



Data WebScraping

Python Fundamentals

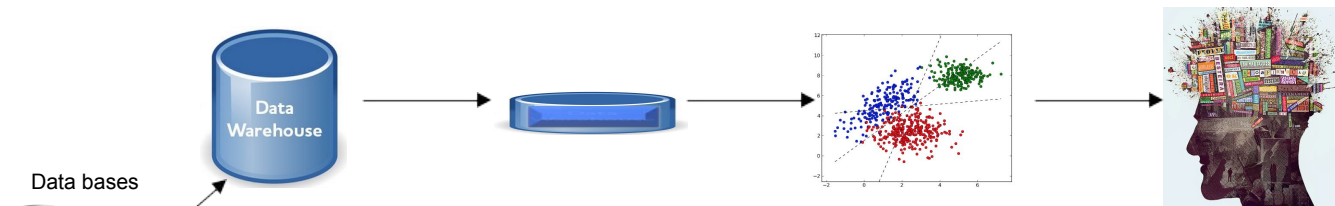
PENSAMIENTO COMPUTACIONAL Y
PROGRAMACIÓN
COLEGIO LA GIROUETTE

\begin{document}

Darío Creado Figueroa

Fuente : Belén Saldías. / IIC2026

Contenidos del curso: *Data Mining Stages*



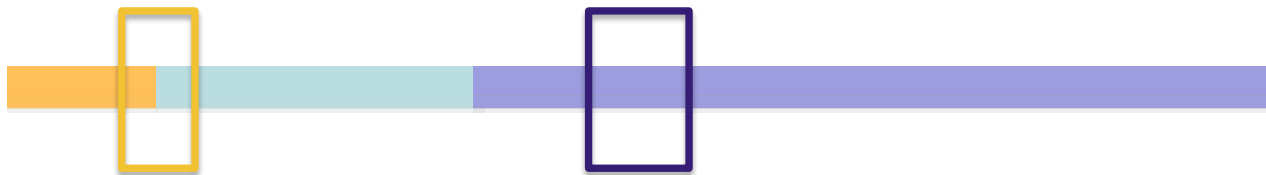
1. Data Integration

2. Data transformation
and selection

3. Data Mining

4. Evaluation and
visualization

Clases



% del curso

Mining the web!

- ¡Una de las **bases de datos** de texto no estructurado **más grandes del mundo!**
- Aplicación de minería de datos para **descubrir patrones en la Web**
- ***Clicking for gold*** → obtener ganancias utilizando los datos en la Web
- No siempre tenemos una API a la cual conectarnos

Source: <https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf>

Datos en la Web - HTML5



```
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <title>Example</title>
5     <link rel="stylesheet" href="style.css">
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>
```

Tag de apertura

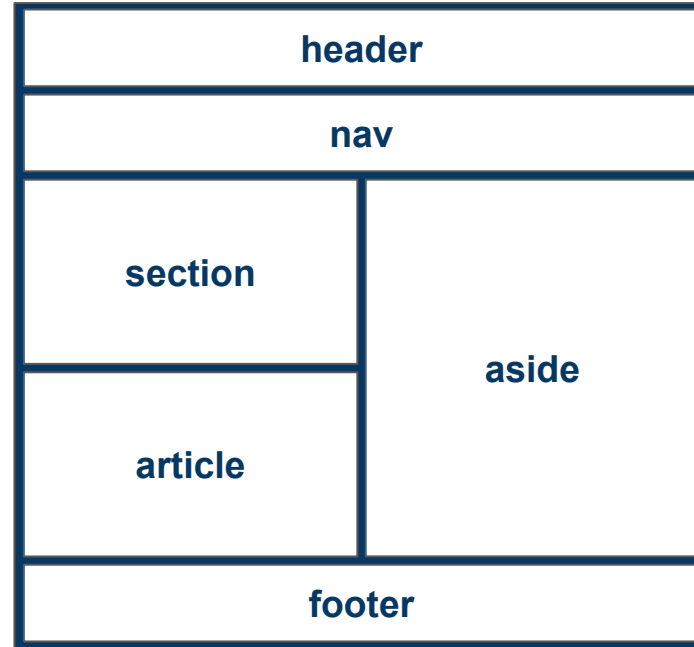
Tag de cierre

Elemento

Elementos semánticos de HTML5



- `<header> </header>`
- `<nav> </nav>`
- `<section> </section>`
- `<article> </article>`
- `<aside> </aside>`
- `<figcaption> </figcaption>`
- `<figure> </figure>`
- `<footer> </footer>`



Source: <https://es.slideshare.net/niharikagupta54966/html5ppt-25757426>



Elementos a nivel de bloque

Estos elemento siempre comienzan en una nueva línea. Toman el ancho máximo disponible.

- `<div> </div>`
- `<h1> </h1>`
- ...
- `<h6> </h6>`
- `<p> </p>`
- `<form> </form>`

```
<!DOCTYPE html>
<html>
<body>

<div style="background-color:orange;color:white;padding:20px;">
  <h2>IIC2433 - Minería de datos</h2>
  <h4>Elementos a nivel de bloque</h4>
  <p>Estos elemento siempre comienzan en una nueva línea. Toman el ancho máximo disponible.</p>
  <p>Se cortan automáticamente en líneas para ajustarse al ancho de la pantalla. Dependiendo del
dispositivo es posible tomar cualquier tamaño máximo de ancho.</p>
</div>

</body>
</html>
```

IIC2433 - Minería de datos

Elementos a nivel de bloque

Estos elemento siempre comienzan en una nueva línea. Toman el ancho máximo disponible.

Se cortan automáticamente en líneas para ajustarse al ancho de la pantalla. Dependiendo del dispositivo es posible tomar cualquier tamaño máximo de ancho.



Elementos en línea

No comienzan en una línea nueva. Toman solo el ancho necesario.

- ` `
- `<a> `
- ` `

```
<!DOCTYPE html>
<html>
<body>

<h1>IIC2433 - <span style="color:orange">Minería de Datos</span> - 2017/2</h1>

</body>
</html>
```

IIC2433 - Minería de Datos - 2017/2

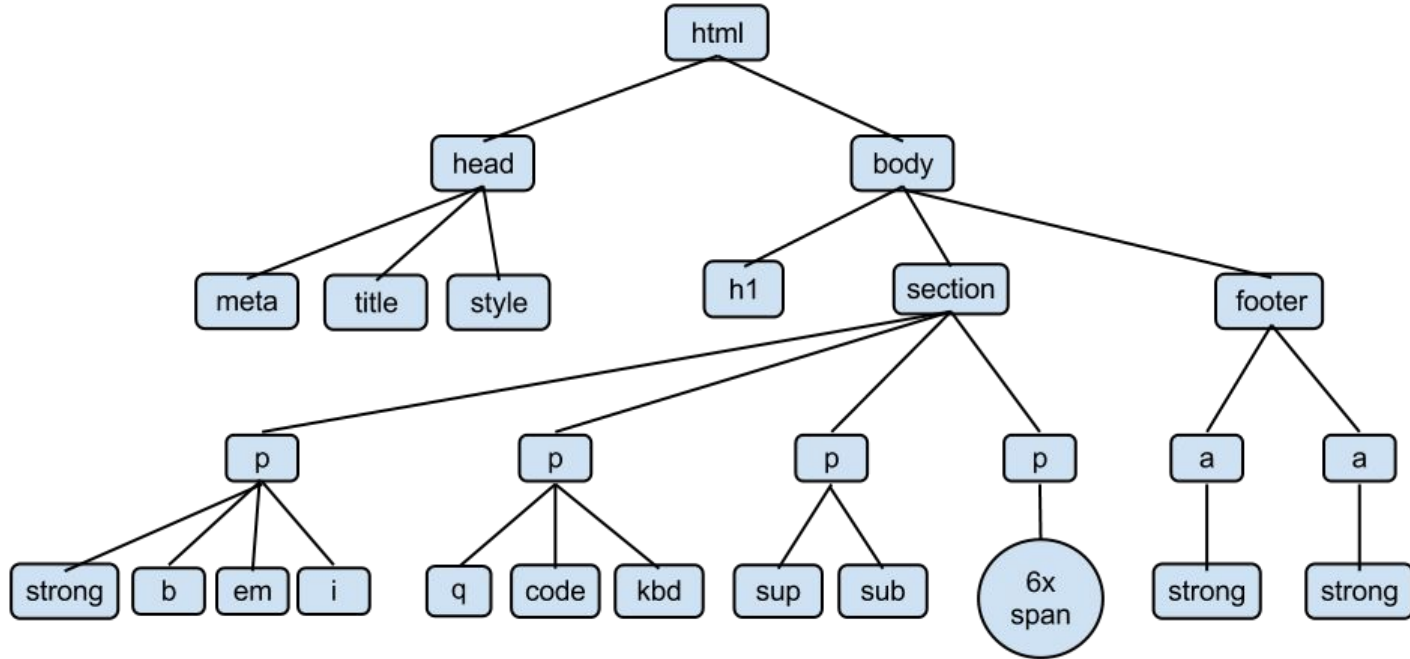
Web Scraping

- Técnica utilizada mediante programas de software para extraer información de sitios web.
- Usualmente se simula la navegación de un humano → **bot**.
- Se puede utilizar un protocolo HTTP para extraer los datos, o navegar y esperar que se cargue todo el sitio.

1. Beautiful soup

- ¿Qué es?
 - Librería para convertir documentos HTML en objetos
 - **Crea un árbol** con la estructura del HTML
 - Permite **extraer** información del árbol y consultarlo por **tags**

1. Beautiful soup



1. BeautifulSoup

```
# paquete que nos permite conectarnos (hacer un request) a un sitio web  
import requests
```

```
# paquete para manipular el código html de una página web  
from bs4 import BeautifulSoup
```

```
# paquete para manipular datos eficientemente  
import pandas as pd
```

```
soup = BeautifulSoup(content, "html.parser")
```

- `content` puede ser un archivo HTML local o consultado en la Web

1. Beautiful soup - comandos básicos

```
<html>
<head>
<title>
  The Dormouse's story
</title>
</head>

<body>
<p class="title">
  <b>
    The Dormouse's story
  </b>
</p>
<p class="story">
  Once upon a time there were three little sisters; and their names were
  <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
  </a>
  ,
  <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
  </a>
  ....
</p>
<p class="story">
  ...
</p>
</body>
</html>
```

```
> soup.title
<title>The Dormouse's story</title>
```

```
> soup.title.string
u'The Dormouse's story'
```

```
> soup.title.parent.name
u'head'
```

```
> soup.p
<p class="title"><b>The Dormouse's
story</b></p>
```

```
> soup.a
<a class="sister" href="http://example.com/elsie"
id="link1">Elsie</a>
```

```
> soup.find_all('a')
[<a class="sister" href="http://example.com/elsie"
id="link1">Elsie</a>,
  <a class="sister" href="http://example.com/lacie"
id="link2">Lacie</a>,
  <a class="sister" href="http://example.com/tillie"
id="link3">Tillie</a>]
```

```
> soup.find(id="link3")
<a class="sister" href="http://example.com/tillie"
id="link3">Tillie</a>
```

```
> for link in soup.find_all('a'):
>     print(link.get('href'))
http://example.com/elsie
http://example.com/lacie
http://example.com/tillie
```

2. Selenium

- ¿Qué es?
 - Automatiza navegadores.
 - Simula el comportamiento de un ser humano.
 - Generalmente usado para testing de sitios web.
 - ¡Puede ser usado para lo que sea! → **uso responsable**.