M2.851. Tipología y ciclo de vida de los datos

Guía de estudio: Bloque 2. Web scraping

Presentación

Esta guía de aprendizaje presenta una breve exposición de los objetivos, contenidos más relevantes, recursos de aprendizaje y actividades de evaluación del bloque "Web Scraping" para orientar en su estudio.

Objetivos

1. Conocer el significado y los potenciales beneficios del web scraping.
2. Ser capaz de evaluar la dificultad de realizar web scraping en un sitio web determinado.
3. Ser capaz de realizar web scraping simple, utilizando Python/R.
4. Ser capaz de extraer contenido audiovisual de un sitio web.
5. Ser capaz de almacenar los datos obtenidos de Internet en un formato interoperable.
6. Conocer los principales repositorios de bases de datos creados a partir de información extraída mediante web scraping.
7. Ser capaz de buscar bases de datos disponibles para un dominio de aplicación determinado.
8. Ser capaz de solucionar los principales obstáculos implementados para evitar el web scraping.
9. Conocer los principales aspectos legales relacionados con el web scraping.
10. Conocer diferentes casos de éxito o usos prácticos de la web scraping.

Contenidos

En este segundo bloque se trata el Web scraping de diferentes tipos de datos. Web scraping es la habilidad de escribir un programa que recoge datos y los guarda en una

base de datos o fichero. No importa cuántas veces has realizado Web scraping que a menudo supone un reto. Obtener datos de Internet es casi tan antiguo como Internet misma. Aunque Web scraping no es un término nuevo, en estos últimos años también se ha denominado como screen scraping, data mining o web harvesting.

En la teoría, Web scraping es una práctica de obtener datos por cualquier medio diferente de interactuar con una API (o de manera obvia, a través de un humano usando un navegador web). Esto se consigue escribiendo un programa que hace consultas a un servidor, pide datos (usualmente en el formato HTML y otros ficheros que tienen páginas web) y parsea los datos para obtener información relevante.

En la práctica, web scraping cubre varias técnicas y tecnologías, así como análisis de datos y seguridad.

Recursos de aprendizaje

- M2: Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- M3: Tutorial de Github https://guides.github.com/activities/hello-world.

Material complementario

1. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
2. Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sueños.

Conceptos más importantes

En este módulo didáctico se han revisado los aspectos fundamentales relacionados con el web scraping. En primer lugar, se ha presentado la utilidad y potencial de esta herramienta, particularmente útil cuando se requiere obtener información de un sitio web

que no dispone de una API para tal efecto, o la API no satisface del todo las necesidades de nuestro proyecto de ciencia de datos.

Después de revisar algunos pasos previos necesarios para planificar cualquier proceso de web scraping de forma óptima, se han revisado las principales herramientas disponibles en las librerías requests y BeautifulSoup que permiten recuperar textos, imágenes o cualquier otro contenido audiovisual procedentes de Internet.

Posteriormente, se han presentado los formatos estandarizados más comúnmente utilizados en el almacenamiento de datos rastreados: CSV y JSON. Así mismo, se han introducido las herramientas Flask mediante las cuales es posible implementar una API que contenga y acceda de manera amigable a estos datos. Finalmente, se han enumerado diferentes repositorios públicos en los que se pueden compartir las bases de datos resultantes.

También se han introducido varias medidas que pueden aplicarse en cualquier sitio web para evitar el web scraping. A continuación, se han presentado varios métodos que permiten simular el comportamiento humano con el objetivo de resolver estas medidas u obstáculos, como el uso de múltiples direcciones IP, la configuración de timeouts, la gestión de logins y galletas de sesión, etc.

A continuación, se han planteado los principales aspectos legales relacionados con la extracción de datos procedentes de Internet, para posteriormente enumerar, una lista de mejores prácticas y consejos que nos permitan implementar un buen uso del web scraping.

Finalmente, con el objetivo de mostrar el potencial de esta herramienta, se han presentado algunos ejemplos de aplicación, así como casos de éxito reales, del web scraping.

Evaluación

La evaluación de este bloque se compone de una actividad práctica:

1. Práctica 1: Se trabajan los conceptos de analizar un sitio web objetivo, realizar web scraping de un sitio web y posteriormente, compartir los datos generados.