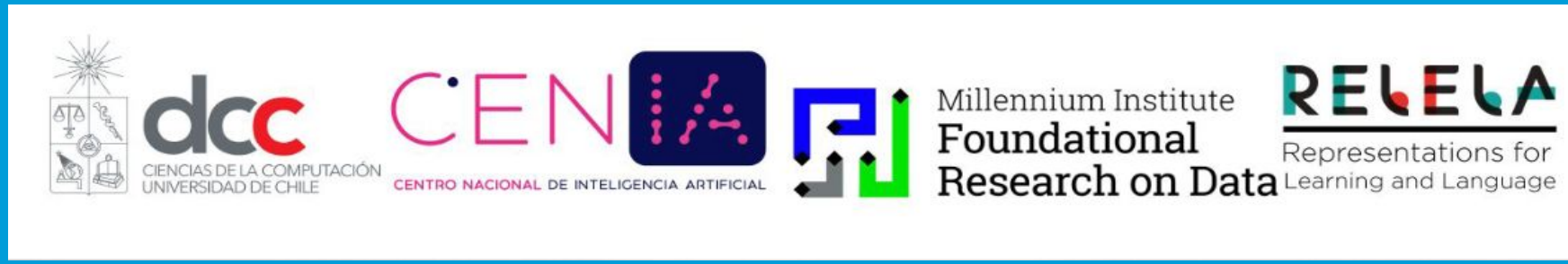
SICSS-Chile: Taller Web Scrapping

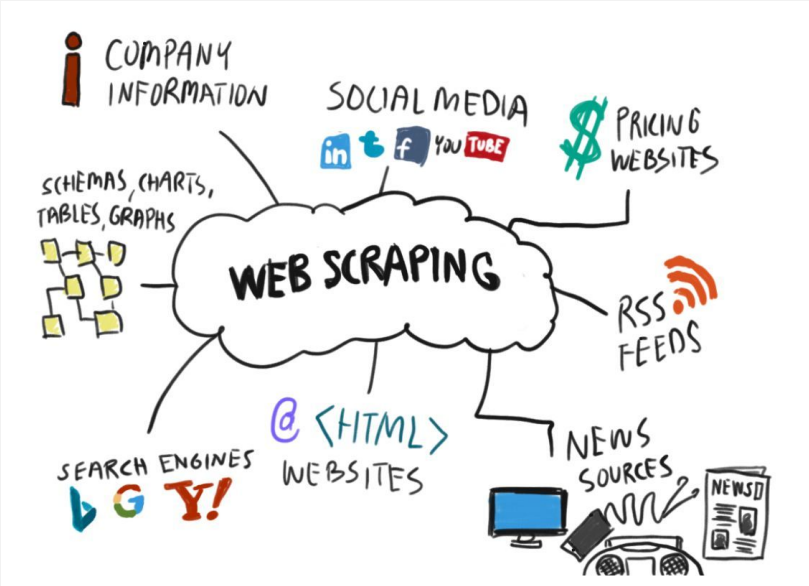
Gabriel Iturra-Bocaz
MSc in Computer Science, Universidad de Chile



¿Qué es Web Scraping?

- El web scraping es una técnica para extraer información de sitios web. Esto se puede hacer de forma manual, pero generalmente es más rápido, eficiente y menos propenso a errores automatizar la tarea.

- Permite adquirir datos no tabulares o mal estructurados de sitios web y convertirlos en un formato utilizable y estructurado, como un archivo .csv o una hoja de cálculo.

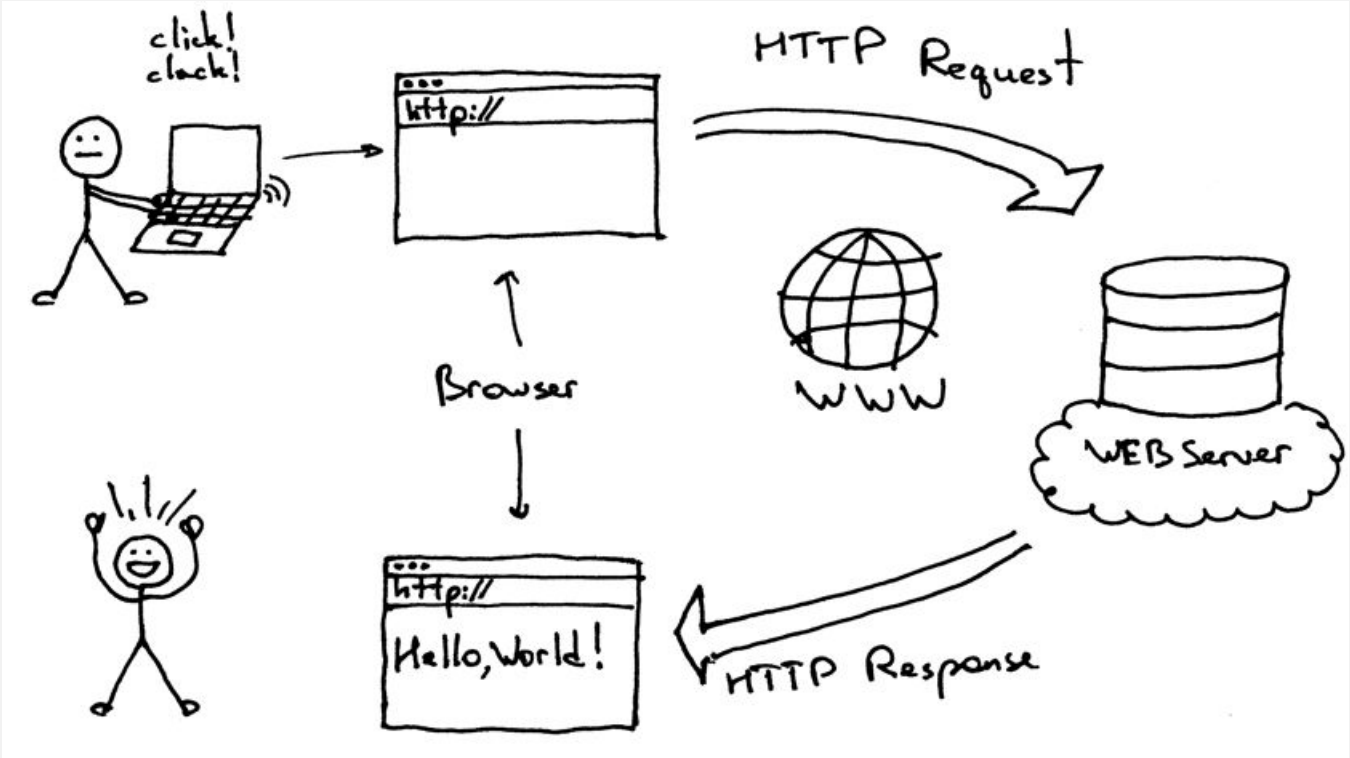


Contenido generado por usuarios en la Web

- ¿Qué tipo de contenido encontramos en la Web?
  - **Texto: html, texto plano, mensajes, post, libros, etc.**
  - Audio: música, podcast.
  - Video: streaming

- ¿Cómo extraer este contenido y convertirlo en información útil?
- ¿Qué podríamos hacer con este contenido?
- ¿Qué dificultades tenemos para acceder a ellos? es posible?

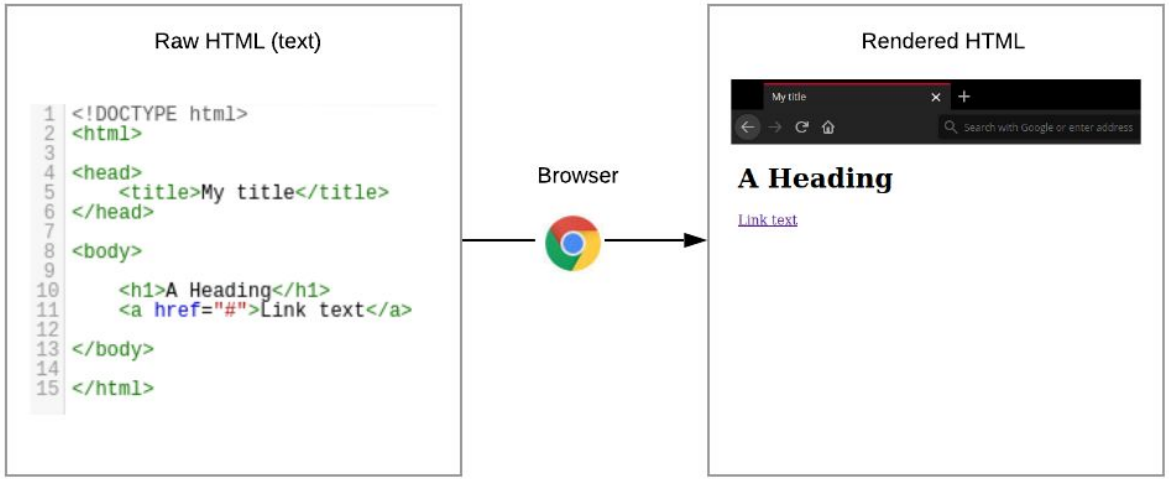
¿Cómo funciona la Web?



¿Qué es HTML?

- HTML significa HyperText Markup Language (Lenguaje de Marcado de Hipertexto).
- Es el lenguaje de marcado estándar para las páginas web que conforman Internet.
- HTML contiene una serie de elementos que conforman una página web y que pueden conectarse con otras páginas web para formar un sitio web.
- Los elementos HTML se representan en etiquetas que le indican al navegador web cómo mostrar el contenido web.

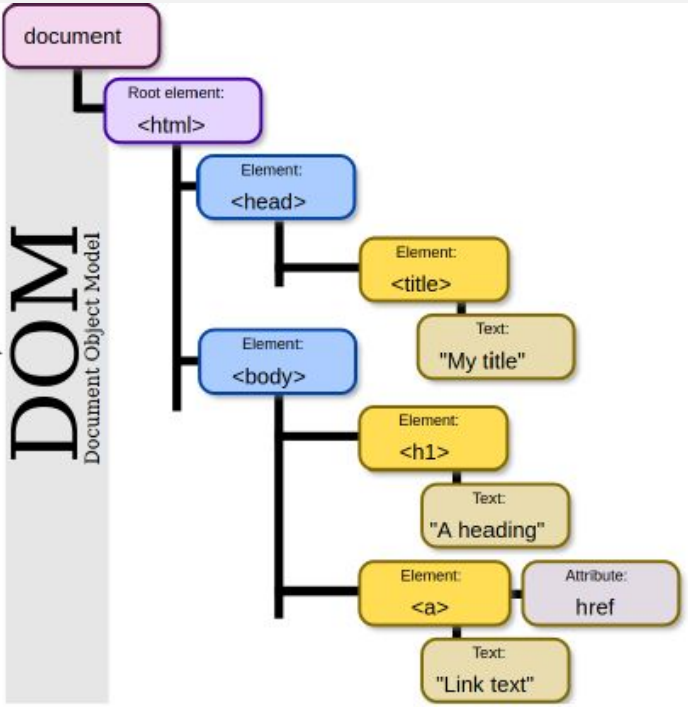
¿Qué es HTML?


Raw HTML (text)
<!DOCTYPE html>
<html>
<head>
<title>My title</title>
</head>
<body>
<h1>A Heading</h1>
Link text
</body>
</html>
Browser
Rendered HTML
My title
Search with Google or enter address
A Heading
Link text

Document Object Model

Raw HTML (text)

```
<!DOCTYPE html>
<html>

<head>
    <title>My title</title>
</head>

<body>

    <h1>A Heading</h1>
    <a href="#">Link text</a>

</body>

</html>
```



Web Crawling

Proceso en el cual un “robot” web navega sistemáticamente a través de internet para inspeccionar páginas web de forma metódica y automatizada.

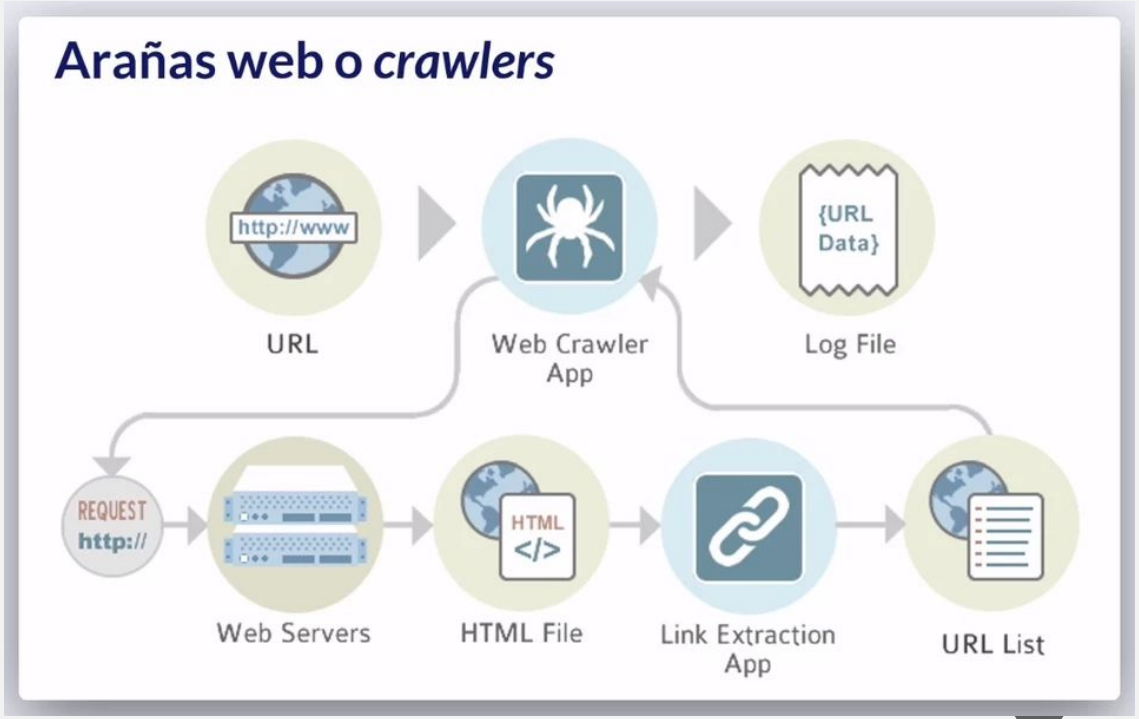
Supuesto: la web está bien conectada a través de links

Objetivo:

- Crear copias de las webs visitadas
- Procesarlas posteriormente para un motor de búsqueda que indexe páginas
- Crear un sistema de búsquedas rápidas

Web Crawling

- Repetir:
  - Tomar una URL de la lista
  - Recuperarla y parsearla
  - Extraer los outlinks de la página
  - Agregar outlinks a la lista semilla



Legal and Ethical Considerations

- Don't break the web: Denial of Service attacks
- Copyright: respetar la propiedad intelectual de otros.
- Más vale prevenir que lamentar.

`requests` Library

- Permite realizar peticiones HTTP a cualquier URL.
- Permite extraer información, principalmente texto, como HTML.

```python
import requests
from pprint import pprint

URL = "https://sicss.io/2023/chile/"
r = requests.get(URL)
pprint(r.content)
```

```
b'\n\n\n\n\n<!DOCTYPE html>\n<html lang="en">\n  <head>\n    <meta charset="utf-8">\n<meta http-equiv="X-UA-Compatible" content="IE=edge">
\n<meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">\n<!-- The above 3 meta tags *must* come first in
the head; any other head content must come *after* these tags -->\n<title>Summer Institute in Computational Social Science</title>\n\n<!--
 Global site tag (gtag.js) - Google Analytics -->\n<script async src="https://www.googletagmanager.com/gtag/js?id=UA-113727080-1"></script
>\n<script>\nwindow.dataLayer = window.dataLayer || [];\nfunction gtag() {\n  dataLayer.push(arguments);\n}\ngtag("js", new Date());\n\ngt
ag("config", "UA-113727080-1");\n</script>\n\n<!-- Bootstrap CSS -->\n<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/boot
strap/4.3.1/css/bootstrap.min.css" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T" crossorigin="anonym
ous">\n<link href="https://fonts.googleapis.com/css?family=Roboto:300,300i,400,400i,700,700i&display=swap" rel="stylesheet">\n<!-- Local S
tylesheet -->\n<link href="/assets/stylesheets/style.css" rel="stylesheet" >\n\n<!-- HTML5 shim and Respond.js for IE8 support of HTML5 el
```

`BeautifulSoup` Library

- Funciona en conjunto con la librería `requests`.
- Es capaz de pasear el contenido obtenido.

```python
import requests
from bs4 import BeautifulSoup

URL = "https://sicss.io/2023/chile/"
r = requests.get(URL)

soup = BeautifulSoup(r.content, 'html5lib')
print(soup.prettify())
```

```
<!DOCTYPE html>
<html lang="en">
 <head>
  <meta charset="utf-8"/>
  <meta content="IE=edge" http-equiv="X-UA-Compatible"/>
  <meta content="width=device-width, initial-scale=1, shrink-to-fit=no" name="viewport"/>
  <!-- The above 3 meta tags *must* come first in the head; any other head content must come *after* these tags -->
  <title>
   Summer Institute in Computational Social Science
  </title>
  <!-- Global site tag (gtag.js) - Google Analytics -->
  <script async="" src="https://www.googletagmanager.com/gtag/js?id=UA-113727080-1">
  </script>
  <script>
```

Gracias por su atención

Summer Institutes of Computational Social Science (SICSS) 2023

References

- https://requests.readthedocs.io/en/latest/
- https://www.crummy.com/software/BeautifulSoup/
- https://monashdatafluency.github.io/python-web-scraping/