Taller de Web Scraping

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

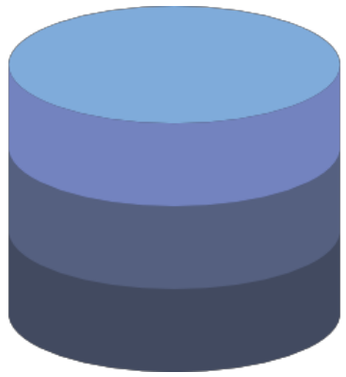
Sobre el profesor

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

¿Por qué Web Scraping?

¿Por qué Web Scraping?

¿Por qué Web Scraping?



¿Por qué Web Scraping?

¿Por qué Web Scraping?

¿Por qué Web Scraping?

¿Por qué Web Scraping?

¿Por qué Web Scraping?

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 Scraping

Preliminares

A lo largo de esta clase vamos a necesitar:

- Python
- HTML
- Beautiful Soup (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


```
    <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


```
    <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 Beautiful Soup

Esta librería permite leer documentos HTML y recorrerlos utilizando Python

Tutorial BS4

Vamos a seguir trabajando con Google Colab para crear nuestros programas