ESTACIÓN LASTARRIA

Web Scraping con R y RStudio
Estación Lastarria – 2022

Descripción

El presente curso tiene como principal objetivo entregar una aproximación general a técnicas avanzadas de web scraping que facilita el lenguaje de programación R. A lo largo de los encuentros se abordarán los diferentes tópicos referidos a las distintas etapas del raspado de páginas de internet (detención del tipo de página a raspar, inspección de su estructura básica, detección de los tags que enmarcan la información de interés, desarrollo de funciones para encadenar las líneas de código necesarias para obtener la información, transformación de datos semiestructurado a datos estructurados en formato 'tidy'). Estos pasos quedarán escritos en scripts enmarcados en un proyecto de Rstudio.

Durante el curso se compartirán (vía GitHub) archivos de datos, código, aplicaciones y referencias bibliográficas de interés. De este modo, cada estudiante contará con el material ya construido en base al cual podrá continuar su aprendizaje a la vez que podrá implementar rápidamente el uso de técnicas avanzadas de web scraping en R y RStudio para sus propios fines de investigación, sea en el área académica o profesional.

Responsable del curso

Agustín Nieto.

Profesor y doctor en historia por la Universidad Nacional de Mar del Plata. Se desempeña como docente en el Departamento de Sociología de la Facultad de Humanidades del UNMdP. En los últimos dos años ha impartido cursos sobre el uso de R en las ciencias sociales y las humanidades encarreras de grado y posgrado. Sus temas de interés se articulan en torno al análisis computacional de la conflictividad social pasada y presente. En el ámbito de la investigación se ha vinculado con proyectos de alto nivel académico y profesional desarrollados por prestigiosas instituciones internacionales y del sistema de investigación científica en Argentina (AUIP, CONICET). Actualmente es investigador independiente del CONICET con lugar de trabajo en el Instituto de Humanidades y Ciencias Sociales. Es autor y desarrollador de "ACEP". Se trata de un paquete de funciones en lenguaje R útiles para la detección y el análisis de eventos de protesta en corpus de textos periodísticos. Sus funciones son aplicables a cualquier corpus de textos (https://agusnieto77.github.io/ACEP/).

Horario

Las sesiones se realizarán los días miércoles, a partir del miércoles 5 de enero de 2022 de 19:00 a 21:00 horas de Chile. Esto será a través de la plataforma Meet, herramienta de Google para celebrar reuniones remotas, permitiendo interactuar a un expositor con el resto de los participantes.

Público objetivo

Estudiantes de pregrado (cuarto año en adelante) y postgrado del área de las Ciencias Sociales, Humanidades, y Ciencias de la Empresa, Marketing y Administración. Profesionales de las mismas áreas mencionadas.

Objetivos de aprendizaje

Objetivo general:

Aproximar a lxs estudiantes a diferentes técnicas de web scraping, al conocimiento de las estructuras de las páginas web y sus etiquetas html, a la transformación de datos semiestructurado a datos estructurado, a la gestión de corpus de texto, en base a las distintas librerías disponibles en CRAN. Se espera que al final del curso cada estudiante haya integrado los conocimientos necesarios y suficientes para adaptar autónomamente las distintas funciones de raspado web desarrolladas en el marco del curso a sus propias instancias de investigación académica y/o profesional.

Objetivos de aprendizaje específicos:

- Identificar las distintas estructuras que presentan las páginas web para definir el enfoque a ser aplicado en el proceso de raspado web.
- Manejar las herramientas básicas para desarrollar una inspección profunda de la estructura de etiquetas html de distintas páginas web.
- Conocer el abanico de librerías y paquetes de funciones disponible en CRAN para desarrollar raspado web con R y Rstudio.

- Desarrollar funciones de raspado web con el enfoque adecuado según sea el contenido y la estructura de la página web.
- Poder almacenar las grandes masas de texto raspadas en la web en un formato tabular.
- Reconocer la bibliografía especializada sobre las técnicas de web scraping.

Requisitos mínimos

- Es deseable un conocimiento básico de internet y de páginas web (no es excluyente).
- Cada estudiante deberá contar con un ordenador operativo y, en lo posible, un manejo intermedio en instalación y configuración de softwares (descarga de archivos, instalación y configuración de programas, etc.).
- Debido a que se trata de un curso de continuidad en la parrilla de cursos de Estación Lastarria, se espera que lxs estudiantes cuenten con las competencias básicas para usar lenguaje de programación R orientado al análisis de datos. Específicamente: lógica general en uso de sintaxis: lectura y manejo de bases de datos (desde formato CSV, TXT, RDS, SQL); gestión de paquetes de funciones especializadas; un manejo fluido de la familia de paquetes *tidyverse* (en particular *dplyr*).
- Es deseable un manejo intermedio del idioma inglés (no es excluyente).

ESTACIÓN LASTARRIA

Contenidos por sesión

Sesión	Contenidos	
1. Introducción al raspado web con R y RStudio	Presentación general del curso. Un repaso sobre el lenguaje de programación R y de RStudio. Organización del directorio de trabajo. Creación de proyectos. Vinculación con GitHub. ¿Qué es el web scraping? ¿Cuándo debemos hacer uso del web scraping y cuando no? API o No API, esa es la cuestión. Algunos ejemplos de uso de APIs: Twitter y el paquete rtweet.	2 horas cronológicas
2. Introducción a la estructura de etiquetas HTML	Una introducción a HTML, CSS y XPath: la importancia de las etiquetas para la recuperación de la información que necesitamos. Inspección de estructuras HTML: herramientas nativas y softwares para la inspección, detección y selección de etiquetas HTML para el rapado web (F12, SelectorGadget, ScrapeMate).	
3. Web scraping en páginas estáticas	¿Cómo hacer web scraping de páginas estáticas en R? Recuperación de información publicada en la web: páginas estáticas. Introducción al paquete rvest. Instalación del paquete desde RStudio. Reconocimiento de las funciones básicas de rvest: read_html(), html_elements(), html_text(), html_table().	
4. Funciones para el raspado masivo de páginas estáticas	¿Cómo llevar a cabo un raspado masivo del contenido de páginas estáticas? Combinar las funciones de rvest con las funciones de otra librería del paquete tidyverse: purrr. Transformación de la información semi-estructurada en datos estructurados.	
5. Web scraping en páginas dinámicas	¿Cómo hacer web scraping del contenido de páginas dinámicas en R? Recuperación de información publicada en la web: páginas dinámicas. Introducción al paquete RSelenium. Instalación del paquete desde RStudio. Reconocimiento de las funciones básicas de RSelenium: remoteDriver(), rsDriver(), navigate(), findElement().	
6. Funciones para el raspado masivo de páginas dinámicas	¿Cómo llevar a cabo un raspado masivo del contenido de páginas dinámicas? Combinar las funciones de rvest con las funciones de rvest y purrr. Guardar la información raspada en formato tabular.	
7. Automatización de las tareas de raspado web: PC, Raspberry Pi, VPS	Tres soportes para una misma tarea: 1) ejecución automática de scripts en R desde una PC de escritorio o portátil; 2) ejecución automática de scripts en R desde una Raspberry Pi; 3) ejecución automática de scripts en R desde una máquina virtual (VPS). Una introducción a crontab y a los paquetes cronR (Linux) y taskscheduleR (Windows).	