Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2115 - Programación como Herramienta para la Ingeniería

Web scraping

Profesor: Felipe Gutiérrez
Prof. Coordinador: Hans Löbel

Que es el Web Scraping?

Web Scraping es el proceso de extraer datos web y, generalmente, convertirlos a datos estructurados.
El proceso de Web Scraping generalmente implica las siguientes etapas:

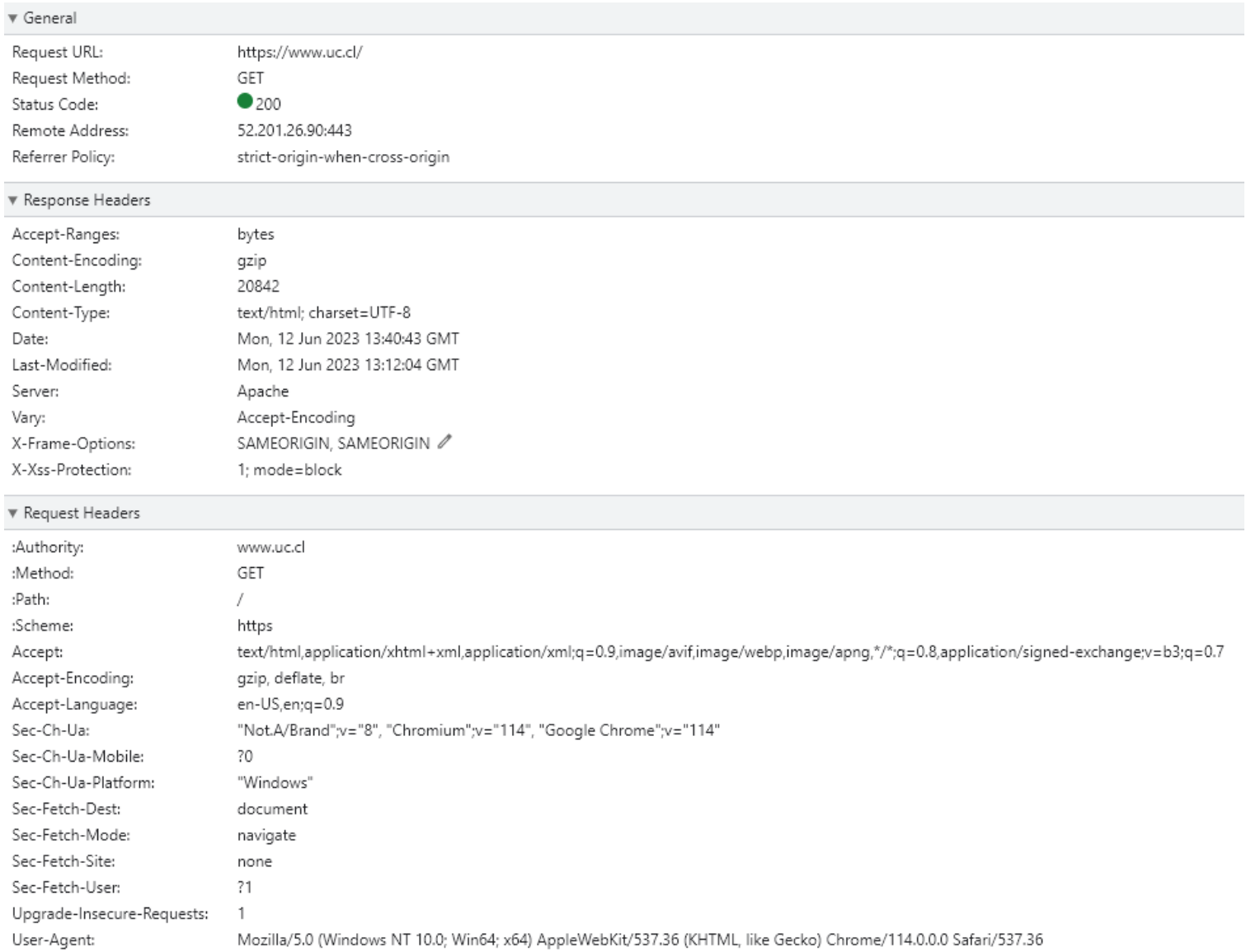
- Obtener acceso a la página web
- Analizar el contenido
- Extraer los datos
- Almacenar los datos

Obtención de acceso: Protocolo de Transferencia de Hipertexto (HTTP)

El HTTP es un protocolo de comunicación utilizado para el intercambio de información entre un cliente y un servidor. Este proporciona una estructura para la solicitud y respuesta de datos, permitiendo que los navegadores web soliciten recursos.

- El cliente, generalmente un navegador, envía solicitud HTTP al servidor que aloja recursos solicitados.
- La solicitud está compuesta por una línea inicial que contiene el método de solicitud (GET, POST, PUT, DELETE, etc.), la URL del recurso y la versión del protocolo HTTP. La solicitud puede incluir encabezados que proporcionan información adicional (como quien solicita los datos)
- El servidor recibe la solicitud y procesa la información proporcionada. Esto implica verificar la validez de la solicitud, autenticar al cliente si es necesario y realizar las acciones correspondientes.
- Posteriormente, el servidor envía su respuesta la cual será analizada por el cliente para determinar si la solicitud fue exitosa.

Como se ve una solicitud



▼ General

Request URL:	https://www.uc.cl/
Request Method:	GET
Status Code:	200
Remote Address:	52.201.26.90:443
Referrer Policy:	strict-origin-when-cross-origin

▼ Response Headers

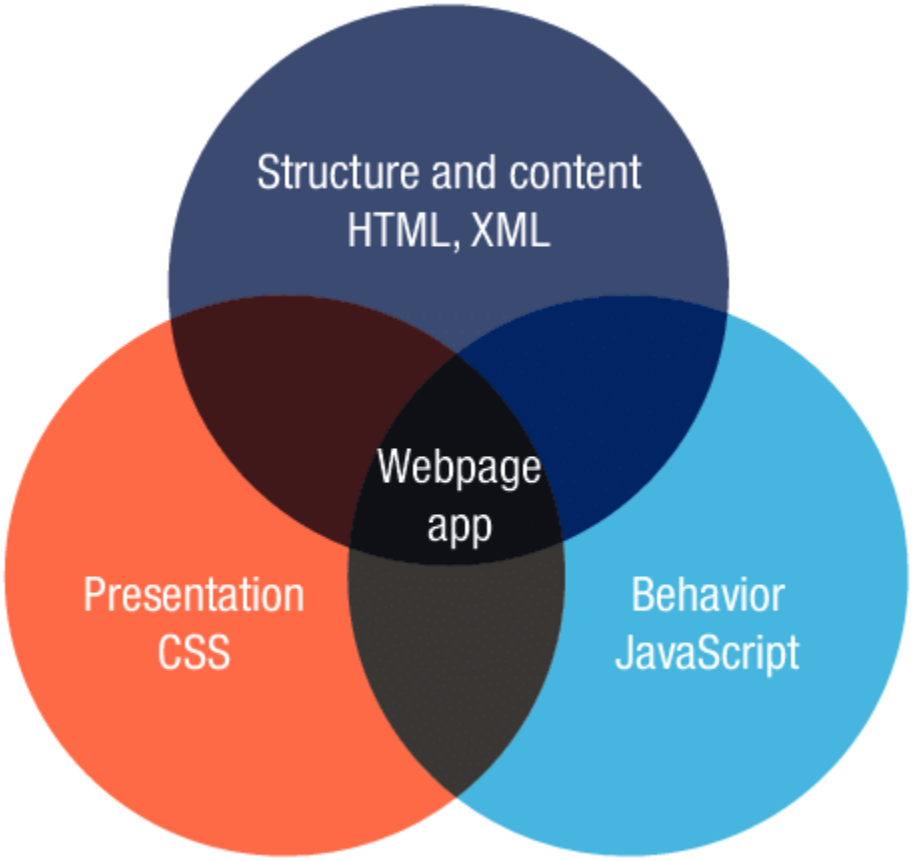
Accept-Ranges:	bytes
Content-Encoding:	gzip
Content-Length:	20842
Content-Type:	text/html; charset=UTF-8
Date:	Mon, 12 Jun 2023 13:40:43 GMT
Last-Modified:	Mon, 12 Jun 2023 13:12:04 GMT
Server:	Apache
Vary:	Accept-Encoding
X-Frame-Options:	SAMEORIGIN, SAMEORIGIN
X-Xss-Protection:	1; mode=block

▼ Request Headers

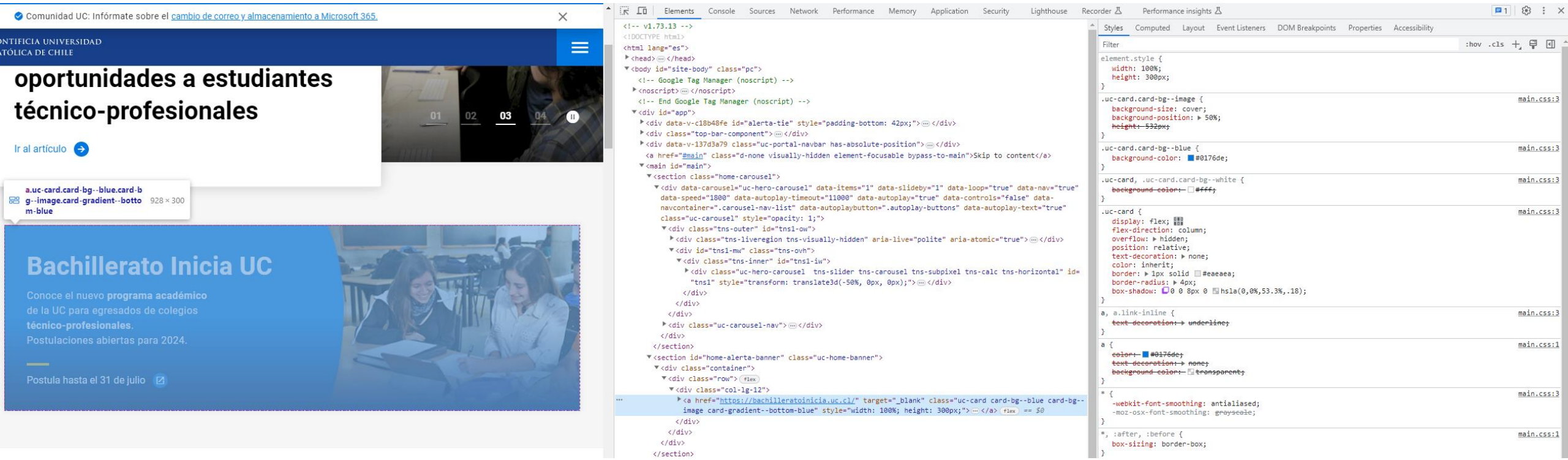
:Authority:	www.uc.cl
:Method:	GET
:Path:	/
:Scheme:	https
Accept:	text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,image/apng,*/*;q=0.8,application/signed-exchange;v=b3;q=0.7
Accept-Encoding:	gzip, deflate, br
Accept-Language:	en-US,en;q=0.9
Sec-Ch-Ua:	"Not.A/Brand";v="8", "Chromium";v="114", "Google Chrome";v="114"
Sec-Ch-Ua-Mobile:	?0
Sec-Ch-Ua-Platform:	"Windows"
Sec-Fetch-Dest:	document
Sec-Fetch-Mode:	navigate
Sec-Fetch-Site:	none
Sec-Fetch-User:	?1
Upgrade-Insecure-Requests:	1
User-Agent:	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/114.0.0.0 Safari/537.36

Contenido web – La triada de la www

- El contenido web está estructurado principalmente por texto con distintas funcionalidades:
- HTML: Lenguaje marcado que estructura y presenta el contenido web y define la estructura a través de elementos y etiquetas.
- CSS: Lenguaje de hojas de estilo utilizado para dar estilo y diseño a páginas web.
- Javascript: Lenguaje de programación que agrega interactividad y funcionalidad dinámica a las páginas web y permite realizar acciones del lado del cliente.
- Extra: La persistencia y almacenamiento se encuentra en bases de datos.



Ejemplo de HTML Y CSS en la web de la uc



¿Y que hay sobre los datos dinámicos?

- El cliente puede hacer solicitudes de información al servidor, por ej: mediante formularios.
- Javascript captura el evento y envía los parámetros de la información solicitada al servidor a través de una solicitud HTTP.
- Tras recibir la respuesta, se inyecta la información en el sitio web (o se realiza la acción solicitada).

Protocolo de Información de Aplicación (API)

- Otra forma de disponibilizar datos que ofrecen los sitios web es a traves de una API
- Las API proporcionan una interfaz de programación que define métodos y funciones para acceder a los datos disponibles
- Estas son muy versátiles y ya que permiten la integración a servicios externos y acceso a bases de datos, entre otras cosas.
- Esto también permite que se desarrollen librerías que faciliten el acceso a una API específica.
- Ejemplos: Google Maps, Twitter, Spotify, Chat GPT cuentan con API para solicitar y procesar información o integrar sus servicios de manera programática para tu propio proyecto.

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2115 - Programación como Herramienta para la Ingeniería

Web Scraping

Profesora: Felipe Gutiérrez
Prof. Coordinador: Hans Löbel