MASTERCLASS
WEB SCRAPING

ANALAIZER.DIGITAL

Comencemos →

ÍNDICE

MASTERCLASS WEB SCRAPING

INTRODUCCIÓN

INTRODUCCIÓN

- ¿Que es el web scraping?

INTRODUCCIÓN

- ¿Que es el web scraping?
  - Web scraping es una técnica utilizada para extraer datos de sitios web de manera automatizada.

INTRODUCCIÓN

- **¿Que es el web scraping?**
  - Web scraping es una técnica utilizada para extraer datos de sitios web de manera automatizada.
  - **Se utiliza para recopilar grandes cantidades de datos que serían tediosos de obtener manualmente.**

INTRODUCCIÓN

- **¿Que es el web scraping?**
  - Web scraping es una técnica utilizada para extraer datos de sitios web de manera automatizada.
  - **Se utiliza para recopilar grandes cantidades de datos que serían tediosos de obtener manualmente.**

INTRODUCCIÓN

- **¿Que es el web scraping?**
  - Web scraping es una técnica utilizada para extraer datos de sitios web de manera automatizada.
  - **Se utiliza para recopilar grandes cantidades de datos que serían tediosos de obtener manualmente.**

Importancia en el análisis de datos

- Permite acceder a datos que no están disponibles a través de APIs o descargas directas.

Importancia en el análisis de datos

- Permite acceder a datos que no están disponibles a través de APIs o descargas directas.
- **Facilita la recopilación de datos de múltiples fuentes en un formato consistente.**

Importancia en el análisis de datos

- Permite acceder a datos que no están disponibles a través de APIs o descargas directas.
- Facilita la recopilación de datos de múltiples fuentes en un formato consistente.
- **Es crucial para el análisis de mercado, investigación académica, monitoreo de precios, etc.**

Importancia en el análisis de datos

- Alguien te las da directamente (como las APIs en el mundo de los datos).

Importancia en el análisis de datos

- Alguien te las da directamente (como las APIs en el mundo de los datos).
- Encuentras un libro lleno de pistas (como las descargas directas).

Importancia en el análisis de datos

- Alguien te las da directamente (como las APIs en el mundo de los datos).
- Encuentras un libro lleno de pistas (como las descargas directas).
- Tienes que buscar por toda la ciudad para encontrar pistas escondidas.

Fundamentos del web scraping

Fundamentos del web scraping

- Cómo funciona el scraping
    - Proceso básico: solicitud HTTP, descarga del contenido HTML, parsing y extracción de datos.

Fundamentos del web scraping

- **Cómo funciona el scraping**
    - Proceso básico: solicitud HTTP, descarga del contenido HTML, parsing y extracción de datos.
    - **Identificación de patrones en la estructura HTML para extraer datos específicos.**

Fundamentos del web scraping

- **Cómo funciona el scraping**
  - Proceso básico: solicitud HTTP, descarga del contenido HTML, parsing y extracción de datos.
  - Identificación de patrones en la estructura HTML para extraer datos específicos.
  - **Manejo de paginación y navegación dentro del sitio web.**

Fundamentos del web scraping

- Herramientas y lenguajes comunes
  - Python: lenguaje popular para scraping debido a sus bibliotecas.

Fundamentos del web scraping

- **Herramientas y lenguajes comunes**
    - Python: lenguaje popular para scraping debido a sus bibliotecas.
    - **BeautifulSoup: biblioteca para analizar documentos HTML y XML.**
    -

Fundamentos del web scraping

- **Herramientas y lenguajes comunes**
  - Python: lenguaje popular para scraping debido a sus bibliotecas.
  - BeautifulSoup: biblioteca para analizar documentos HTML y XML.
  - **Trafilatura: biblioteca para analizar documentos HTML y XML, orientada a blogs.**
  -

Fundamentos del web scraping

- **Herramientas y lenguajes comunes**
    - Python: lenguaje popular para scraping debido a sus bibliotecas.
    - BeautifulSoup: biblioteca para analizar documentos HTML y XML.
    - Trafilatura: biblioteca para analizar documentos HTML y XML, orientada a blogs.
    - **Scrapy: framework para crear bots de scraping.**

Fundamentos del web scraping

- **Herramientas y lenguajes comunes**
  - Python: lenguaje popular para scraping debido a sus bibliotecas.
  - BeautifulSoup: biblioteca para analizar documentos HTML y XML.
  - Trafilatura: biblioteca para analizar documentos HTML y XML, orientada a blogs.
  - Scrapy: framework para crear bots de scraping.
  - **Selenium: para sitios web que requieren interacción dinámica.**
  -

Fundamentos del web scraping

- **Herramientas y lenguajes comunes**
  - Python: lenguaje popular para scraping debido a sus bibliotecas.
  - BeautifulSoup: biblioteca para analizar documentos HTML y XML.
  - Trafilatura: biblioteca para analizar documentos HTML y XML, orientada a blogs.
  - Scrapy: framework para crear bots de scraping.
  - Selenium: para sitios web que requieren interacción dinámica.
  - **Playwright: para sitios web que requieren interacción dinámica, últimamente en tendencia ya que de forma nativa tiene también soporte en JS.**

Fundamentos del web scraping

PYTHON

Fundamentos del web scraping

PYTHON

BS4

Fundamentos del web scraping

PYTHON

BS4

SELENIUM

VENTAJAS

Ventajas del web scraping para el análisis de datos

- **Acceso a grandes volúmenes de datos**
  - **Capacidad de recopilar datos de múltiples fuentes rápidamente.**

Ventajas del web scraping para el análisis de datos

- **Acceso a grandes volúmenes de datos**
  - Capacidad de recopilar datos de múltiples fuentes rápidamente.
  - **Posibilidad de crear conjuntos de datos personalizados y únicos.**

Ventajas del web scraping para el análisis de datos

- **Automatización de la recolección de datos**
  - **Reducción del tiempo y esfuerzo en la recopilación manual.**

Ventajas del web scraping para el análisis de datos

- **Automatización de la recolección de datos**
  - Reducción del tiempo y esfuerzo en la recopilación manual.
  - **Disminución de errores humanos en la entrada de datos.**

Ventajas del web scraping para el análisis de datos

- **Obtención de datos en tiempo real.**
  - **Capacidad de obtener información actualizada constantemente.**

Ventajas del web scraping para el análisis de datos

- **Obtención de datos en tiempo real.**
    - Capacidad de obtener información actualizada constantemente.
    - **Útil para análisis de tendencias y toma de decisiones rápidas.**

Ventajas del web scraping para el análisis de datos

- **Análisis de mercado y competencia**
  - **Monitoreo de precios y productos de la competencia.**

Ventajas del web scraping para el análisis de datos

- **Análisis de mercado y competencia**
  - Monitoreo de precios y productos de la competencia.
  - **Análisis de sentimientos en redes sociales y reseñas de productos.**

Almacenamiento y procesamiento de datos

Almacenamiento y procesamiento de datos

- **Data Warehouse**
  - **Definición: Sistema centralizado para almacenar y gestionar datos estructurados de múltiples fuentes.**

Almacenamiento y procesamiento de datos

- **Data Warehouse**
  - Definición: Sistema centralizado para almacenar y gestionar datos estructurados de múltiples fuentes.
  - **Características:**
    - **Orientado a temas específicos del negocio.**
    - **Datos integrados y consistentes.**
    - **No volátil (conserva historial).**
    - **Variante en el tiempo (mantiene diferentes versiones de los datos).**

Almacenamiento y procesamiento de datos

- **Data Warehouse**
  - Definición: Sistema centralizado para almacenar y gestionar datos estructurados de múltiples fuentes.
  - Características:
    - Orientado a temas específicos del negocio.
    - Datos integrados y consistentes.
    - No volátil (conserva historial).
    - Variante en el tiempo (mantiene diferentes versiones de los datos).
  - **Relación con datos de scraping:**
    - **Almacenamiento estructurado de datos extraídos.**
    - **Facilita el análisis histórico y la generación de informes.**

Almacenamiento y procesamiento de datos

- **Data Mart**
  - **Definición: Subconjunto de un data warehouse enfocado en un área específica del negocio.**

Almacenamiento y procesamiento de datos

- **Data Mart**
  - Definición: Subconjunto de un data warehouse enfocado en un área específica del negocio.
  - **Diferencias con Data Warehouse:**
    - **Más pequeño y especializado.**
    - **Diseñado para un departamento o función específica.**
    - **Generalmente más rápido para consultas.**

Almacenamiento y procesamiento de datos

- **Data Mart**
  - Definición: Subconjunto de un data warehouse enfocado en un área específica del negocio.

  - Diferencias con Data Warehouse:
    - Más pequeño y especializado.
    - Diseñado para un departamento o función específica.
    - Generalmente más rápido para consultas.

  - **Uso en análisis de datos scrapeados:**
    - **Creación de marts específicos (ej: precios de competidores, sentiment analysis).**
    - **Permite un acceso más rápido a datos relevantes para equipos específicos.**

Almacenamiento y procesamiento de datos

- **Data Lake**
  - **Definición: Repositorio centralizado que permite almacenar todos los datos estructurados y no estructurados a cualquier escala.**

Almacenamiento y procesamiento de datos

- **Data Lake**
  - Definición: Repositorio centralizado que permite almacenar todos los datos estructurados y no estructurados a cualquier escala.
  - **Ventajas:**
    - **Flexibilidad para almacenar datos en su formato original.**
    - **Capacidad para manejar grandes volúmenes de datos diversos.**
    - **Ideal para análisis big data y machine learning.**

Almacenamiento y procesamiento de datos

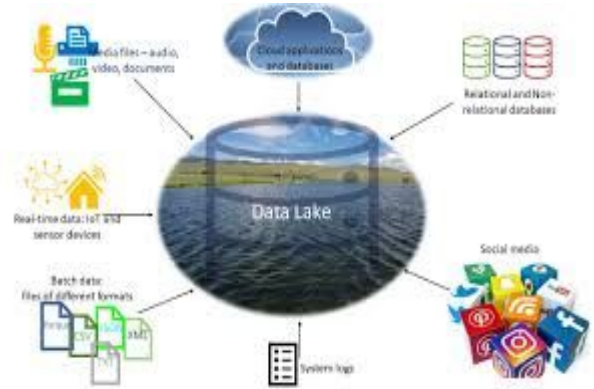
- **Data Lake**
  - Definición: Repositorio centralizado que permite almacenar todos los datos estructurados y no estructurados a cualquier escala.
  - Ventajas:
    - Flexibilidad para almacenar datos en su formato original.
    - Capacidad para manejar grandes volúmenes de datos diversos.
    - Ideal para análisis big data y machine learning.
  - **Almacenamiento de datos no estructurados del scraping:**
    - **Perfecto para guardar HTML crudo, textos no estructurados, imágenes.**
    - **Permite el análisis posterior con técnicas de big data.**

Almacenamiento y procesamiento de datos

Data Warehouse

Data Mart

Data Lake



Flujo de trabajo

Flujo de trabajo: del scraping al análisis

- **Extracción de datos mediante scraping**
  - **Diseño del scraper basado en la estructura del sitio web objetivo.**

Flujo de trabajo: del scraping al análisis

- **Extracción de datos mediante scraping**
  - Diseño del scraper basado en la estructura del sitio web objetivo.
  - **Implementación del scraper (código, configuración de herramientas).**

Flujo de trabajo: del scraping al análisis

- **Extracción de datos mediante scraping**
  - Diseño del scraper basado en la estructura del sitio web objetivo.
  - Implementación del scraper (código, configuración de herramientas).
  - **Ejecución y monitoreo del proceso de scraping.**

Flujo de trabajo: del scraping al análisis

- **Limpieza y transformación de datos**
  - **Identificación y manejo de datos faltantes o incorrectos.**

Flujo de trabajo: del scraping al análisis

- **Limpieza y transformación de datos**
  - Identificación y manejo de datos faltantes o incorrectos.
  - **Normalización y estandarización de formatos (fechas, números, etc.).**

Flujo de trabajo: del scraping al análisis

- **Limpieza y transformación de datos**
  - Identificación y manejo de datos faltantes o incorrectos.
  - Normalización y estandarización de formatos (fechas, números, etc.).
  - **Eliminación de duplicados y datos irrelevantes.**

Flujo de trabajo: del scraping al análisis

- **Carga en el sistema de almacenamiento apropiado**
  - **Elección del sistema basado en las necesidades de análisis:**
    - **Data Warehouse para datos estructurados y análisis histórico.**

Flujo de trabajo: del scraping al análisis

- **Carga en el sistema de almacenamiento apropiado**
  - **Elección del sistema basado en las necesidades de análisis:**
    - Data Warehouse para datos estructurados y análisis histórico.
    - **Data Mart para análisis departamental específico.**

Flujo de trabajo: del scraping al análisis

- **Carga en el sistema de almacenamiento apropiado**
  - **Elección del sistema basado en las necesidades de análisis:**
    - Data Warehouse para datos estructurados y análisis histórico.
    - Data Mart para análisis departamental específico.
    - **Data Lake para datos diversos y análisis futuro no definido.**

Flujo de trabajo: del scraping al análisis

- **Carga en el sistema de almacenamiento apropiado**
  - Elección del sistema basado en las necesidades de análisis:
    - Data Warehouse para datos estructurados y análisis histórico.
    - Data Mart para análisis departamental específico.
    - Data Lake para datos diversos y análisis futuro no definido.
  - **Proceso ETL (Extract, Transform, Load) o ELT (Extract, Load, Transform).**

Flujo de trabajo: del scraping al análisis

- **Análisis y visualización de datos**
  - **Aplicación de técnicas estadísticas y de machine learning.**

Flujo de trabajo: del scraping al análisis

- **Análisis y visualización de datos**
  - Aplicación de técnicas estadísticas y de machine learning.
  - **Creación de dashboards y reportes interactivos.**

Flujo de trabajo: del scraping al análisis

- **Análisis y visualización de datos**
  - Aplicación de técnicas estadísticas y de machine learning.
  - Creación de dashboards y reportes interactivos.
  - **Identificación de patrones, tendencias y anomalías en los datos.**

CONCLUSIÓN

CONCLUSIÓN

- **Importancia del web scraping en la era del big data.**

CONCLUSIÓN

- Importancia del web scraping en la era del big data.
- **Consideraciones éticas y técnicas en la implementación.**

CONCLUSIÓN

- Importancia del web scraping en la era del big data.
- Consideraciones éticas y técnicas en la implementación.
- **Valor del almacenamiento adecuado para maximizar el potencial de los datos.**

FUTURO DEL WEBSCRAPING

- **Tendencias emergentes (ej: scraping de apps móviles, IoT).**

FUTURO DEL WEBSCRAPING

- Tendencias emergentes (ej: scraping de apps móviles, IoT).
- **Desafíos futuros (ej: aumento de medidas anti-scraping, regulaciones).**

FUTURO DEL WEBSCRAPING

- Tendencias emergentes (ej: scraping de apps móviles, IoT).
- Desafíos futuros (ej: aumento de medidas anti-scraping, regulaciones).
- **Oportunidades para innovación y nuevos casos de uso.**

GRACIAS!