Elaboración de datasets basados en la web y aplicaciones



Índice

Web Scraping

- ¿Qué es Web Scraping?
- Historia del Web Scraping
 - 1. Origen
 - 2. Aparición de la Web 2.0
 - 3. Democratización y estándares
- Usos comunes
- Herramientas más utilizadas
- Ejemplos práctico

Datasets

- ¿Qué son los datasets?
- Historia de los datasets
 - 1. Inicios en Bases de Datos
 - 2. Era de Internet
 - 3. IA y Big Data
- Usos comunes
- Herramientas más utilizadas
- Ejemplos práctico

¿Qué es Web Scraping?

Web scraping es una técnica de extracción de datos de páginas web de forma automática.

Esta técnica emplea programas o scripts que navegan por páginas web, acceden a su contenido y recogen información a almacenar.



Historia del Web Scraping Origen

Matthew Gray, en 1993, lanzó World Wide Web Wanderer



JumpStation II Results Page

Search for "sunsite" in document titles

The following list is a verbose list of documents in the JumpStation database the have the word "sunsite" in their titles. Please note that this data is transitory, and not stored on this server.

http://www.sun.com/f0/sunsite/

Title SunSITE

Last Modification: Friday 12-Aug-94 19 17 22 GMT

Type test/little

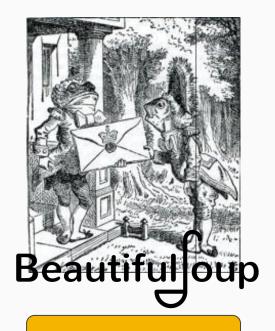
Links 19

A search for the criteria you requested yielded I document whose details are stored in the JumpStation II document database.

JumpStation Copyright 1993, 1994 by Jonathon Fletcher

j.flotchor@string.ac.uk

En 1993 surge el primer motor de búsqueda basado en un robot web Historia del Web Scraping Aparición de la Web 2.0



Sin automatizar



Automatizando

Historia del Web Scraping

Democratización y estándares

Aparecen nuevas herramientas de código abierto y en la nube



Usos comunes

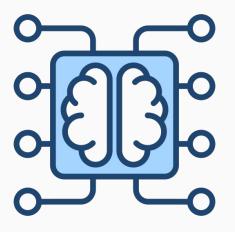
- Análisis de mercados -



- Medios comunicativos -



- Entrenamiento de modelos -



Herramientas más utilizadas

- Python y Selenium -



- ParseHub -



- Octoparse -



Creación de dataset a partir de Web Scraping



¿Qué son los datasets?

Un dataset es un conjunto de datos organizados utilizados para el análisis de datos, el entrenamiento de IA y la toma de decisiones.

Un dataset se compone de variables, registros y metadatos. Un dataset puede ser estructurado, no estructurado o semiestructurado.



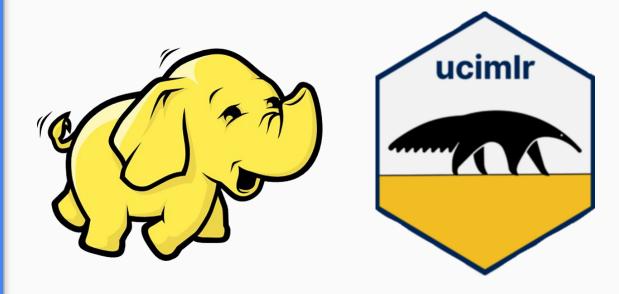
Historia de los datasets Inicios en Bases de Datos





Dominio de SQL entre los años 60 y 80

Historia de los datasets Era de Internet



Ayudan a manejar el Big Data proporcionar datasets para el aprendizaje automático

Historia de los datasets IA y Big Data



El crecimiento de los datos y el auge de la IA han fomentado el cambio en los datasets

Usos comunes

- Ciencia de datos e IA -



- Negocio y Marketing -



- Investigación -



Herramientas más utilizadas

- Recopilación de datos -



- Procesamiento de datos -



- Almacenamiento de datos -



Tratamiento de datasets en Python



The end.