

Taller de Web Scrapping

Adrián Soto

adrian.soto@ing.puc.cl - @alanezz

Sobre el profesor

- Alumno de doctorado
- Profesor de IIC2413 - Bases de Datos
- Desarrollador Web

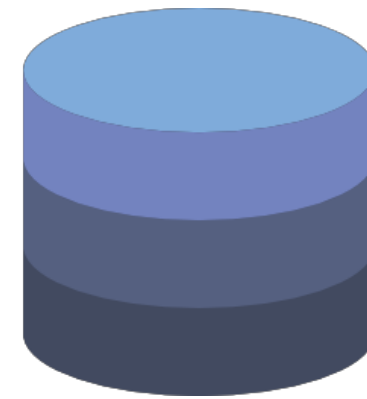
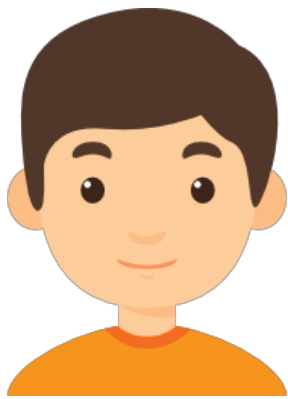
¿Por qué Web Scraping?



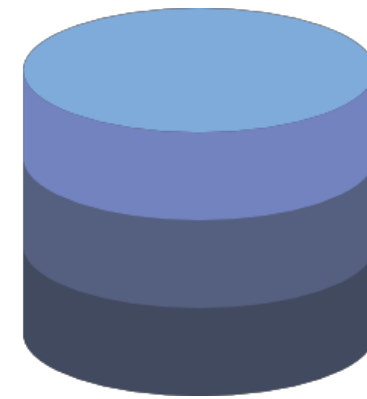
¿Por qué Web Scrapping?



¿Por qué Web Scraping?



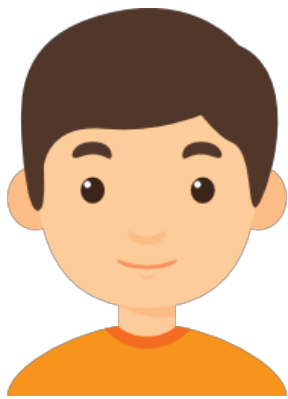
¿Por qué Web Scraping?



¿Por qué Web Scrapping?



¿Por qué Web Scrapping?



¿Por qué Web Scrapping?



¿Por qué Web Scrapping?



Web Scraping

Web Scraping es extraer datos desde el código HTML de algún sitio web

Web Scraping

Usos

Usamos Web Scraping cuando:

- Necesitamos extraer datos de manera automática
- No tenemos acceso directo a una base de datos
- Necesitamos simular el comportamiento de un usuario

Web Scrapping

Preliminares

A lo largo de esta clase vamos a necesitar:

- Python
- HTML
- BeautifulSoup (librería de Python)
- Google Colab: (<https://colab.research.google.com/>)

HTML

Es el *markup language* para crear páginas y aplicaciones web

Los documentos HTML tienen estructura de árbol

En general, se complementa con CSS y JavaScript

HTML

Ejemplo

[EMOL](#) [EL MERCURIO](#) [BLOGS](#) [LEGAL](#) [CAMPO](#) [INVERSIONES](#) [AUTOS](#) [PROPIEDADES](#) [EMPLEOS](#) [ECONÓMICOS.CL](#) [AUTOS- CASAS](#) [LA SEGUNDA](#) [LUN](#)


Ingresar | Registrarse

Santiago: Jueves 10 de enero del 2019 | Actualizado 22:57

[e.](#)

[Noticias](#)

[Economía](#)

[Deportes](#)

[Espectáculos](#)

[Tendencias](#)

[Autos](#)

[Servicios](#)

[360°](#)

[Chile](#) | [Mundo](#) | [Tecnología](#) | [Educación](#) | [Documentos](#) | [Multimedia](#) |

[amarillas.com](#)



DEPORTES



Bomba en el tenis mundial: Andy Murray anuncia su retiro para este año debido a las lesiones

22:17 | El británico señaló que espera jugar en Wimbledon su último torneo, aunque recalcó que probablemente no podrá competir más allá del Abierto de Australia que arranca esta semana.

 13

Piñera conversa con líder de Asamblea Nacional venezolana y entrega su absoluto respaldo

21:26 | El Mandatario, que aseguró que el Gobierno de Chile no reconoce el nuevo régimen de Maduro, solidarizó con Juan Guaidó y le expresó su admiración "por la lucha que están dando por la democracia".

94

327

4

221

28

93

60















OEA y EE.UU. declaran la ilegitimidad del nuevo Gobierno de Nicolás Maduro

EL COMENTARISTA OPINA

La moda de lo desechable

5 0

 **Aoryu**

AHORA SE DEBATE



Piñera no incluyó a Carabineros en conmemoración del conflicto del Beagle. ¿Qué piensas?

563



```
<div id="noticias_caja_texto">
  <div class="contenedor-titulo">
    <a href="/noticias/Deportes/2019/01/10/933914/Bomba-en-el-
    tenis-mundial-Andy-Murray-anuncia-su-retiro-para-este-ano-debido-a-las-
    lesiones.html" id="ucHomePage_cuNoticiaDeporte_titular">Bomba en el
    tenis mundial: Andy Murray anuncia su retiro para este año debido a las
    lesiones</a>
  </div>
  <p id="ucHomePage_cuNoticiaDeporte_bajada"><span
  class="color_hora2008">22:17 | </span>El británico señaló que espera
  jugar en Wimbledon su último torneo, aunque recalcó que probablemente
  no podrá competir más allá del Abierto de Australia que arranca esta
  semana. </p>
  <div id="ucHomePage_cuNoticiaDeporte_ContadorComentarios"
  class="cont_contador_comentarios" data-id="933914">
    <a href="/noticias/Deportes/2019/01/10/933914/Bomba-en-el-
    tenis-mundial-Andy-Murray-anuncia-su-retiro-para-este-ano-debido-a-las-
    lesiones.html#comentarios"
    id="ucHomePage_cuNoticiaDeporte_LinkContadorComentarios">
      <span class="fb_comments_count"></span>
    </a>
  </div>
</div>
```

HTML

Ejemplo

[EMOL](#) [EL MERCURIO](#) [BLOGS](#) [LEGAL](#) [CAMPO](#) [INVERSIONES](#) [AUTOS](#) [PROPIEDADES](#) [EMPLEOS](#) [ECONÓMICOS.CL](#) [AUTOS- CASAS](#) [LA SEGUNDA](#) [LUN](#)


Ingresar | Registrarse

Santiago: Jueves 10 de enero del 2019 | Actualizado 22:57

[e.](#)

[Noticias](#)

[Economía](#)

[Deportes](#)

[Espectáculos](#)

[Tendencias](#)

[Autos](#)

[Servicios](#)

[360°](#)

[Chile](#) | [Mundo](#) | [Tecnología](#) | [Educación](#) | [Documentos](#) | [Multimedia](#) |

[amarillas.com](#)

NEW CITROËN BERLINGO



COMO TÚ,
PUEDE HACER DE TODO.

DESDE:
\$11.990.000
+ IVA

COTIZA **AQUÍ**

DEPORTES



Bomba en el tenis mundial: Andy Murray anuncia su retiro para este año debido a las lesiones
22:17 | El británico señaló que espera jugar en Wimbledon su último torneo, aunque recalcó que probablemente no podrá competir más allá del Abierto de Australia que arranca esta semana.

Piñera conversa con líder de Asamblea Nacional venezolana y entrega su absoluto respaldo
21:26 | El Mandatario, que aseguró que el Gobierno de Chile no reconoce el nuevo régimen de Maduro, solidarizó con Juan Guaidó y le expresó su admiración "por la lucha que están dando por la democracia".

94

327

4

221

28

93

60

Gobierno no reconoce nuevo régimen de Maduro: "Llega al poder de forma ilegítima"

Chile Vamos pide a Piñera que reconozca a AN de Venezuela como único órgano legítimo

Maduro se compromete a seguir con el mandato de Chávez

Perú llama a consultas a su encargada de negocios

Paraguay rompe relaciones diplomáticas con Venezuela

Duque llama a "cercar diplomáticamente" a Venezuela



OEA y EE.UU. declaran la ilegitimidad del nuevo Gobierno de Nicolás Maduro



563

EL COMENTARISTA OPINA

La moda de lo desechable

5 0

Aoryu

AHORA SE DEBATE



Piñera no incluyó a Carabineros en conmemoración del conflicto del Beagle. ¿Qué piensas?

```
<div id="noticias_caja_texto">
  <div class="contenedor-titulo">
    <a href="/noticias/Deportes/2019/01/10/933914/Bomba-en-el-
    tenis-mundial-Andy-Murray-anuncia-su-retiro-para-este-ano-debido-a-las-
    lesiones.html" id="ucHomePage_cuNoticiaDeporte_titular">Bomba en el
    tenis mundial: Andy Murray anuncia su retiro para este año debido a las
    lesiones</a>
  </div>
  <p id="ucHomePage_cuNoticiaDeporte_bajada"><span
  class="color_hora2008">22:17 | </span>El británico señaló que espera
  jugar en Wimbledon su último torneo, aunque recalcó que probablemente
  no podrá competir más allá del Abierto de Australia que arranca esta
  semana. </p>
  <div id="ucHomePage_cuNoticiaDeporte_ContadorComentarios"
  class="cont_contador_comentarios" data-id="933914">
    <a href="/noticias/Deportes/2019/01/10/933914/Bomba-en-el-
    tenis-mundial-Andy-Murray-anuncia-su-retiro-para-este-ano-debido-a-las-
    lesiones.html#comentarios"
    id="ucHomePage_cuNoticiaDeporte_LinkContadorComentarios">
      <span class="fb_comments_count"></span>
    </a>
  </div>
</div>
```


Tags HTML

- html
- h1, h2, h3, ...
- p
- ul
- div, span
- table
- a
- img
- ...

Atributos en HTML

- class
- id
- href
- ...

HTML

Para aprender HTML en detalle puedes ingresar a:

<https://www.w3schools.com/html/default.asp>

Ejemplos HTML

Actividad

Escriba un documento HTML que contenga una lista con links a sitios de noticias (al menos 3)

Escriba un documento HTML que contenga al menos 3 fotos de volcanes de Chile, junto con su nombre y una descripción (puede obtenerla desde Wikipedia)

Python

Es un lenguaje de programación ampliamente utilizado

Usaremos este lenguaje para construir un programa que haga Web Scraping

Tutorial Python

Durante esta clase utilizaremos Google Colab para crear
nuestros programas

Web Scraping con Python

Para hacer Web Scraping vamos a utilizar la librería BeautifulSoup

Esta librería permite leer documentos HTML y recorrerlos utilizando Python

Tutorial BS4

Vamos a seguir trabajando con Google Colab para crear
nuestros programas