SCRAPING WEB

Scraping con Webscrapper

¿Qué es el scraping?

El Web scraping nos permite extraer información de la web de forma estructurada para poder manejarla como datos y tratarla para su análisis y visualización.

Existen diferentes técnicas para hacer scraping. Todas ellas utilizan algún software o aplicación, bien directamente en el navegador bien usando algún lenguaje con Python que, mediante script, captura la información disponible en una web.

El scraping usa técnicas automatizadas para extraer la información publicada en la web en una escala de automatización que va desde las acciones manuales ("copiar y pegar") hasta sistemas de reconocimiento de información semántica, pasando por aplicaciones de web scraping, parsers HTML, algoritmos de minería de datos, etc.

No todas las web permiten ser "scrapedas". En ocasiones, los administradores de los sitios web limitan esta posibilidad mediante técnicas como bloqueos de IPs, limitar accesos con el fichero robots.txt, etc.

Qué es Web scraper y cómo usarlo

Web Scraper es una extensión para navegador que resulta muy útil para raspar contenido de una web ya que puede ejecutarse directamente desde el navegador web sin necesidad de instalar otro software adicional.

Para utilizarlo debemos:

1. Descargarnos el plugin para Chrome e instalarlo.
2. Para acceder al plugin debemos abrir en Chrome la consola de herramientas de desarrolladores. Para acceder al DevTools debes seccionar el menú de la esquina superior derecha del explorador, luego seleccionando More Tools > Develope Tools. Alternativamente puedes usar Ctrl+Shift+I o Cmd+Opt+I en Mac, y en Windows presionar la tecla F12

Acceder a la consola de herramientas de desarrolladores mediante menú

Ejemplo de uso: Capturar datos de la Wikipedia

Veamos un caso práctico de uso. Pongamos por caso que queremos descargar un conjunto de datos de una entrada de la wikipedia de un país. Habitualmente, encontramos información en tablas como, por ejemplo, de la población. Vamos a ver cómo extraer una tabla.

Web Scraper en DevTools

- Accede a la página de la wikipedia de Marruecos
- Localiza la tabla Populations (in thousands)
- Abre del DevTools y accede a la pestaña de Web Scraper

- Crea un mapa del sitio seleccionado “Create new sitemap”. Indica el nombre que quieras en “Sitemap name”, por ejemplo “wiki-marruecos” y en “Start URL” incluye la URL de la página web, en este caso “https://en.wikipedia.org/wiki/Morocco”
- Una vez creado el sitemap que hemos llamado wiki-marruecos debemos crear el selector desde el que vamos a extraer la información, en este caso, la tabla «Populations (in thousands)».
- Pon un nombre a Id, por ejemplo «tabla-poblacion».
- En Type selecciona «Table».
- En Selector, pulsa el botón Select y luego, en la página web, pulsa sobre toda la tabla. Quedará seleccionada en rojo. En el popup que aparece, pulsa sobre «Done selecting»
- En «Header row selector» seleccionar Select y, en la página web, marcar la fila de encabezado (Year, Pop…)y en el popup que aparece, pulsa sobre “Done selecting”
- Marca el cuadro de «Multiple» para que cargue todas las demás filas de datos
- Salva el selector con «Save selector»
- Nos aparecerá la información del selector guardado y podremos comprobar qué datos se van a cargar pulsado el botón «Data preview»

- Si queremos, podríamos repetir el proceso creando nuevos selectores. Estos pueden ser tablas, pero también podemos crear selectores html, imágenes, etc.
- En este caso vamos a «raspar» los datos y a descargarlos directamente en csv.
- En el menú de Web Scraper de nuestro Sitemap “Sitemap wiki-marruecos” seleccionamos la opción «Scrape»
- Cuando finalice, en ese mismo menú seleccionamos la opción «Export Data as CSV». Se descargará la tabla y ahora ya podremos trabajar con ella en un editor de hojas de cálculo o importarla a un servicio web para crear una infografía, etc., como en este ejemplo creado con Infogram.

Evolución de la población de Marruecos desde 1960

Aprender más

La web de Web Scraper tiene un **completo manual con videotutoriales** en el que puede seguirse, con más detalle, cómo trabajar con esta herramienta.

Navigate multi-level navigation to scrape all items in an e-commerce site.

Scrape an e-commerce site with pagination.

Grupo Ciberimaginario | Manuel Gertrudix - Alejandro Carbonell | 2025/2026 | Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional. Los contenidos citados se ajustan a lo regulado en el art. 32 del TRLPI de España