



Web Scrapping y acceso a datos desde la web

Cristián Ayala (caayala@uc.cl)

3 de mayo de 2022

Curso	Web Scrapping y acceso a datos desde la web
Nombre en inglés	Web Scrapping and web data access
Sigla	sin sigla
Créditos	5 UC
Módulos	2
Requisitos	SOL4001 Procesamiento Avanzado de Bases de Datos
Carácter	Optativo
Profesores	Cristián Ayala , Ingeniero Civil de la Pontificia Universidad Católica de Chile. Magister en Sociología UC. Director DESUC
Fechas	Primera versión 2022
Horario	Clases: martes y jueves de 18 a 20 horas; Cápsulas web
Lugar	Clases sincrónicas: Clases expositivas online, vía zoom

1 Descripción

Internet es una fuente importante de datos para las ciencias sociales y humanidades. Puede tratarse de datos alojados en páginas web, habitualmente en formato **html** o accediendo a servicios como Google Sheets, YouTube o Twitter mediante APIs provistas por esas empresas. En varios casos se han desarrollado paquetes de R para recuperar información desde ellos de manera fácil e intuitiva.

Este curso explorará distintas formas de acceder a ellos de manera programática utilizando R. Se mostrará también técnicas para limpiar, tabular y crear bases de datos para análisis posteriores.

Al final de este curso los alumnos debiesen tener la capacidad de acceder a nuevas fuentes de datos para su análisis. Esta habilidad es de gran utilidad práctica porque más y más información es generada, almacenada y —de alguna manera— disponible en Internet.

Este curso tiene como requisito haber realizado *Procesamiento Avanzado de Bases de Datos* (SOL4001) para estar familiarizado con el lenguaje y manejo de datos estructurados en R.

2 Público objetivo

El curso está dirigido a profesionales o licenciados de diversas áreas de las ciencias sociales, humanidades, comunicaciones o educación, que deseen ampliar el repertorio de fuentes de datos a su disposición para posteriores análisis cuantitativos.

3 Requisitos de ingreso

- Grado académico o título profesional, obtenido en universidades chilenas o extranjeras, equivalente al grado de licenciado que confiere la Pontificia Universidad Católica de Chile.
- Currículum vitae con antecedentes curriculares.
- Haber aprobado el curso **SOL4001 *Procesamiento Avanzado de Bases de Datos***.
- Es deseable conocimiento intermedio del idioma inglés.

4 Objetivos de aprendizaje

1. Entender el funcionamiento y estructura de una página web con miras a identificar información posible de capturar programáticamente.
2. Seleccionar y dominar distintas técnicas de captura de datos en la web según las finalidades de investigación que se desee enfrentar.
3. Obtener información de servicios web mediante sus APIs y librerías diseñadas para ello.
4. Ampliar y mejorar la capacidad de limpieza de datos no estructurados mediante expresiones regulares y programación funcional utilizando funciones y paquetes del **tidyverse**.

5 Contenidos

Los contenidos cubiertos en el curso son los siguientes:

- Comprensión de la estructura y funcionamiento de una página web.
- Realizar *web scrapping* mediante el paquete **rvest** y **polite**.
- Programación funcional para manejar sobre estructuras de datos como JSON o listas mediante funciones del paquete **purrr**.
- Limpiar y modificar caracteres mediante expresiones regulares (**stringr**).
- Acceder y modificar información en planillas de Google Sheets mediante el **googlesheets4**.
- Capturar información y tweets en Twitter mediante el paquete **rtweet**.

6 Metodologías de aprendizaje

El curso cuenta con tres componentes pedagógicos:

- Uno *expositiva* en donde se mostrará algunos aspectos teóricos y prácticos a considerar para comprender y diseñar técnicas de *web scrapping* o acceso a APIs de algunos servicios.
- El segundo es *práctico* en donde se le pedirá a los alumnos que apliquen lo aprendido hasta el momento en ejercicios concretos.
- El tercero serán *cápsulas a-sincrónicas* en las que se mostrarán conceptos o ejemplos a través de videos pregrabados.

7 Evaluación

La nota final del curso se calcula a partir de dos componentes:

Tareas (60%) 3 tareas prácticas relacionadas con los temas vistos en clases.

Proyecto final (40%) Proyecto de captura y análisis de datos individual a partir de un objetivo de investigación propuesto por el alumno.

8 Equipo docente

Cristián Ayala

Ingeniero Civil Industrial de la Pontificia Universidad Católica de Chile.

Magíster en Sociología de la Pontificia Universidad Católica de Chile.

Director de la Dirección de Estudios Sociales (DESUC) del Instituto de Sociología UC.

9 Modalidad

Remota, con clases y talleres sincronicos y cápsulas a-sincrónicas.

10 Bibliografía

El material del curso es auto-contenido, no siendo necesarias lecturas obligatorias. Se sugieren el siguiente material para reforzar el uso de R y revisar la documentación en línea de cada uno de los paquetes que se utilizarán en este curso.

Wickham, Hadley. 2019. *Advanced r*. CRC press. <https://adv-r.hadley.nz>.

Wickham, Hadley, y Garrett Grolemond. 2017. «R for data science». <https://r4ds.had.co.nz>.

Xie, Yihui, J. J. Allaire, y Garrett Grolemond. 2021. *R Markdown: The Definitive Guide*. Chapman & Hall/CRC. <https://bookdown.org/yihui/rmarkdown/>.