

Introducción al Web Scraping con R

Escuela de Verano 2025

Tania Hutt

Bienvenidos

Objetivos de la Clase

- Entender qué es el web scraping y sus aplicaciones.
- Aprender a extraer datos de una página web usando R.
- Practicar con un ejemplo real: **Cartelera de conciertos 2025 en Chile.**

Acceso a materiales

<https://github.com/taniahutt/EV2025>

Preparación: Instalación de R y RStudio

Pasos:

1. Descargar e instalar R

- Ir a <https://cran.r-project.org/>.
- Seleccionar el sistema operativo de su computador.
- Descargar la versión más reciente y seguir las instrucciones de instalación.

2. Descargar e instalar RStudio

- Ir a <https://posit.co/download/rstudio-desktop/>.
- Descargar la versión gratuita de RStudio Desktop.
- Instalar siguiendo las instrucciones.

3. Verificar la instalación

- Abrir RStudio.
- En la consola, escribir `version` y presionar Enter para confirmar que todo funciona correctamente.

¿Qué es el Web Scraping?

El **web scraping** es una técnica para extraer información de sitios web de forma automatizada.

Aplicaciones:

- Crear bases de datos para análisis.
- Monitorear cambios en páginas web.
- Obtener datos actualizados en tiempo real.

Herramientas en R

Para hacer web scraping, utilizaremos las siguientes librerías:

```
# Instalamos las librerías si no las tienen
if (!require("rvest")) install.packages("rvest")
if (!require("tidyverse")) install.packages("tidyverse")
```

Pasos Básicos para el Web Scraping

1. Leer el HTML de la página web.
2. Identificar los elementos HTML que contienen los datos deseados.
3. Extraer y limpiar los datos.

Ejemplo: Cartelera de Conciertos 2025 en Chile

Paso 1: Leer el HTML

Usaremos la URL:

<https://cooperativa.cl/noticias/magazine/musica/shows-en-vivo/cartelera-de-conciertos-2025-en-chile/2024-07-30/115338>

Paso 2: Inspeccionar los Elementos HTML

- 1. Abrir la página en un navegador.**
- 2. Hacer clic derecho en el elemento que contiene la información deseada.**
- 3. Seleccionar "Inspeccionar" para ver el código HTML.**

En este caso, buscaremos los elementos que contienen la lista de conciertos.