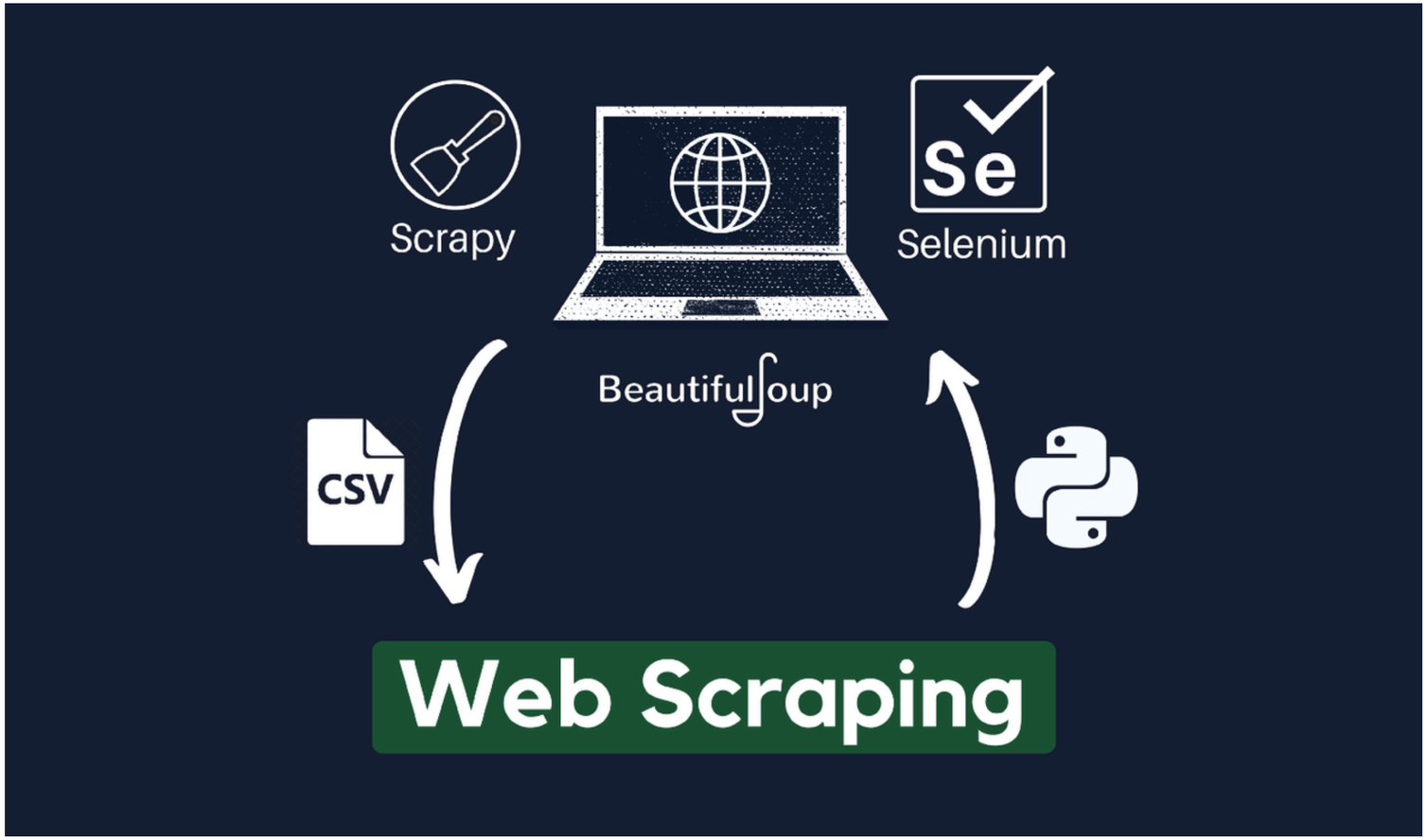
AI F5 SCHOOL

factoria F5
POWERED BY SIMPLON

Co-funded by
the European Union

Web Scraping

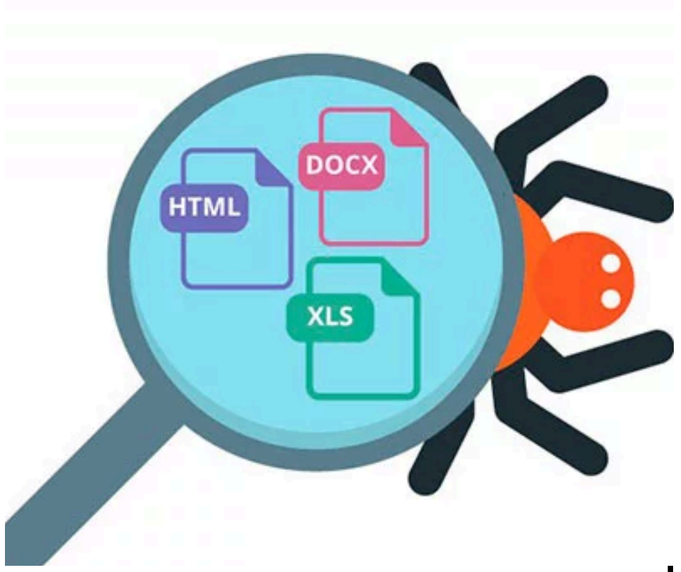


AI F5 SCHOOL

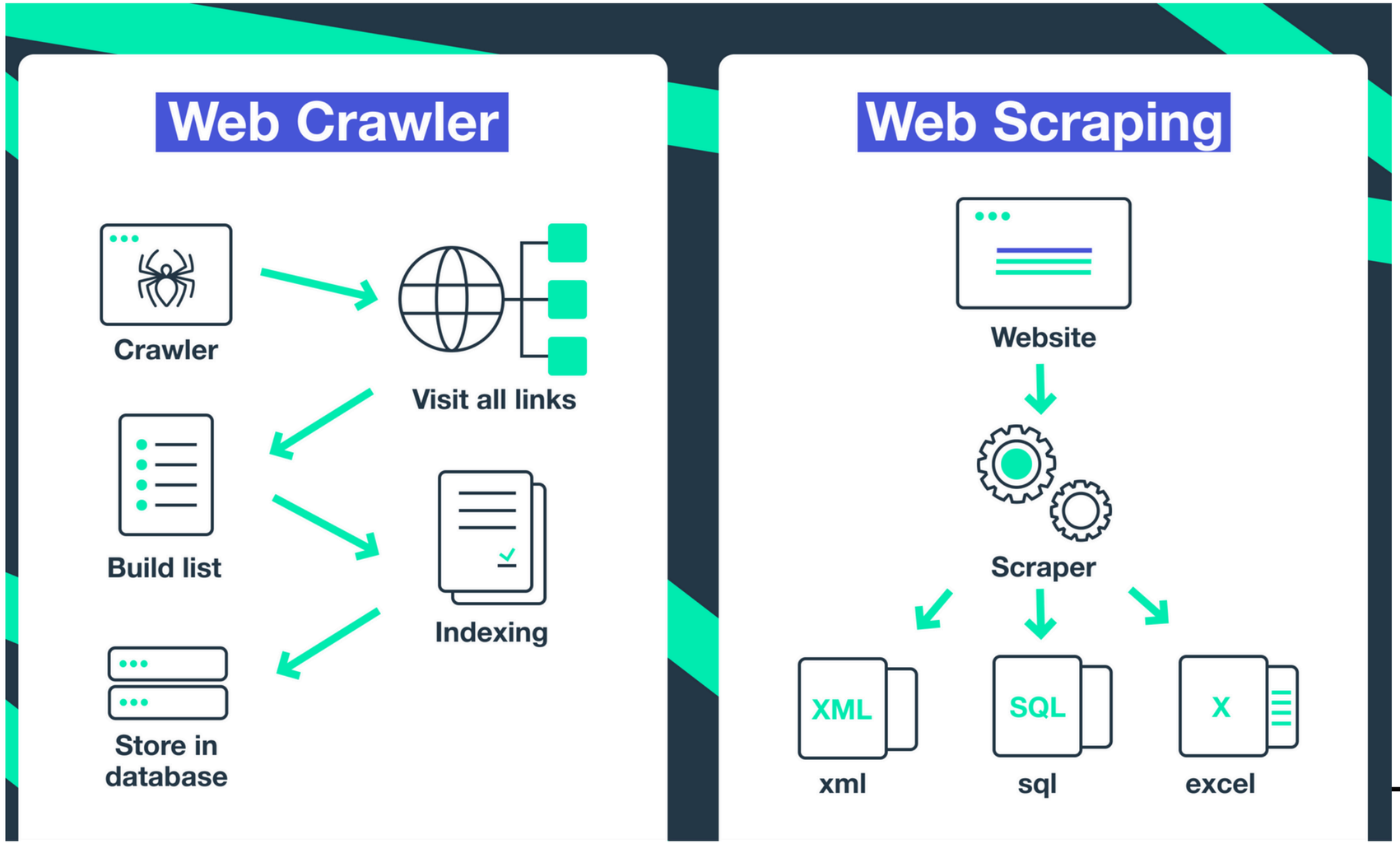
¿Que es Web Scraping?

El web scraping, también conocido como raspado web o extracción de datos web, es una técnica que utiliza software para extraer información de sitios web.

Funciona de manera similar a como lo haría una persona: el programa envía una solicitud al sitio web, recibe el código HTML de la página y luego extrae los datos específicos que le interesan.

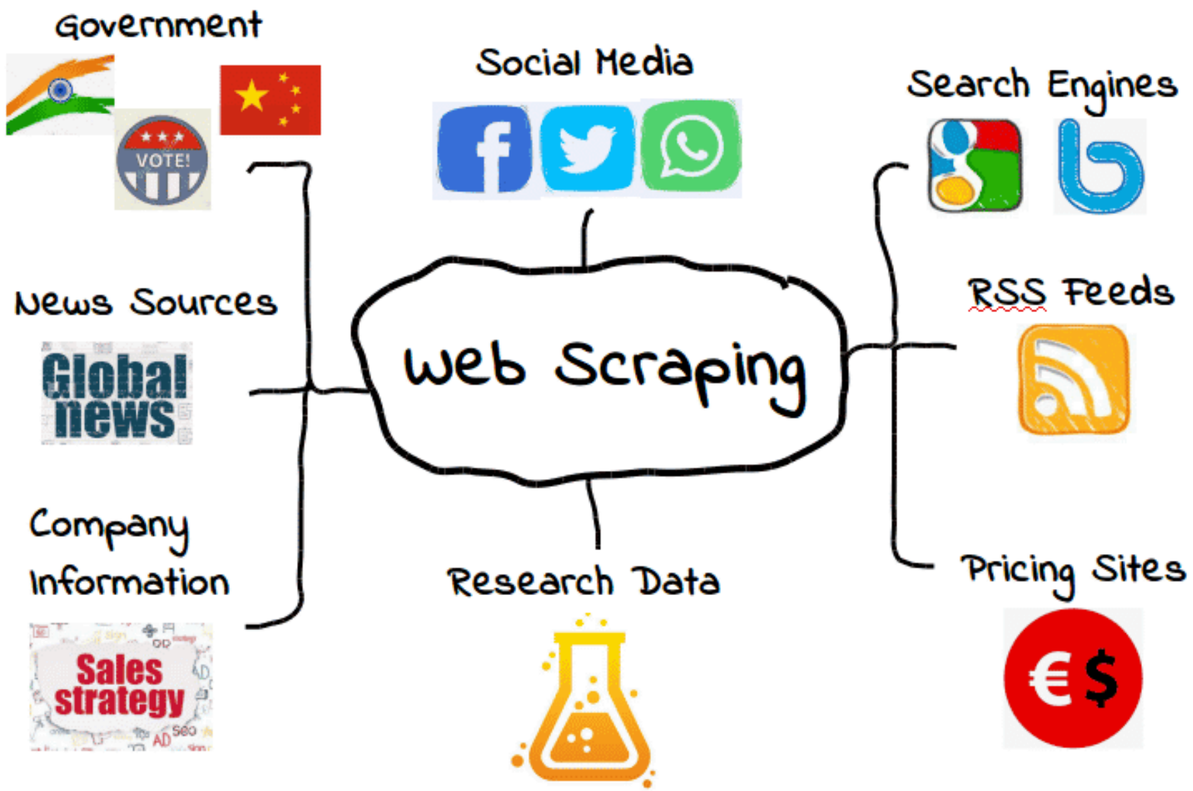


Web Scraping y web Crawling



Usos del Web Scraping

- Brand monitoring and competition analysis
- Machine learning data acquisition
- Financial data analysis
- Social media trend analysis
- SEO monitoring for website ranking insights.



Historia del Web Scraping

1989: Nace la World Wide Web, sentando las bases para la necesidad de organizar y automatizar la recopilación de información.

1993: El primer robot web, WorldWideWeb Wanderer, y el primer motor de búsqueda basado en crawler, JumpStation, aparecen para indexar y organizar la web.
2000: Las Web APIs y APIs de crawler proporcionan métodos estructurados para acceder y extraer datos.
2004: Python Beautiful Soup facilita el análisis y parseo de HTML para la extracción de datos específicos.

Las herramientas visuales de web scraping como ParseHub permiten que usuarios sin conocimientos de programación creen scrapers web.

Tipos de Web Scraping

- Scraping basado en el HTML (parsing de páginas web)
- Scraping basado en APIs
- Scraping con técnicas de DOM (Document Object Model)
- Headless browsers (Navegadores sin interfaz gráfica)

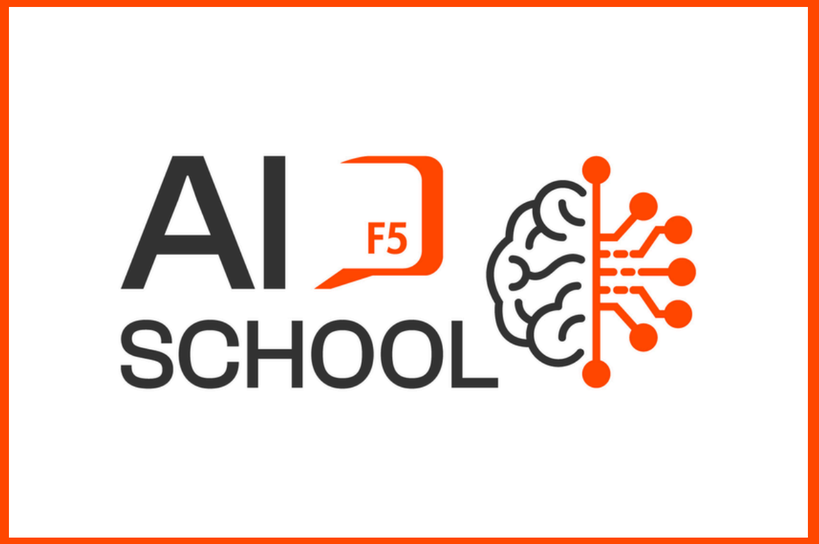
Tecnologías Principales

- BeautifulSoup: Librería para parsing de HTML y XML
- Scrapy: Framework completo de scraping
- Selenium: Automatización de navegadores web
- Requests: Librería para hacer solicitudes HTTP
- Pandas: Manipulación y análisis de datos extraídos

Aspectos Legales y Éticos

- Revisión de términos de servicio de los sitios web
- Respeto a la privacidad y uso responsable de los datos
- Buenas prácticas y técnicas para evitar la sobrecarga del servidor

Vamos al Notebook



F5
factoría
POWERED BY SIMPLON

Gracias