empezar la estrategia de los videos y en general, para quizás descubrir otras oportunidades.

Usted se da cuenta que debe construir la base de datos, pero no hay problema porque sabe que hay un sitio web llamado IMDb de donde puede sacar información. Y eso es lo que vamos a hacer.

¡Muchas gracias por su atención!

