Web scraping

Para economistas

Alejandro Acosta León

1 de junio de 2024

Economía | udla

Material

El material de este taller está disponible en:
`https://github.com/alejo-acosta/web-scraping`

Pueden descargar el material en la pestaña de código y luego en "Download ZIP".
O pueden clonar el repositorio con el siguiente comando:
`git clone https://github.com/alejo-acosta/web-scraping`

Economía | udla

Expectativas

¿Qué esperas aprender en este taller?

Escríbanlo en el siguiente Link

Economía

Agenda - Teoría

Conceptos

- ¿Qué es web scraping?
- Problemas éticos, legales y mal uso del web scraping.
- Alternativas al web scraping.
- Herramientas y librerías necesarias.

Entorno

- Instalación de Python y librerías necesarias (requests, BeautifulSoup, pandas, selenium).
- Configuración del entorno de desarrollo (puede ser Jupyter Notebook, VSCode, etc.).

Economía

Agenda - Práctica

Nivel 1: data estructurada en HTML

- Pandas :)

Nivel 2: data semi-estructurada

- Inspección de elementos HTML.
- Extracción de datos con BeautifulSoup.
- Descarga de archivos.

Nivel 3: data no estructurada y que no usa HTML

- Automatización web con Selenium.
- Extracción de información relevante.

Economía

¿Qué es web scraping?

Web scraping son técnicas que consisten en extraer información de sitios web. Estas técnicas pueden ser manuales o automatizadas.

- El problema es que los sitios web no están diseñados para ser leídos por máquinas, sino por humanos. Por lo tanto, debemos 'enseñar' a nuestra computadora a leer estos sitios web.
- Además, el internet puede ser muy ordenado o completamente caótico. Por lo tanto, debemos ser creativos y flexibles en la forma en que extraemos la información.
- Normalmente el web scraping se puede resumir en 3 'proceso'
  1. Parsing o análisis del código.
  2. Automatización (bots).
  3. Extracción de la información.

Economía | udla

Imaginémonos que vivimos dentro del internet, en la gran nación democrática y soberana de Weblandia.

Así como en la vida real, Weblandia tiene bienes públicos y privados. Normalmente, el web scraping se hace sobre 'bienes públicos'.

vs

Los bienes públicos son no excluyentes y no rivales.
¿Pero realmente es así en Weblandia o incluso en la vida real?

Economía | udla

Problemas éticos y legales

¿Cuáles?

Ejemplos de uso del web scraping en la literatura

- Cavallo, Alberto, and Roberto Rigobon. 2016. "The Billion Prices Project: Using Online Prices for Measurement and Research." *Journal of Economic Perspectives, 30 (2): 151-78.*
- Glaeser, Edward L., Hyunjin Kim, and Michael Luca. 2018. "Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change." *AEA Papers and Proceedings, 108: 77-82.*

Economía | udlb

Cuando no usar web scraping

Hacer web scraping es lento, tedioso y muchas veces poco eficiente.

- Si la información está disponible en un formato estructurado (CSV, Excel, SQL, etc.).
- Si la información está disponible a través de una API.
- Si la información es sensible o privada.
- Con fines maliciosos (spam, phishing, DDoS, brute force hacking).
- Si nuestro proceso acapara todos los recursos del servidor.
- Si se hace por 'moda'.

Revisar:

Foerderer, J. (2023). Should we trust web-scraped data?. *arXiv preprint arXiv:2308.02231.*

Economía | udla

Manos a la obra, preparemos nuestro entorno de trabajo.