# Web Scraping con R

## Introducción

World Wide Web está compuesta por muchos millones de documentos enlazados entre sí, conocidos también como páginas web. Para extraer del texto fuente únicamente la información que le interesa al usuario, se utiliza un tipo software especial. Se trata de los programas llamados web scrapers, crawlers, spiders o, simplemente, bots, que examinan el texto fuente de las páginas en busca de patrones concretos y extraen la información que contienen. Los datos conseguidos mediante web scraping posteriormente se resumen, combinan, evalúan o almacenan para ser usados más adelante.

## Objetivos

El objetivo de este curso es dotar a los usuarios de los conocimientos básicos sobre la técnica de adquisición de datos conocida como web scraping. Con ello y aprenderán a extraer datos textuales publicados en una página web.

## A quien va dirigido

El curso está dirigido a profesionales o personas interesadas en el análisis de datos y en el mundo de la programación

**Los alumnos deben disponer de conocimientos y/o experiencia en programación R**

## Temario

Módulo 1: Introducción al Webscraping

1.¿Qué es el Webscraping?

- Definición y aplicaciones.
- Consideraciones legales y éticas.

2.Instalación y configuración del entorno de trabajo

- Instalación de R y RStudio.
- Instalación de paquetes necesarios: rvest, httr, xml2, tidyverse.

3.Conceptos básicos de HTML y CSS

- Estructura de un documento HTML.
- Selectores CSS.

Módulo 2: Fundamentos de Webscraping con R

1.Uso del paquete rvest

- Introducción a rvest.
- Lectura de páginas web con read_html().
- Selección de nodos HTML con html_nodes() y html_node().
- Extracción de texto y atributos con html_text() y html_attr().

2.Navegación y extracción de datos

- Manejo de datos tabulares con html_table().
- Scraping de múltiples páginas.

3.Manejo de sesiones con httr

- Introducción a httr.
- Realización de solicitudes GET y POST.

Módulo 3: Scraping Avanzado

1.Scraping de sitios dinámicos

- Introducción a sitios web dinámicos.
- Uso de rvest y httr con JavaScript.
- Introducción a RSelenium para scraping de contenido dinámico.

2.Automatización de tareas de scraping

- Uso de RSelenium para interactuar con formularios y botones.
- Extracción de datos de sitios web que requieren autenticación.

## Duración y Desarrollo

12 horas teórico-practicas
Del 7 al 9 de Octubre de 9 a 13 horas
Modalidad Presencial-Virtual

## Condiciones

Curso enmarcado en el Digital Tatent Hub. Gratuito para empresas socias de GAIA. Otro tipología de empresas pueden ponerse en contacto con dth-academy@gaia.es.

**Cancelaciones**: Si cancelas tu inscripción con un margen mínimo de 4 días laborables previos al inicio del curso, no se aplicará ninguna penalización.
En caso de cancelar tu inscripción con un margen menor a 4 días laborables, se estudiará el % de penalización aplicable.
No informar, y/o no presentarse a la formación puede suponer un cargo de entre 150-300€.