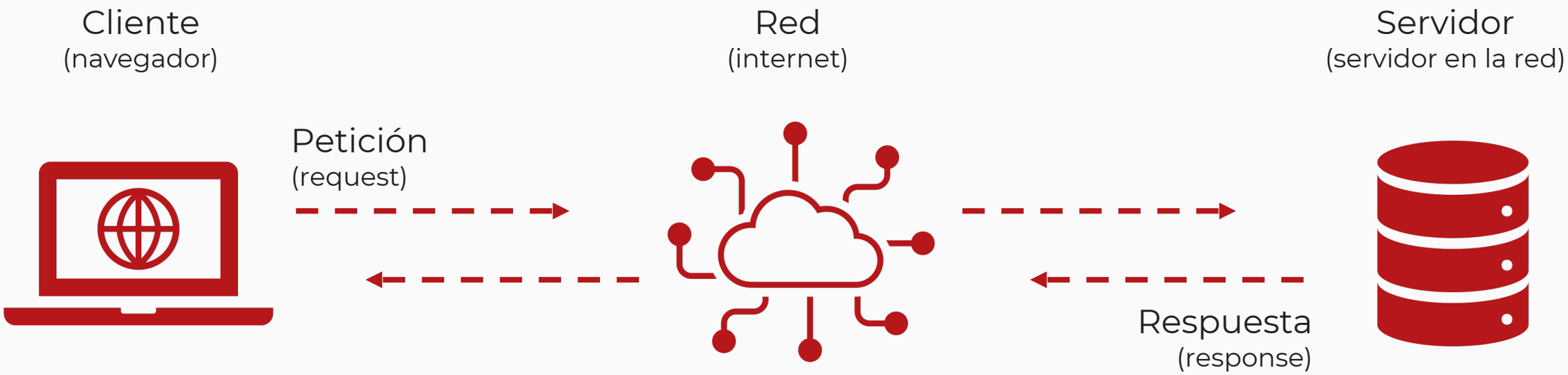
Tecnológico de Monterrey
Escuela de Ciencias Sociales
y Gobierno

Introducción a Raspado Web (webscraping)

Mtro. René Rosado González
Director de Programa LTP

Hypertext Transfer Protocol (HTTP)

La base de la comunicación de datos en la red



Hypertext Transfer Protocol (HTTP)

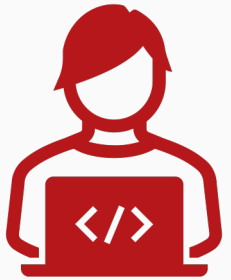
Recursos Elementales

Hypertext

Hyperlinks

Hypermedia

Scripts



Lenguaje de Marcado de HiperTexto

HyperText Markup Languaje (HTML)

Lo que vemos

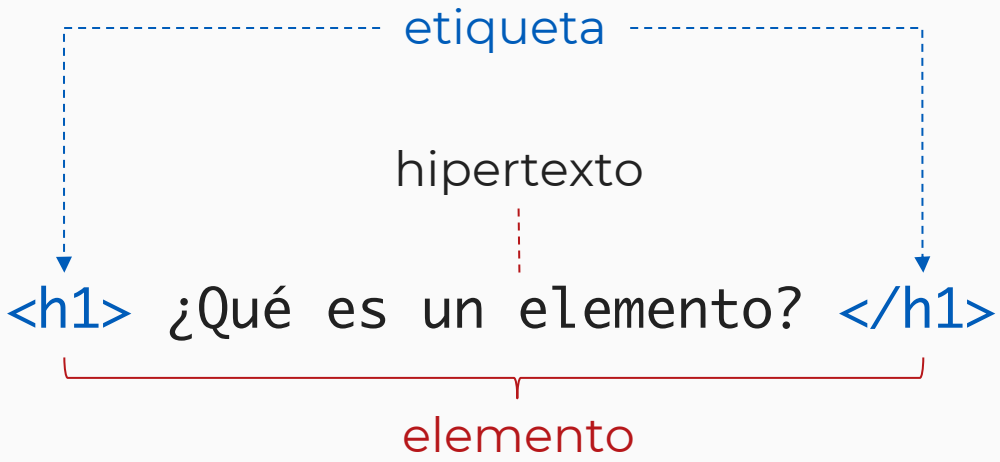
Lo que es

```
  </head>
  <body class="no-admin path-node page-node-type-adm-carrera-final">
    <!-- Google Tag Manager (noscript) -->
    <noscript><iframe src="https://www.googletagmanager.com/ns.html?id=GTM-TD7F4V7" height="0" width="0" style="display:n
    <!-- End Google Tag Manager (noscript) -->
    <a href="#main-content" class="visually-hidden focusable">
      Pasar al contenido principal
    </a>

      <div class="dialog-off-canvas-main-canvas" data-off-canvas-main-canvas>
        <div>


<div  id="block-displayhamburger" class="block">



<div class="menu-overlay display-hamburger">
    <div class="zone-logo">
        <i class="material-icons close">close</i>
        <!-- logo -->
        <section id="" class="logo is-active_logo top-menu-logo">

<a x-ms-format-detection="none"

    href="/"
>      <svg version="1.1" viewBox="0 0 195 52" height="42px" width="195px" xmlns:xlink="http://www.w3.org/1999/xlink" xml
    </a>
</section>
    </div>
    <div class="search-close">
        <div class="enlace-llave">

              <ul   class="menu-menu-llave">
              <li>
        <a href="">Llave</a>
                              <i class="material-icons">vpn_key</i>
      <ul  class="menu-menu-llave submenu-llave">
              <li>
        <a href="https://mitec.itesm.mx/">Alumnos</a>
              </li>
          <li>
        <a href="https://mitecpadres.itesm.mx/">Padres</a>
              </li>
          <li>
        <a href="https://exatec.itesm.mx/">Egresados</a>
              </li>
          <li>
        <a href="http://miespacio.itesm.mx/">Profesores</a>
              </li>
          <li>
```

Lenguaje de Marcado de HiperTexto

HyperText Markup Languaje (HTML)

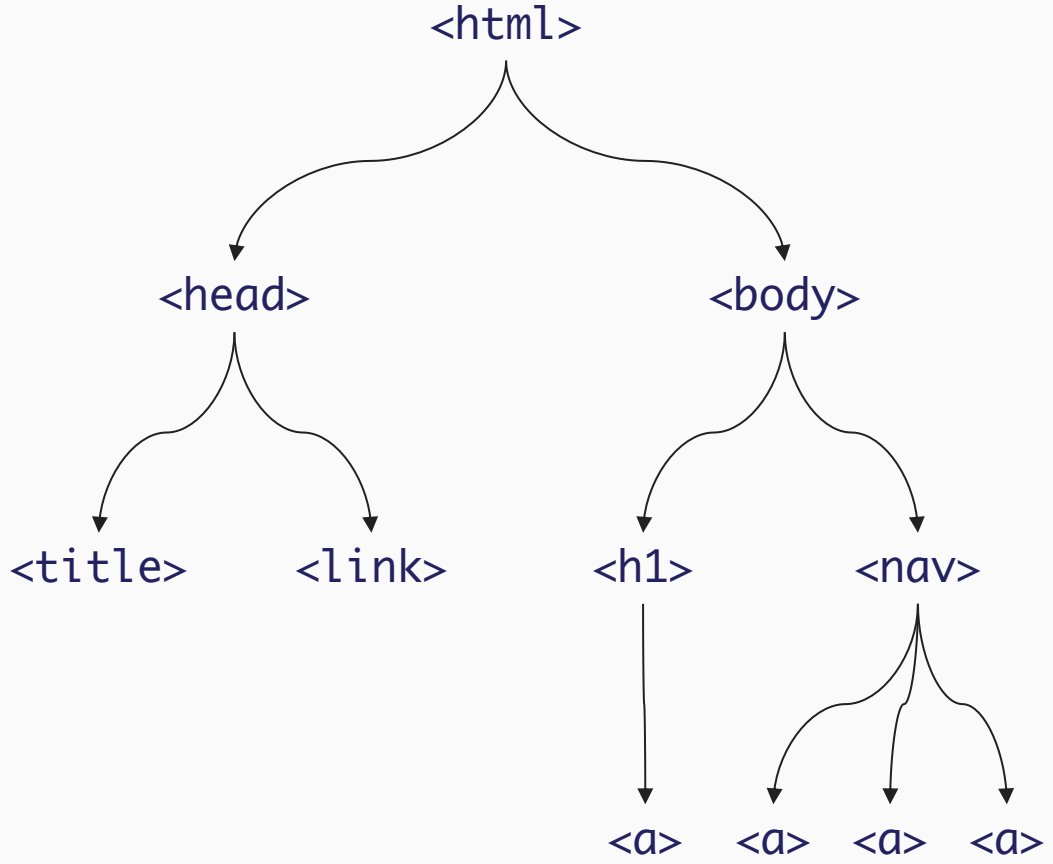
- Hipertexto (*hypertext*) se refiere a enlaces que conectan páginas web entre sí, ya sea dentro de un único sitio web o entre sitios web.
- HTML utiliza "marcado (markup)" para anotar contenido a mostrar en un navegador web.
- El marcado HTML incluye *elementos* diferenciados por *etiquetas*, que consisten en el nombre del elemento rodeado por "<" y ">".



Modelo de Objetos de Documento

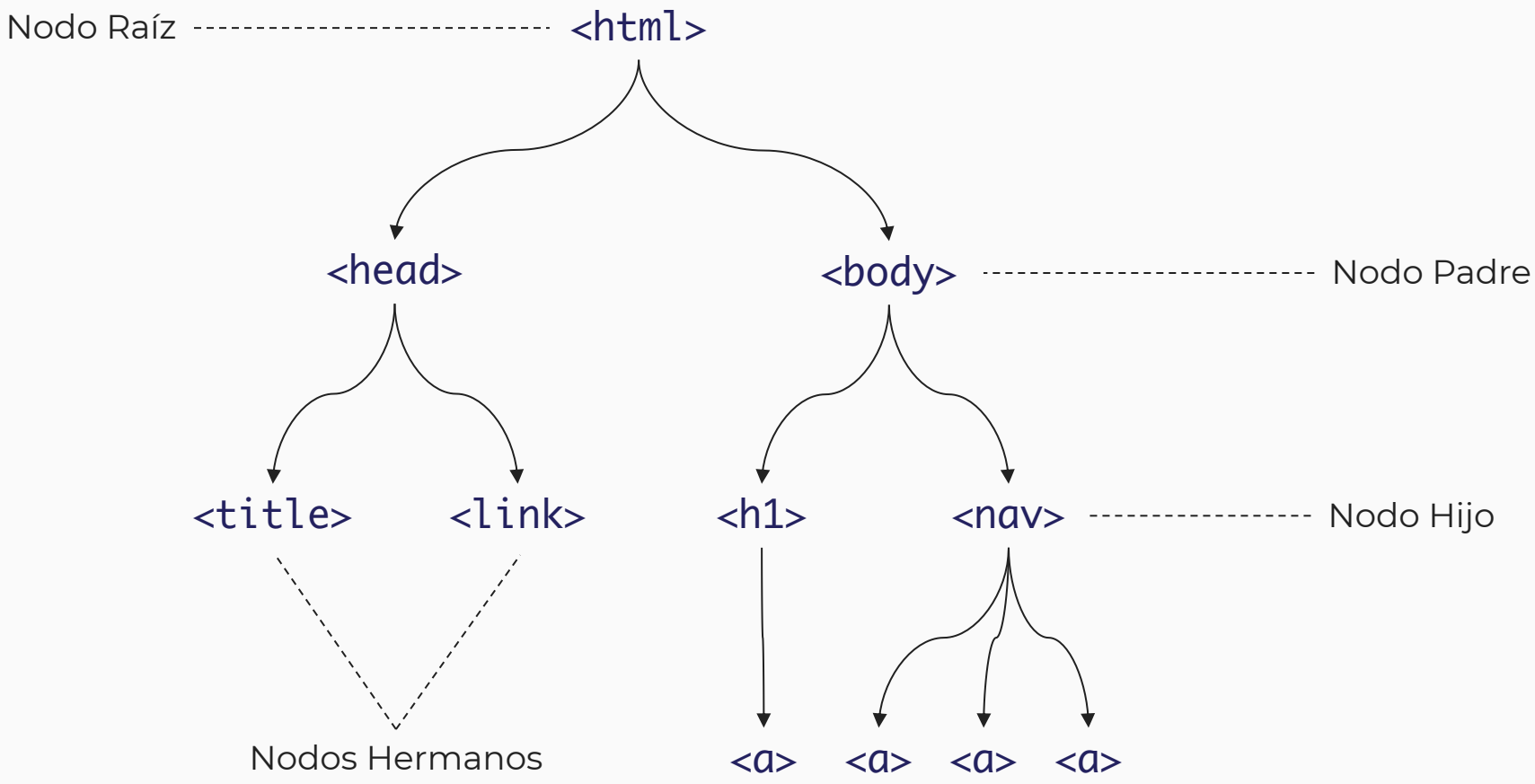
Document Object Model (DOM)

```
1  <!DOCTYPE html>
2  <html>
3      <head>
4          <title>Example</title>
5          <link rel="stylesheet" href="st
6      </head>
7      <body>
8          <h1>
9              <a href="/">Header</a>
10         </h1>
11         <nav>
12             <a href="one/">One</a>
13             <a href="two/">Two</a>
14             <a href="three/">Three</a>
15         </nav>
```



Modelo de Objetos de Documento

Document Object Model (DOM)



Atributos de HTML

HTML

Identificador único

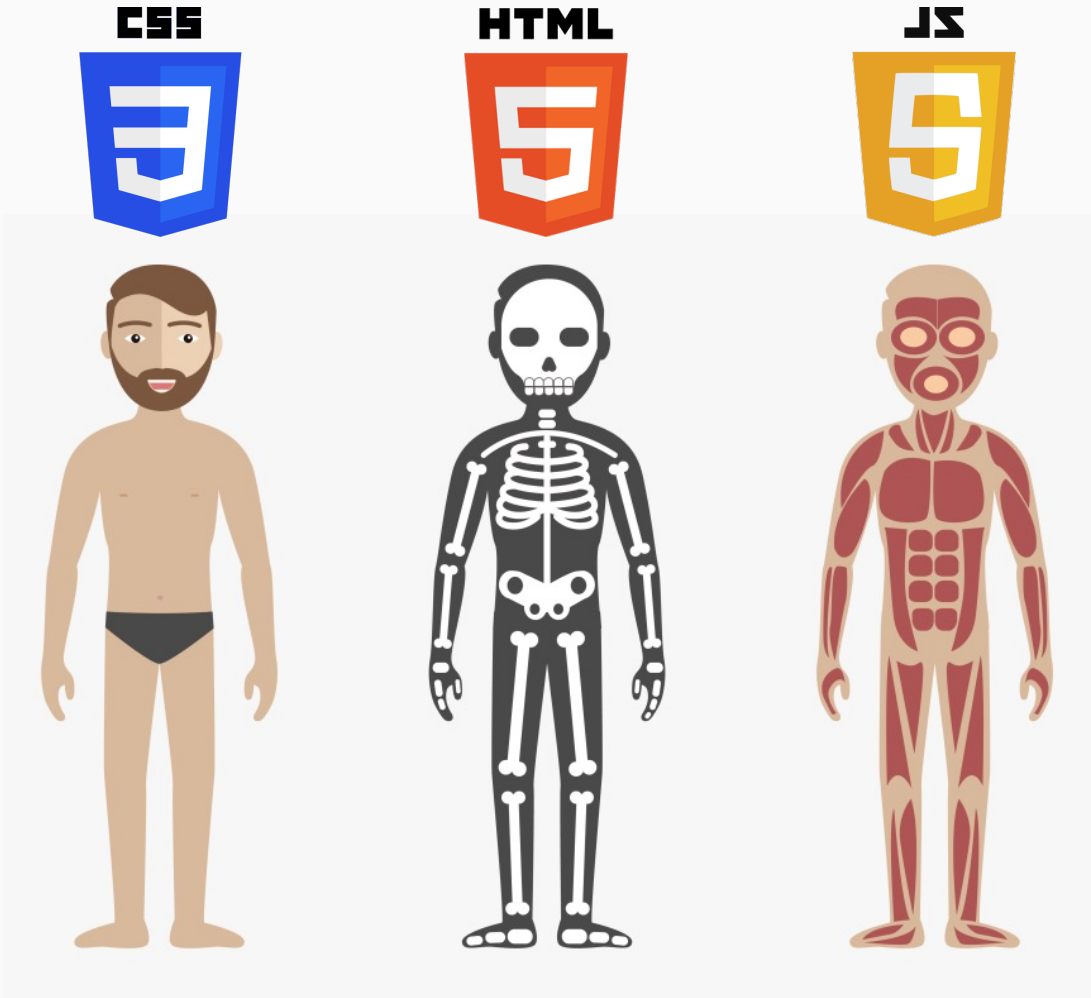
Clase para grupos

Estilo

```
<h1 id="ejemplo" class="bueno" style="font:Montserrat"
href="https://www.w3schools.com/html/html_attributes.asp">

    Ejemplo

</h1>
```

Hypervínculo

La Triada del Diseño Web



Hoja de Estilos en Cascada

Cascading Style Sheets (CSS)

HTML

```
<p>
        Esto es una etiqueta
</p>
<p id="MiID">
        Esto es una etiqueta cool
</p>
<p class="MiClase">
        Esto es una etiqueta aun más cool
</p>
```

CSS

```
#MiID { font-size: 25px; }
.MiClase { font-family:Montserrat;
           font-size: 50px; }
p { color:"darkred"; }
```

Output

Esto es una etiqueta

Esto es una etiqueta cool

Esto es una etiqueta aun más cool

Hoja de Estilos en Cascada

Cascading Style Sheets (CSS)

Selector ----▸ p {

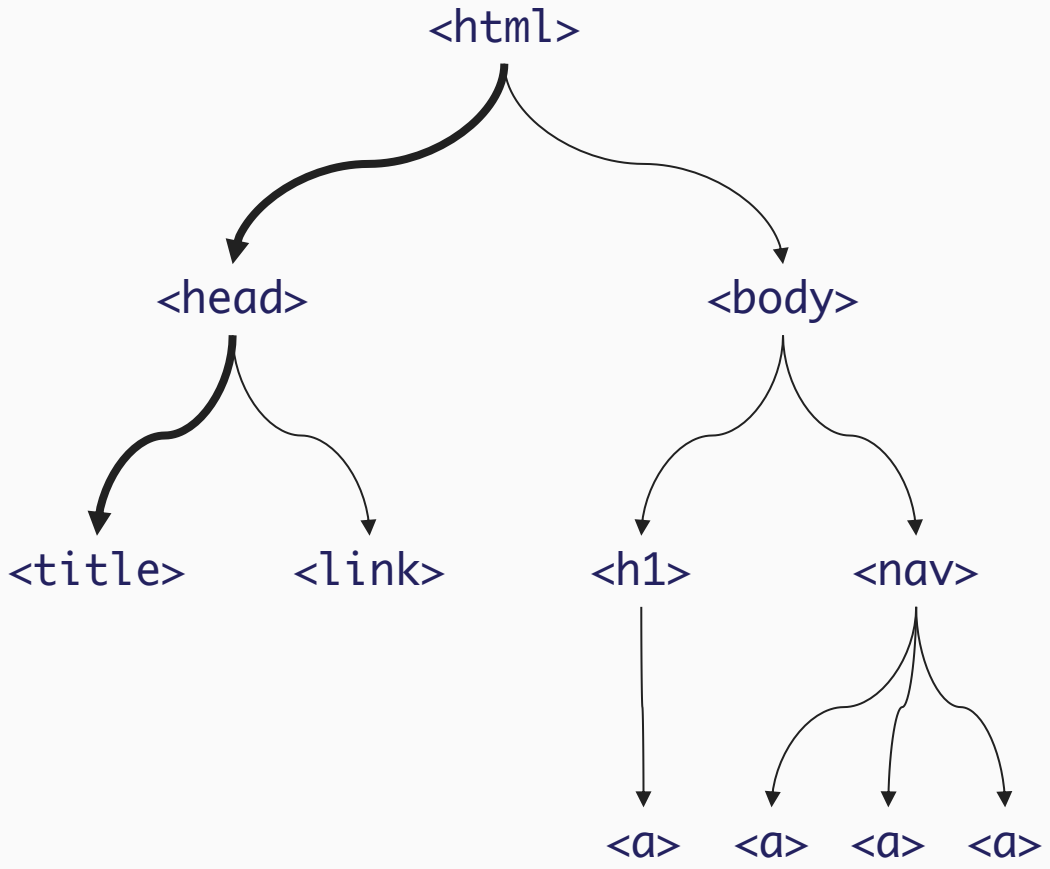
Declaración -----▸ color : "darkred" ;

Propiedad Valor

}

Hoja de Estilos en Cascada

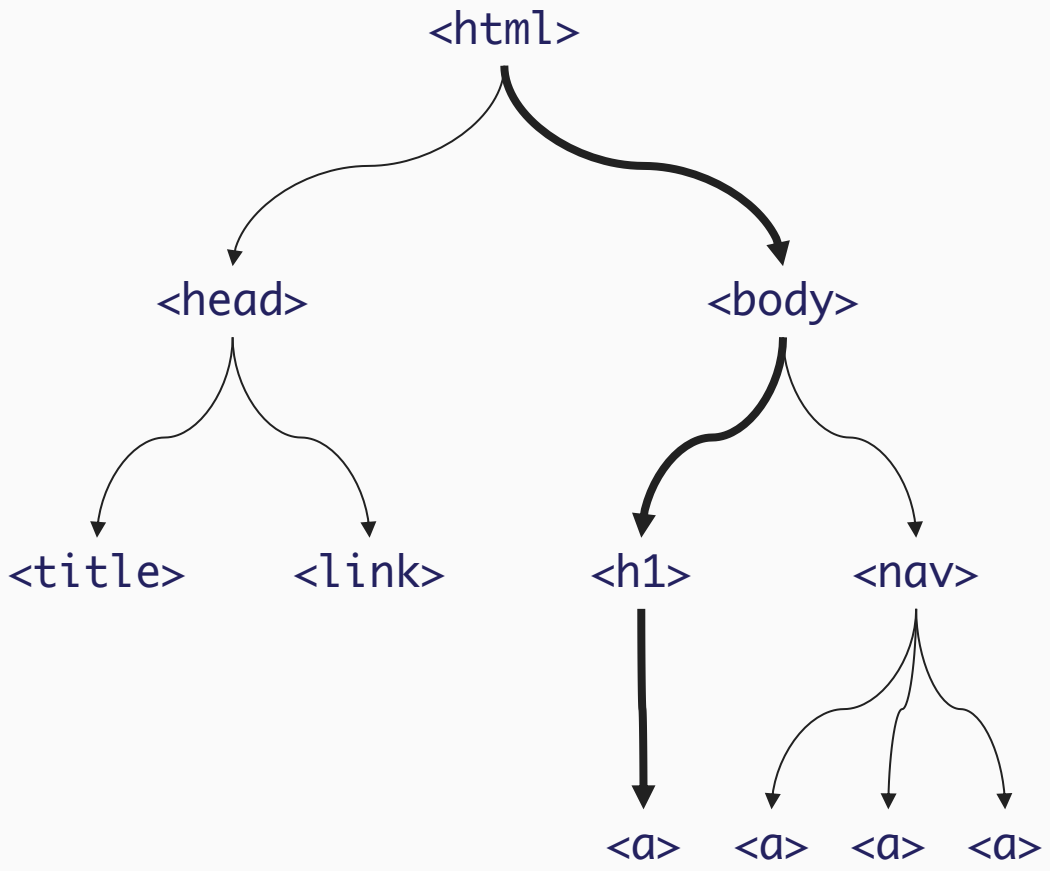
Cascading Style Sheets (CSS)




```
css = html head title
```

Hoja de Estilos en Cascada

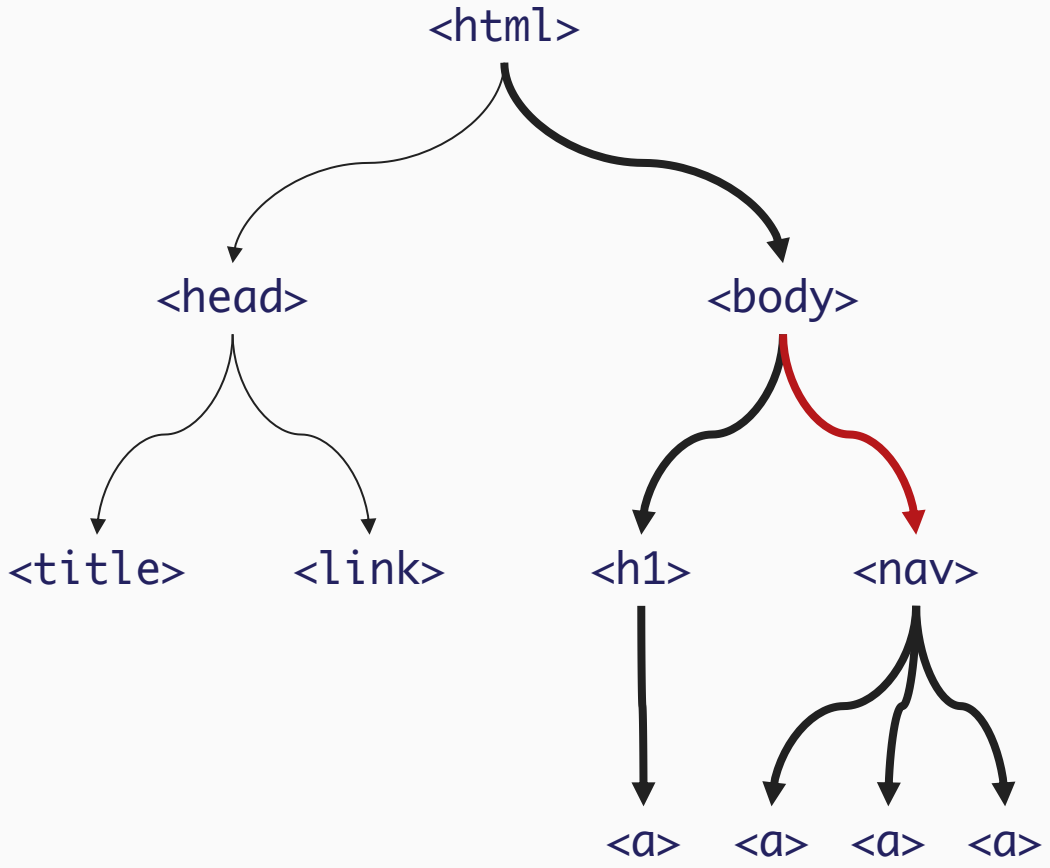
Cascading Style Sheets (CSS)




```
css = html body h1 a
```

Hoja de Estilos en Cascada

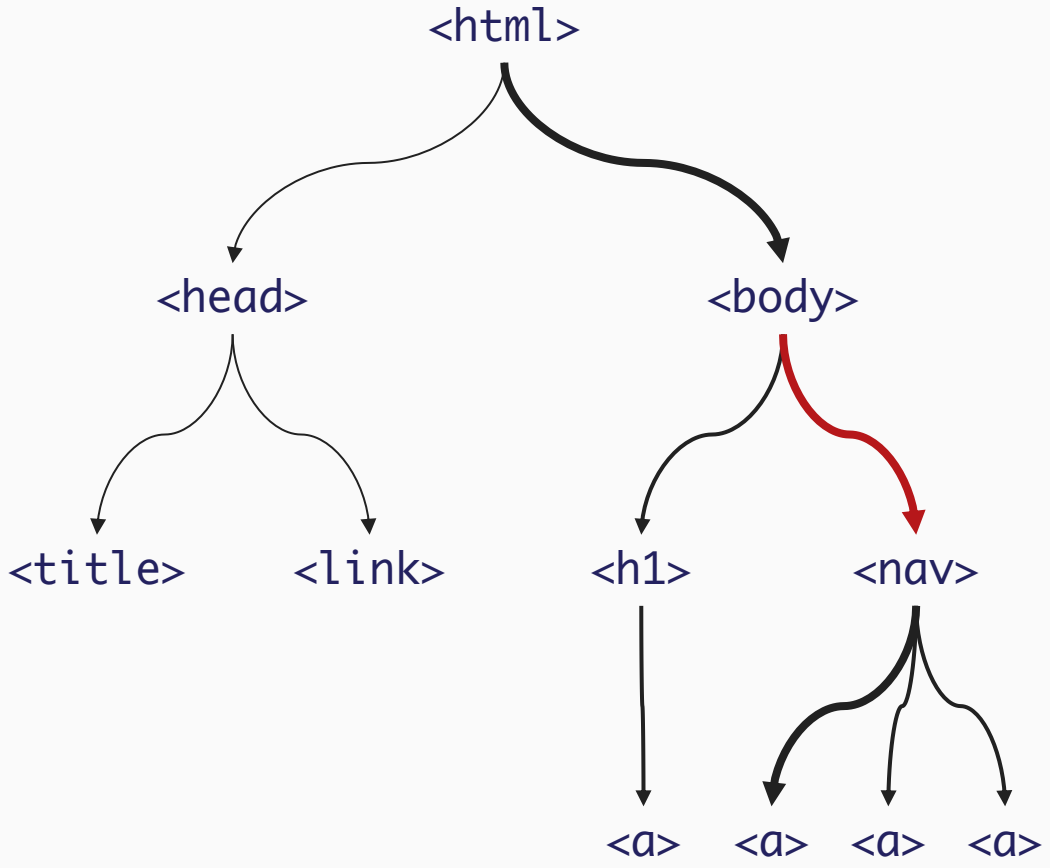
Cascading Style Sheets (CSS)




```
css = html body * a
```

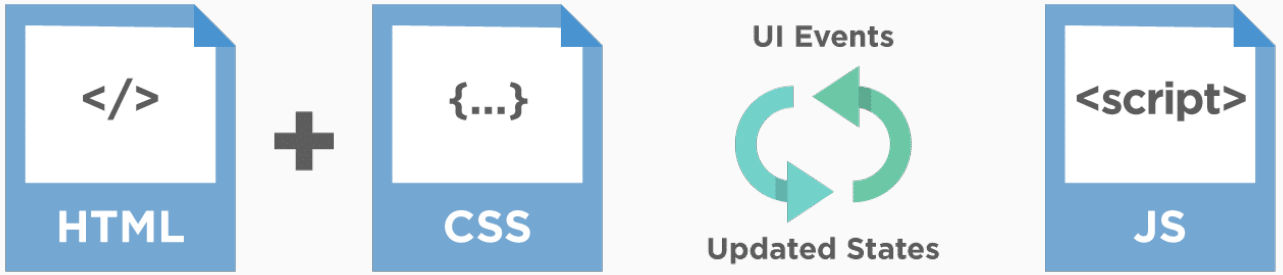
Hoja de Estilos en Cascada

Cascading Style Sheets (CSS)



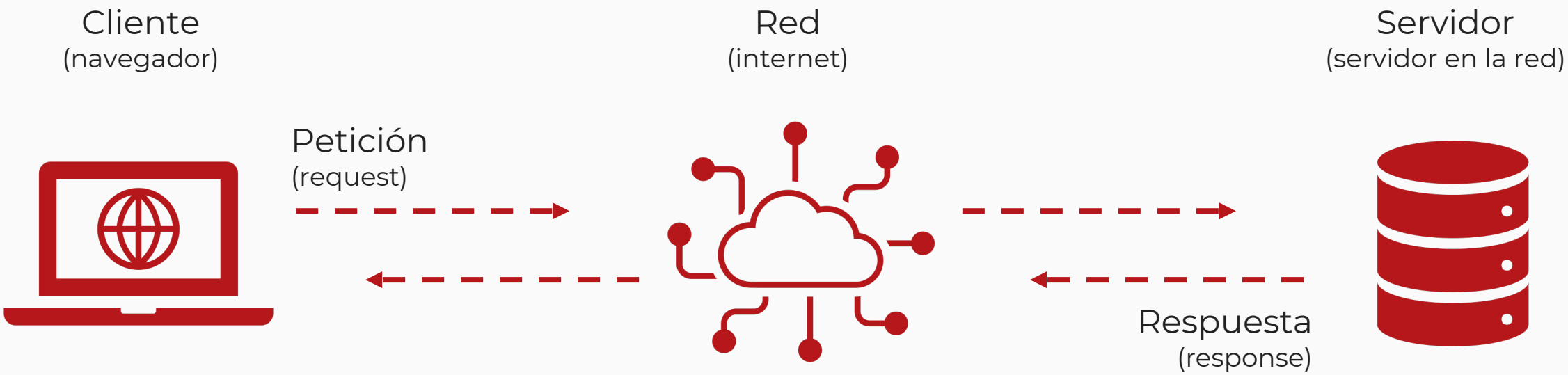

```
css = html body a:nth-child(2)
```

JavaScript



Petición - Respuesta

Request Response



Peticiónes en HTTP

HTTP Request

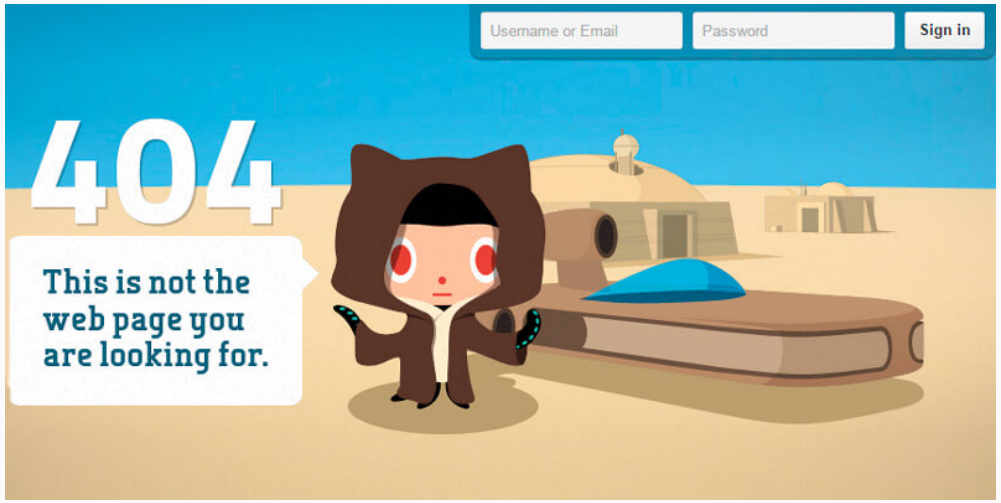
- **GET :** solicita una representación del recurso especificado.
- **HEAD:** solicita una respuesta idéntica a la de una solicitud GET, pero sin el cuerpo de la respuesta.
- **POST:** se utiliza para enviar una entidad al recurso especificado, lo que a menudo provoca un cambio de estado o efectos secundarios en el servidor.
- **PUT:** reemplaza todas las representaciones actuales del recurso de destino con la carga útil de la solicitud.
- **DELETE:** elimina el recurso especificado.
- **CONNECT:** establece un túnel al servidor identificado por el recurso de destino.
- **OPTIONS:** se utiliza para describir las opciones de comunicación para el recurso de destino.
- **TRACE:** realiza una prueba de bucle de mensajes a lo largo de la ruta al recurso de destino.
- **PATCH:** se utiliza para aplicar modificaciones parciales a un recurso.

Respuestas en HTTP

HTTP Response

Los códigos de estado de respuesta HTTP indican si una solicitud HTTP específica se ha completado correctamente. Las respuestas se agrupan en cinco clases:

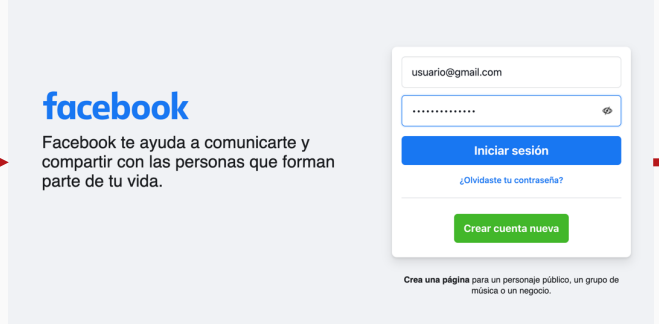
- Respuestas informativas (100-199)
- Respuestas satisfactorias (200-299)
- Redirecciones (300–399)
- Errores del cliente (400–499)
- Errores del servidor (500–599)



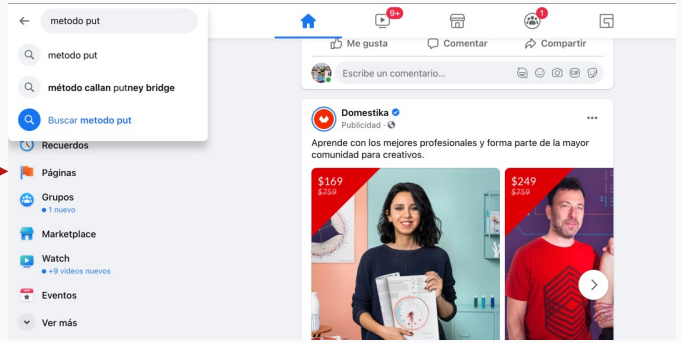
Navegación Web

GET

POST



PUT

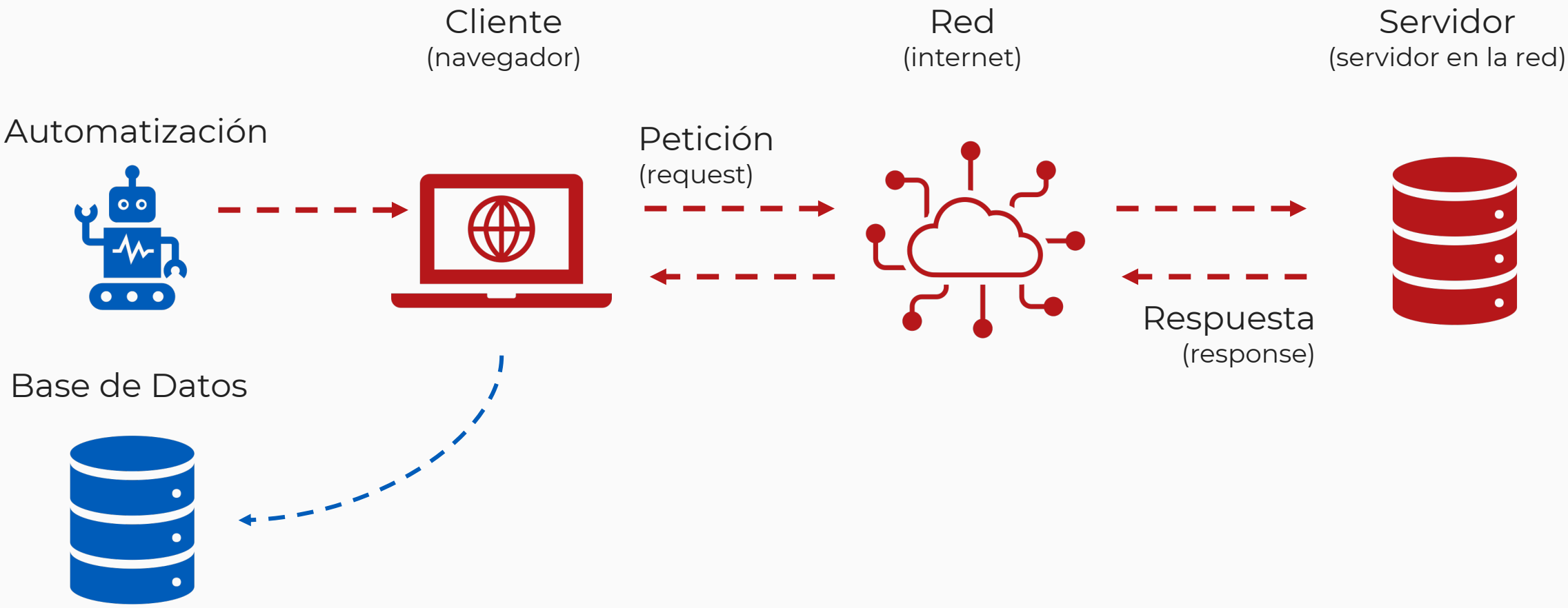


COPY

PASTE

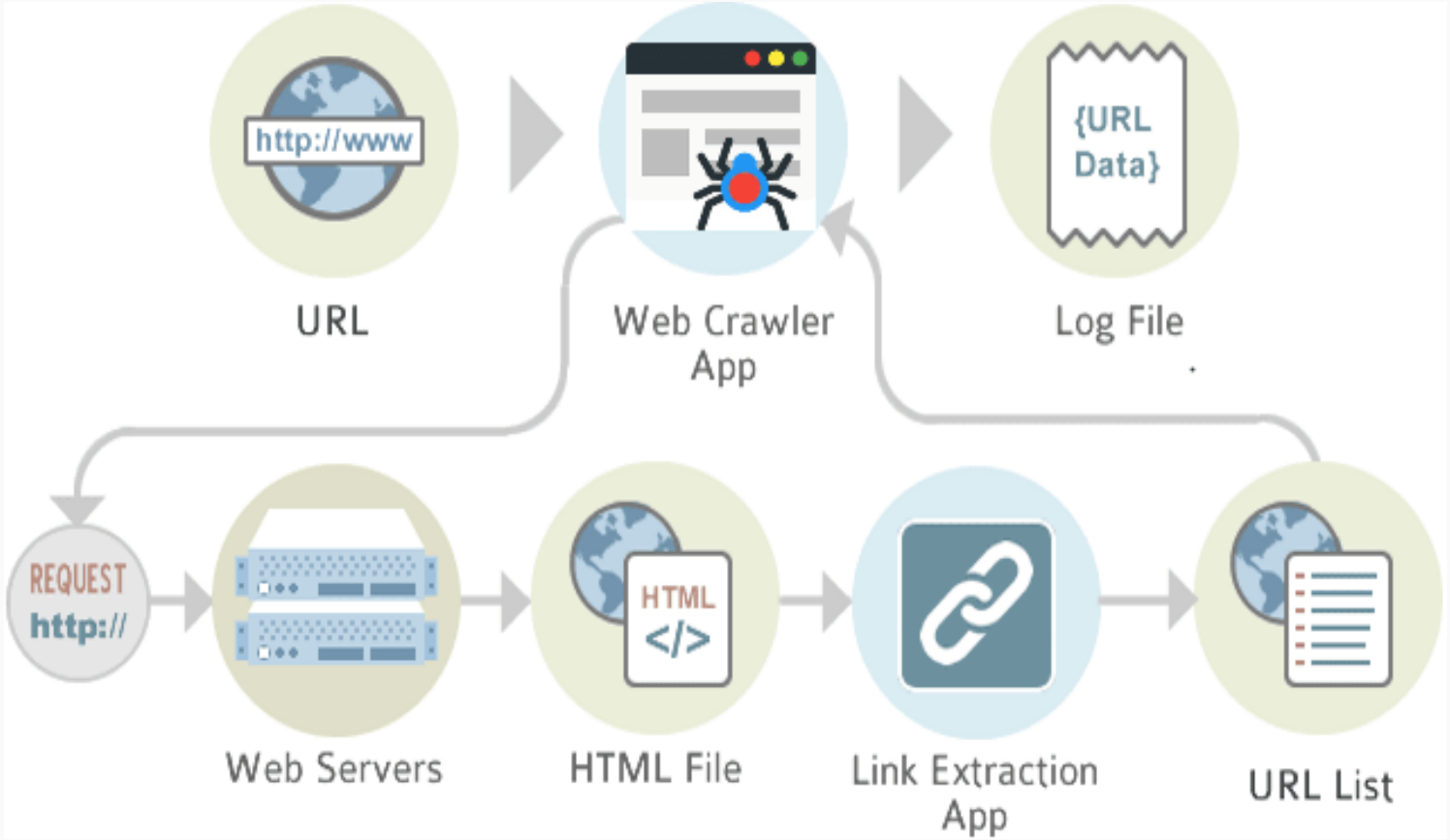
Raspado Web

Web Scrapping



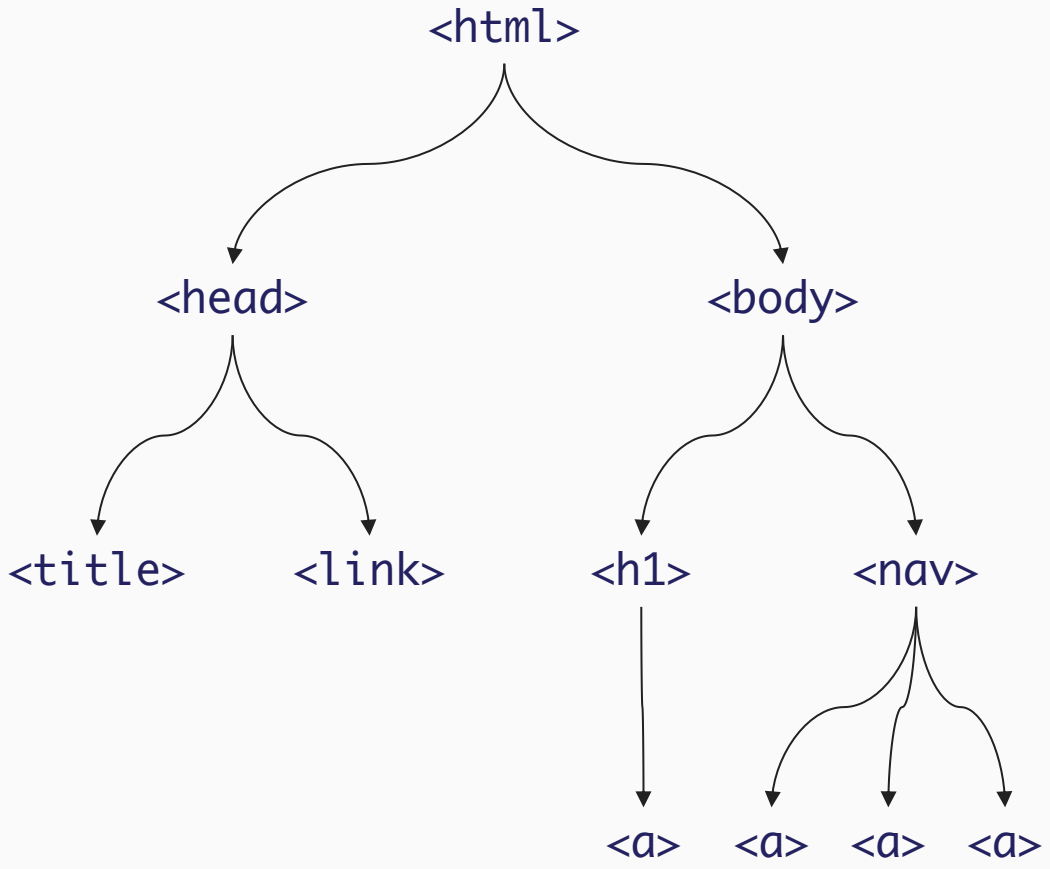
Rastreo Web

Web Crawling



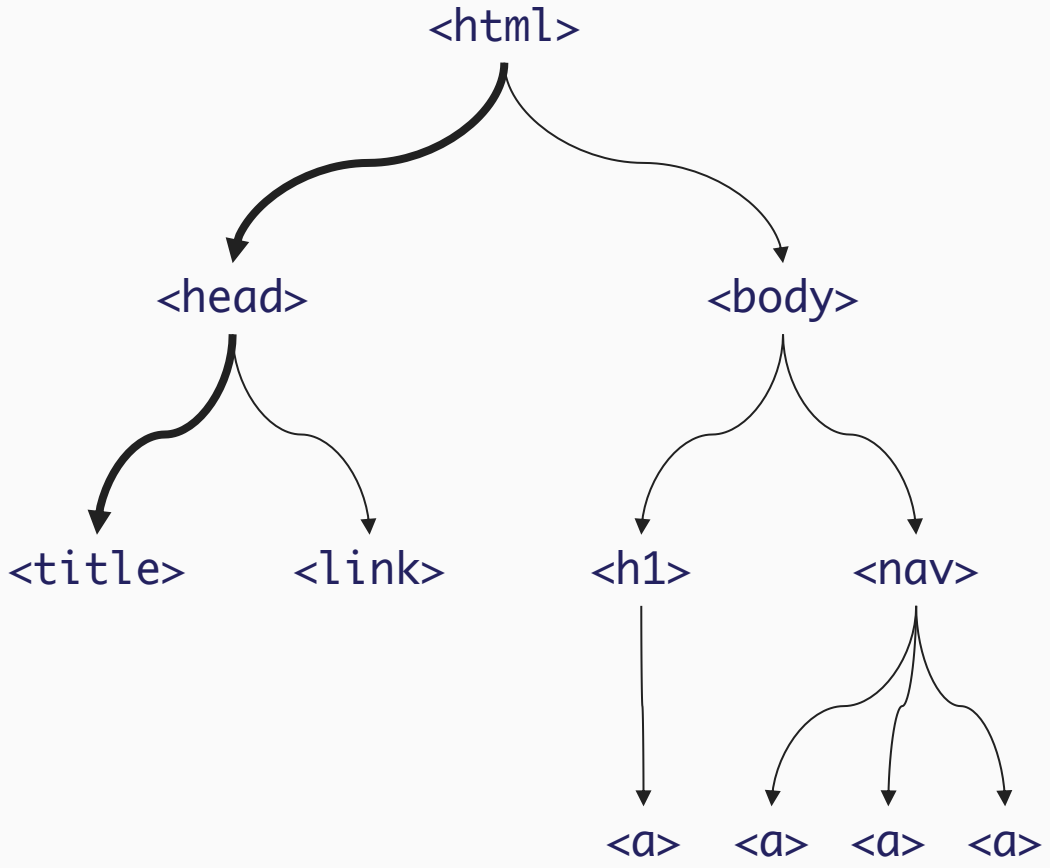
Lenguaje de Ruta XML

XML Path Languaje (XPATH)



Lenguaje de Ruta XML

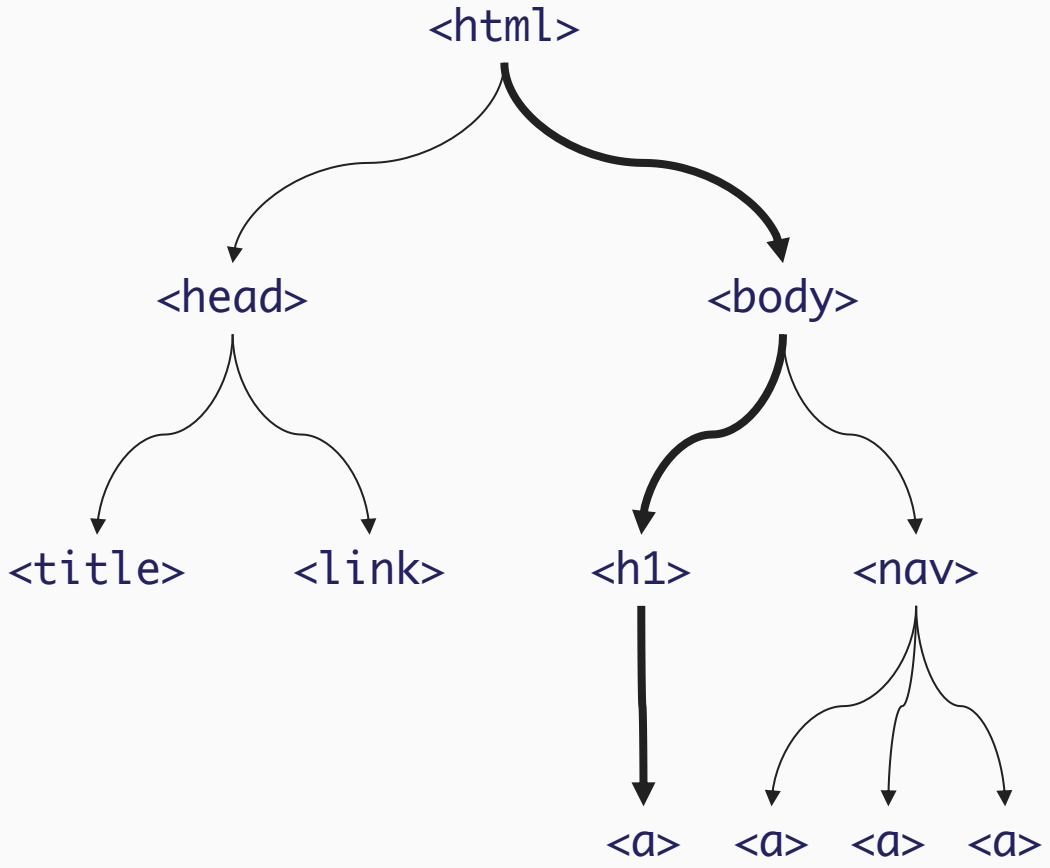
XML Path Languaje (XPATH)




```
xpath = html/head/title
```

Lenguaje de Ruta XML

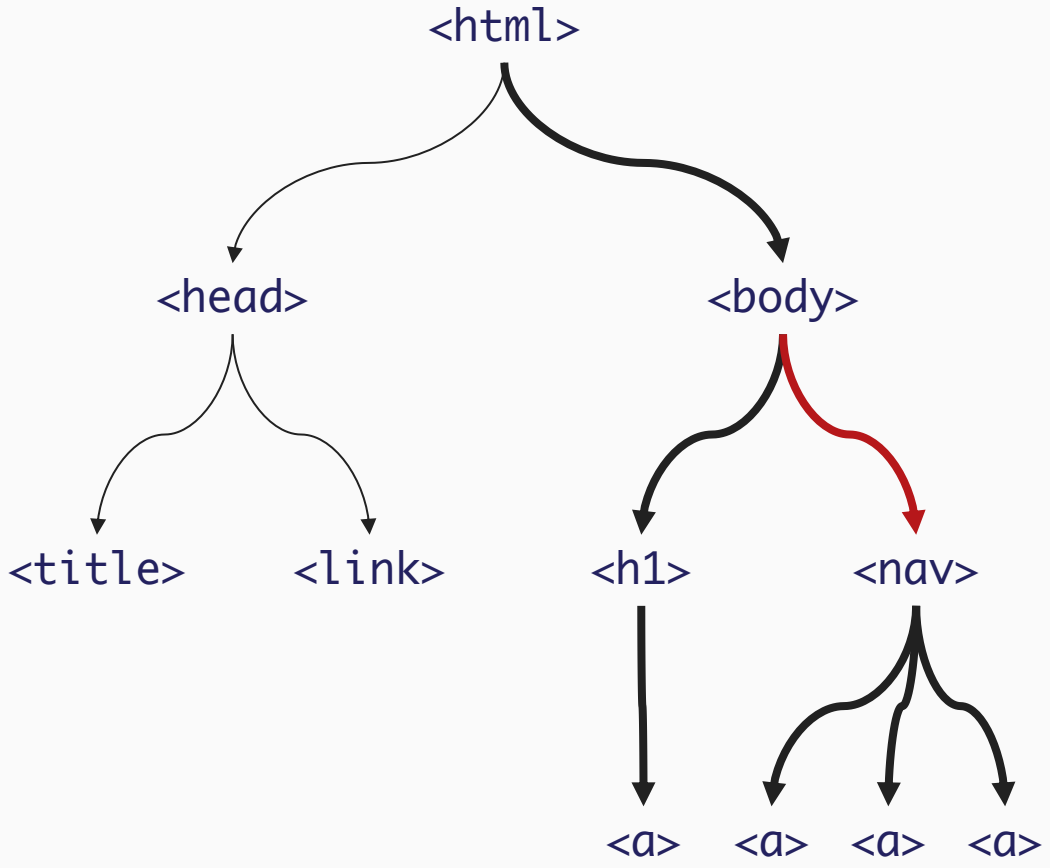
XML Path Languaje (XPATH)




```
xpath = html/body/h1/a
```

Lenguaje de Ruta XML

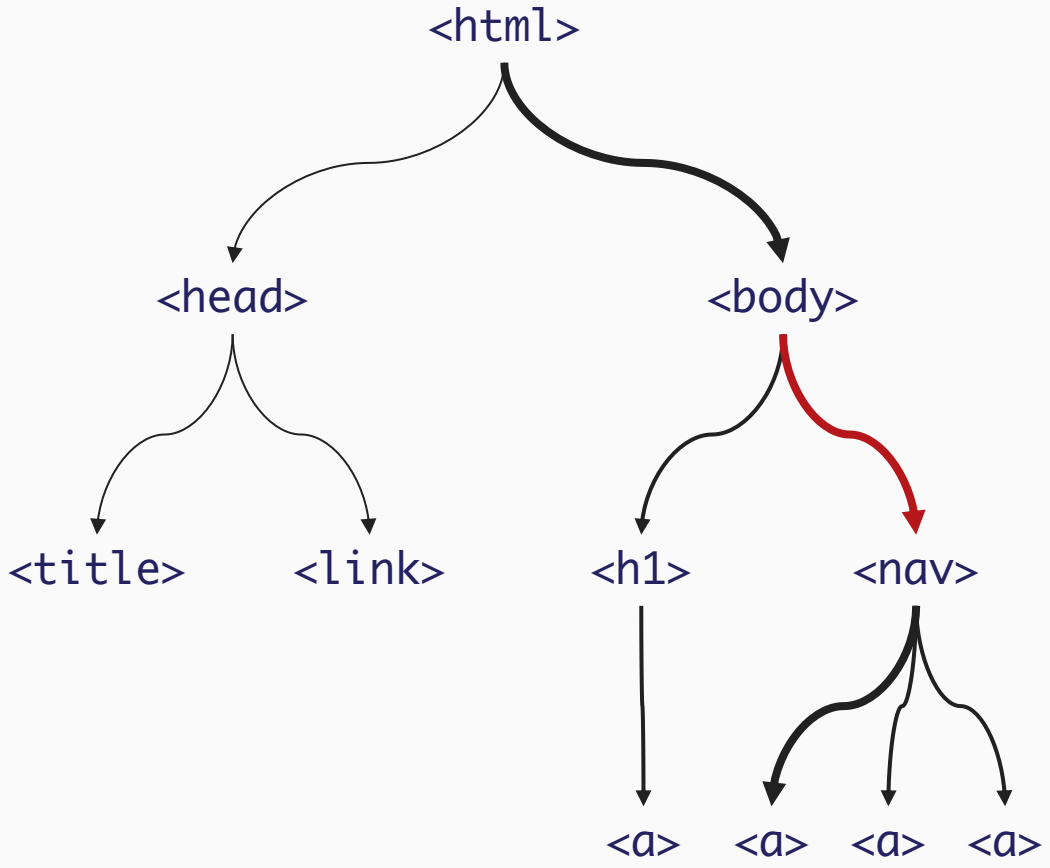
XML Path Languaje (XPATH)




```
xpath = html/body//a
```

Lenguaje de Ruta XML

XML Path Languaje (XPATH)




```
xpath = html/body//a[2]
```

¡Gracias por tu atención!

Escríbeme: r.rosado@tec.mx