# UNIVERSIDAD MAYOR REAL Y PONTIFICA DE SAN FRANCISCO XAVIER DE CHUQUISACA

# FACULTAD DE CIENCIAS Y TECNOLOGÍA



**DOCUMENTACION WEB SCRAPING**

**CARRERA :** Ing. En Ciencias de la Computación

**MATERIA :** Inteligencia Artificial ll

**DOCENTE :** Ing. Pacheco L. Carlos W.

**UNIVERSITARIO :** Vidaurre Mejía Christian Paul

Sucre-Bolivia

# Web Scraping en el Campo de la Inteligencia Artificial

## Introducción

El Web Scraping es una técnica automatizada para extraer datos de páginas web. Su relevancia en el campo de la Inteligencia Artificial (IA) radica en la capacidad de generar datos estructurados esenciales para entrenar modelos, mejorar la precisión y potenciar el rendimiento de los sistemas de IA.

## Propósito

* Proveer un panorama general sobre el Web Scraping.
* Identificar sus aplicaciones en la IA.
* Describir herramientas, tecnologías y consideraciones éticas involucradas.

## Conceptos Fundamentales

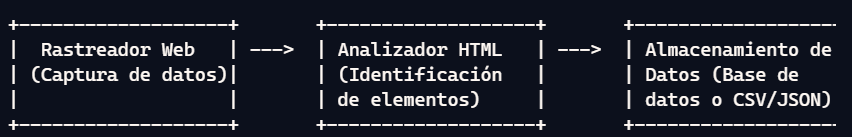
### ¿Qué es el Web Scraping?

El Web Scraping es el proceso automatizado de extracción de datos de páginas web, transformándolos en un formato estructurado para su análisis y aplicación.

1. Extracción Automática: Uso de software para recopilar datos de sitios web.
2. Formato Estructurado: Conversión de los datos a formatos como CSV o JSON para su análisis.
3. Aplicaciones Diversas:

* Análisis de mercado.
* Investigación académica.
* Desarrollo y entrenamiento de modelos de IA.

### Arquitectura del Sistema



1. Captura de Datos: Rastreo y recolección del contenido HTML.
2. Análisis de HTML: Identificación de los elementos clave.
3. Extracción de Datos: Almacenamiento de la información en estructuras como listas o diccionarios.
4. Almacenamiento y Procesamiento: Uso en bases de datos o integración en sistemas de IA.

### Aplicaciones del Web Scraping en IA

El Web Scraping es un pilar fundamental en múltiples aplicaciones dentro del campo de la IA, incluyendo:

* Entrenamiento de Modelos: Obtención de grandes volúmenes de datos para algoritmos de aprendizaje automático y redes neuronales.
* Análisis de Sentimientos: Extracción de opiniones y comentarios en redes sociales para estudiar percepciones públicas.
* Investigación y Desarrollo: Generación de datos para estudios avanzados en áreas como visión por computadora y procesamiento de lenguaje natural.
* Automatización de Procesos: Simplificación de tareas repetitivas como el monitoreo de precios o el seguimiento de noticias.
* Análisis de Tendencias: Monitoreo de mercados para detectar patrones emergentes, Identificación de áreas de oportunidad para la innovación.

### Tipos de Datos para Entrenamiento de IA

Los datos extraídos a través del Web Scraping pueden clasificarse en:

* Textuales: Reviews, comentarios y artículos.
* Numéricos: Precios, estadísticas e indicadores económicos.
* Visuales: Imágenes, videos y otros contenidos multimedia.

### Herramientas y Tecnologías

* Lenguajes de Programación: Python, JavaScript, Java.
* Bibliotecas de Web Scraping: Beautiful Soup, Scrapy, Selenium.
* Herramientas de Extracción: Octoparse, ParseHub, Import.io.
* Almacenamiento de Datos: Amazon S3, Google Cloud Storage, MongoDB.

### Proceso de Construcción

1. Definición de Objetivos:

* Establecer metas claras para el proyecto.
* Identificar las páginas web relevantes y los datos necesarios.

1. Análisis de la Estructura Web:

* Examinar el código HTML y CSS de las páginas.
* Identificar elementos clave como etiquetas, clases y atributos.

1. Desarrollo del Sistema:

* Seleccionar el lenguaje y las herramientas adecuadas.
* Implementar scripts que extraigan y estructuren los datos.

1. Pruebas y Refinamiento:

* Realizar pruebas iterativas para corregir errores.
* Optimizar el rendimiento del sistema.

1. Integración con IA:

* Preparar los datos para el entrenamiento de modelos.
* Utilizar algoritmos específicos para extraer valor de los datos.

### Consideraciones Éticas y Legales

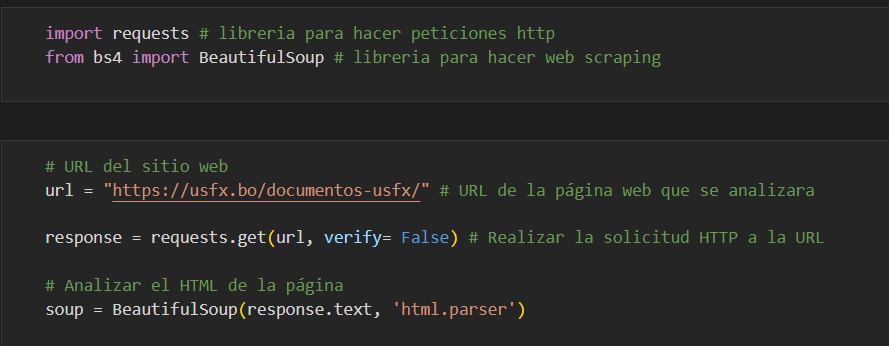
Privacidad: Respetar la confidencialidad de los usuarios y evitar la recopilación no autorizada.

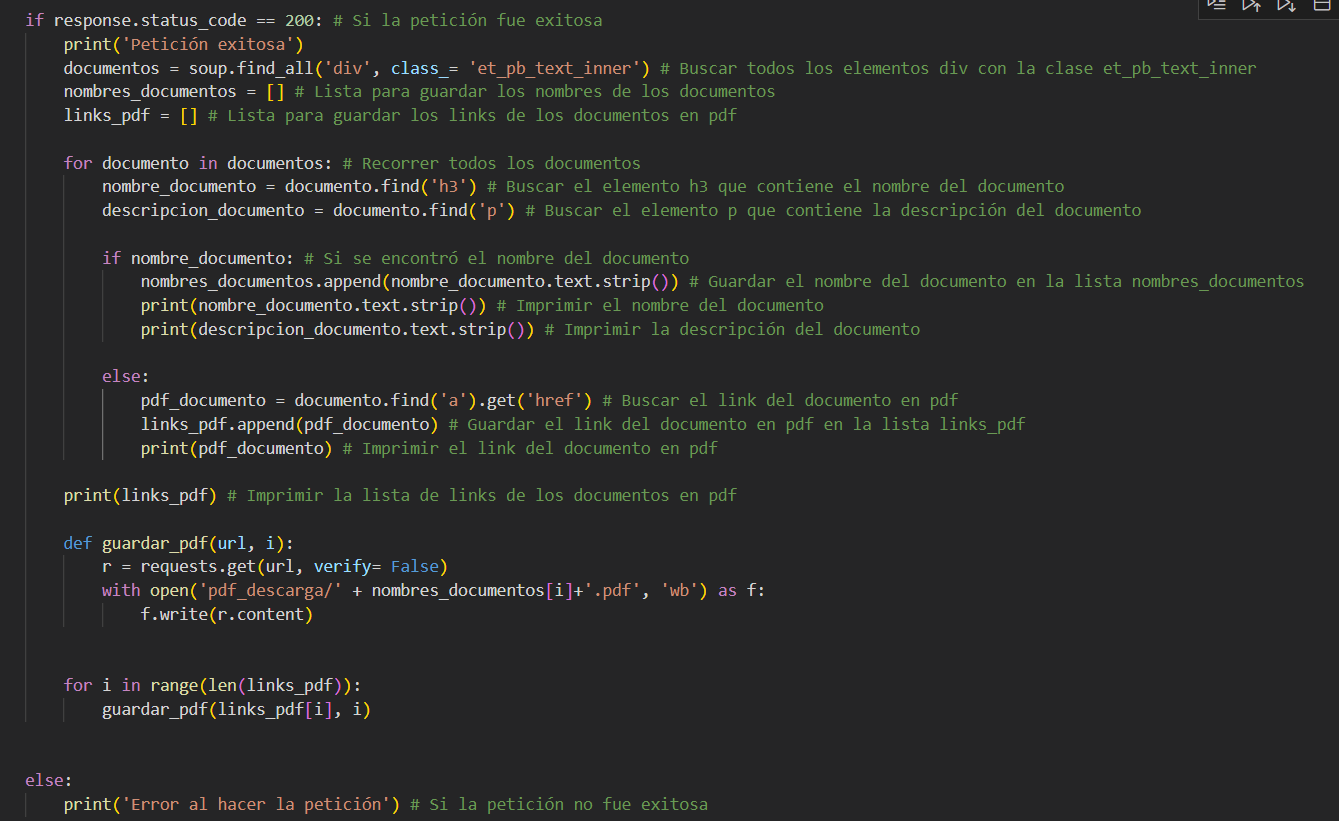
Legislación: Cumplir con las leyes de propiedad intelectual y términos de uso.

Ética: Aplicar prácticas responsables para minimizar impactos negativos en los sitios web, Evitar prácticas abusivas que puedan impactar negativamente en los servidores de los sitios web, Garantizar un uso responsable y beneficioso del Web Scraping.

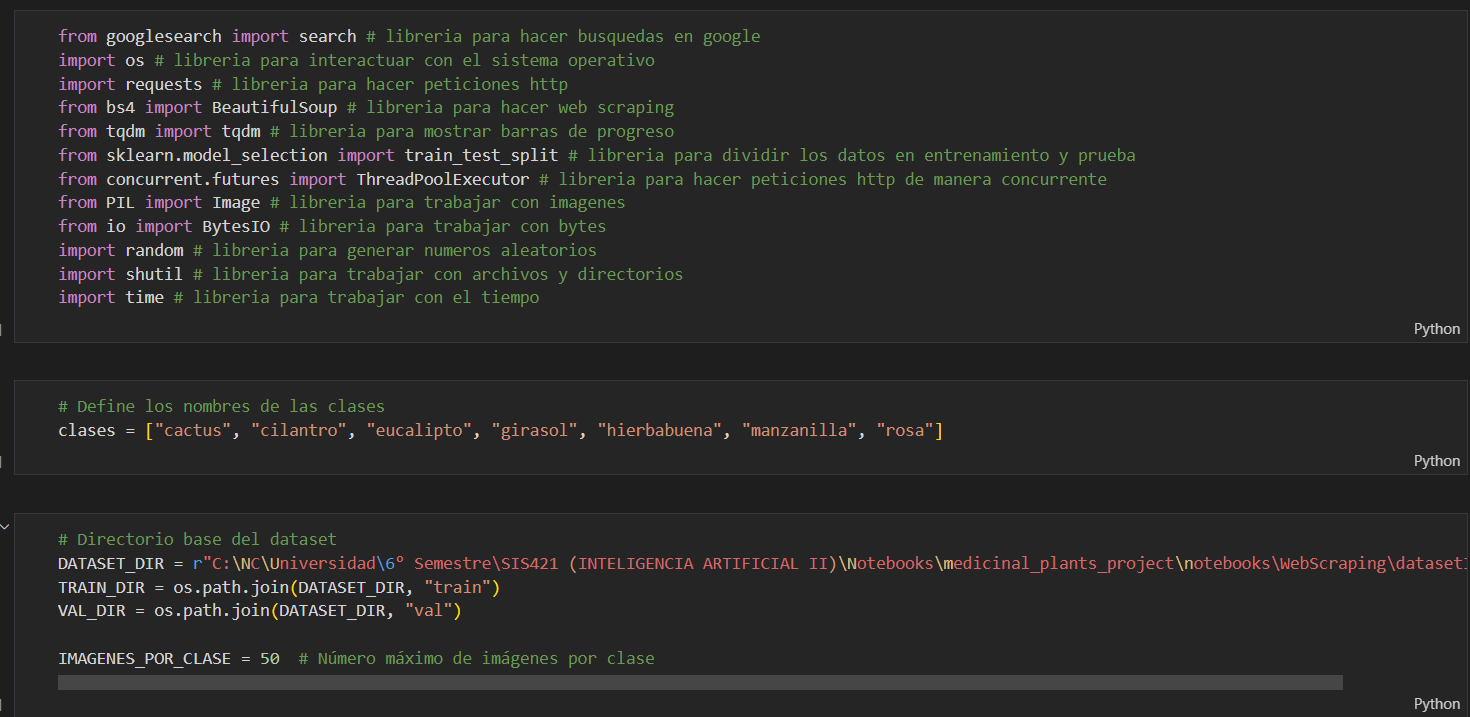
### Ejemplos Practicos

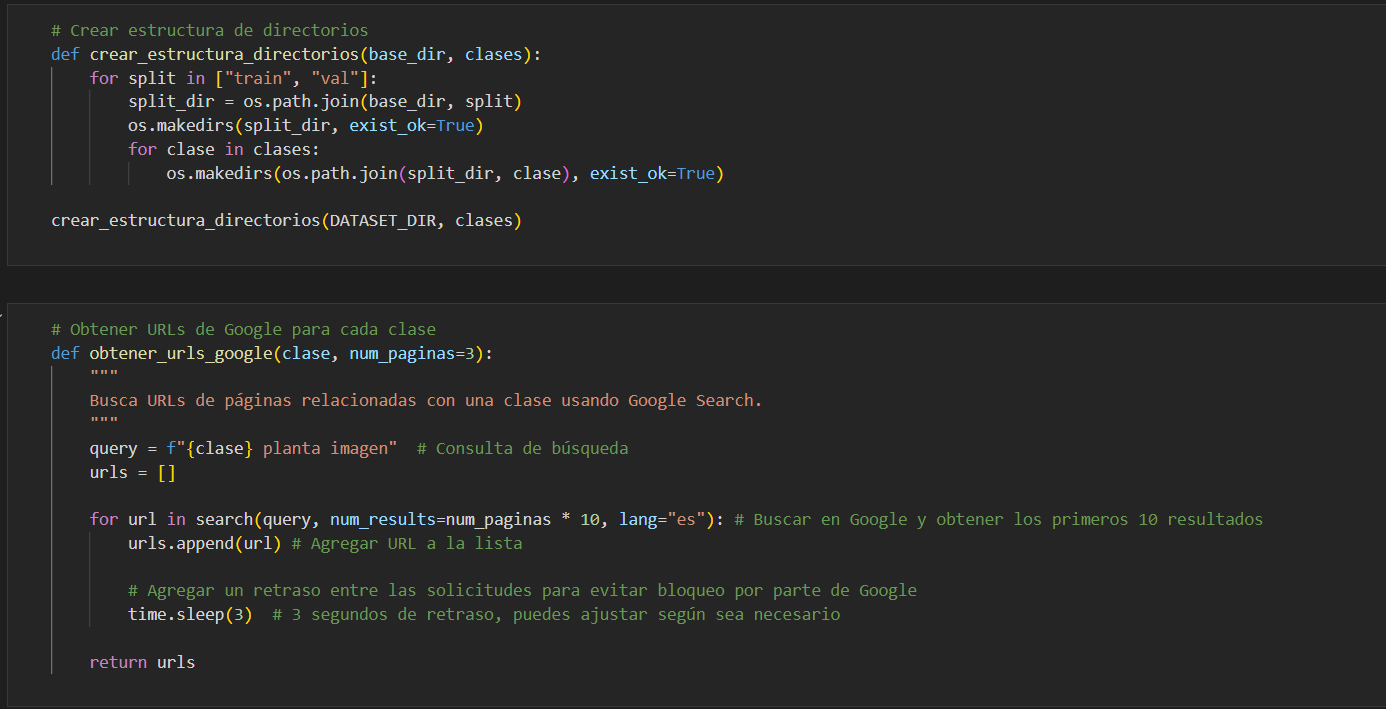
#### Extracción de documentación de la universidad en formato pdf de la página de la universidad

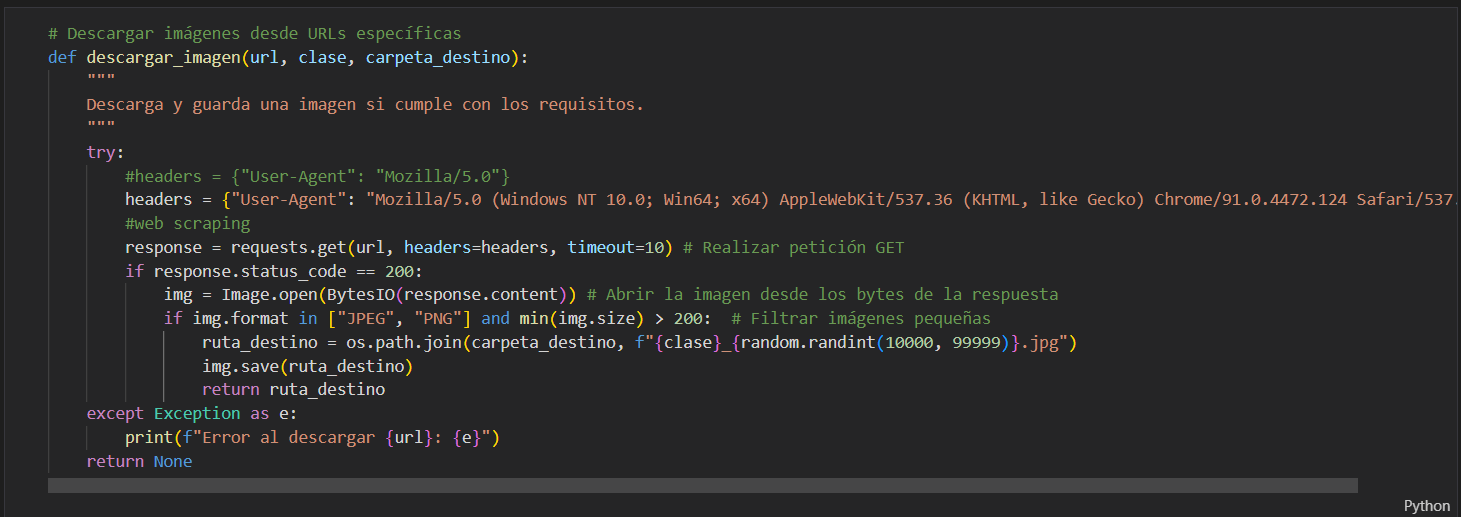


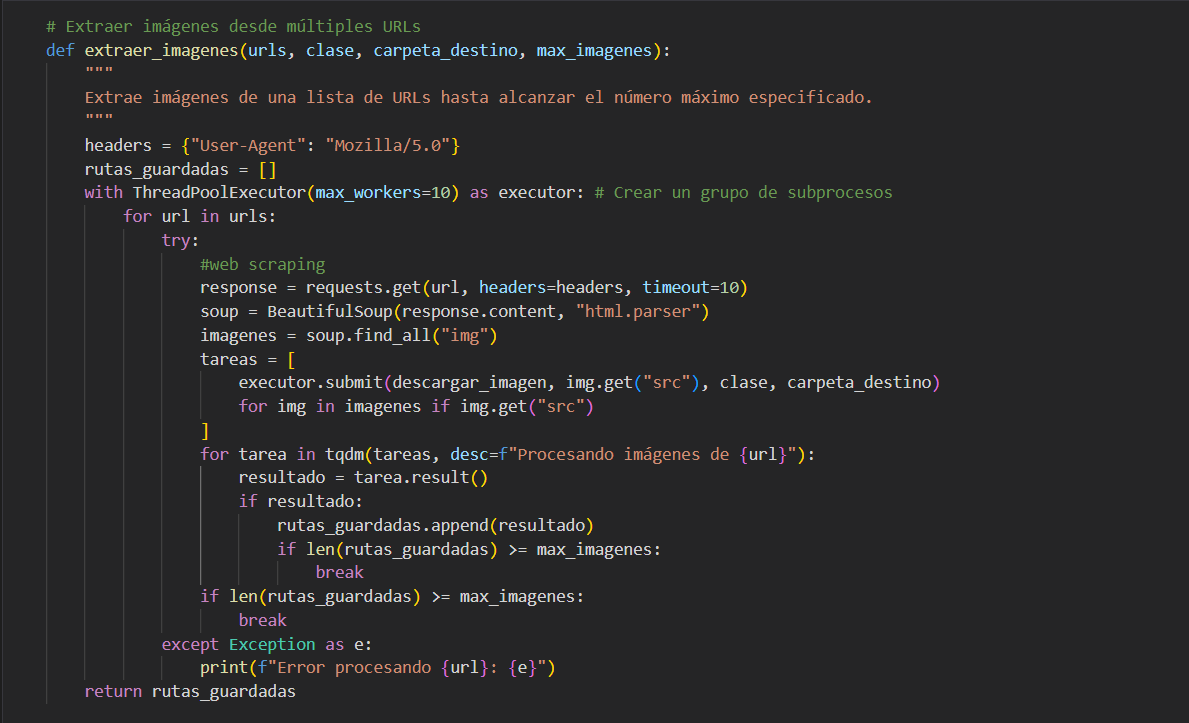


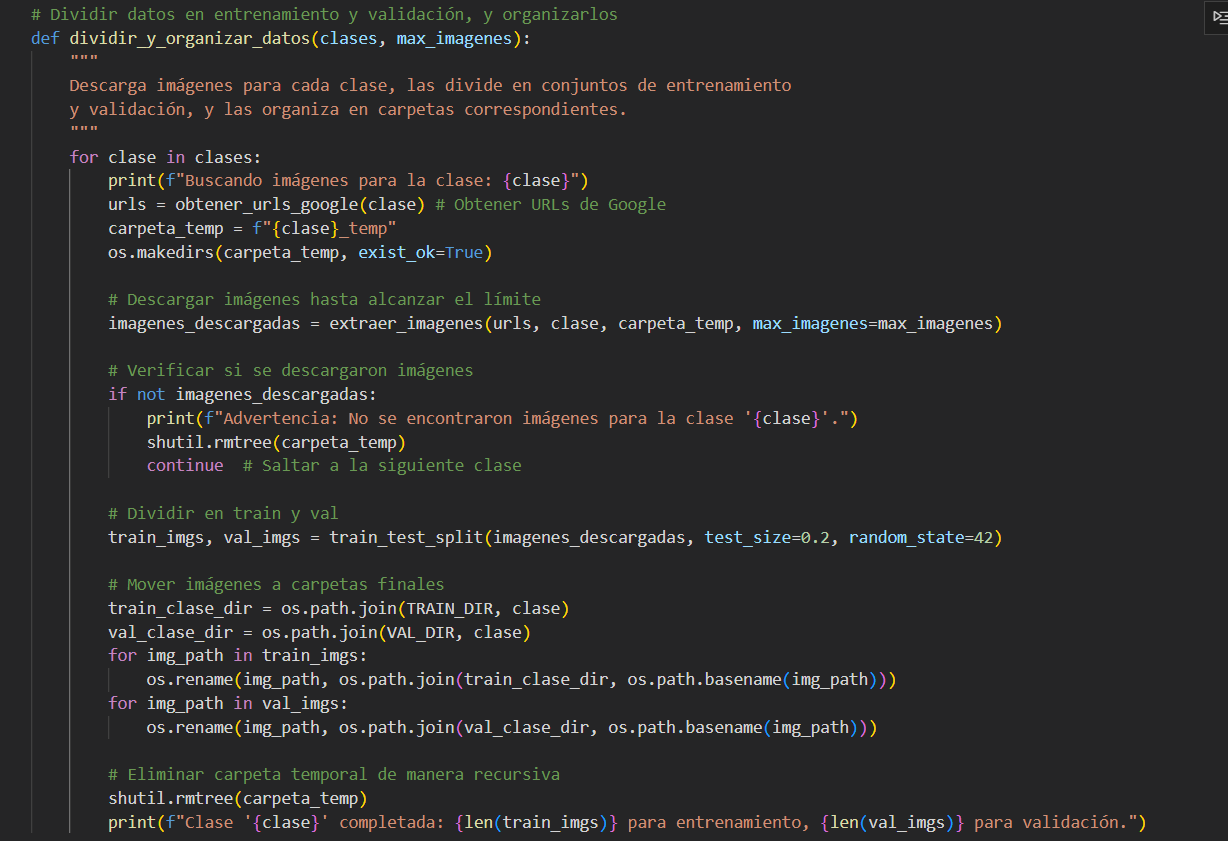
#### Recoleccion de imágenes de plantas para un dataset

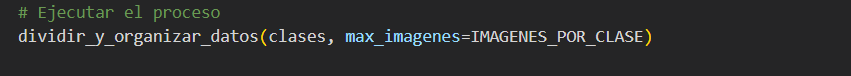












### Recomendaciones

Utilizar APIs oficiales como Graph API, Google Custom Search JSON API o Bing Image Search API para recopilar grandes volúmenes de datos.

Garantizar un uso ético del Web Scraping, evitando abusos y respetando la normativa vigente.

Implementar límites de frecuencia en las solicitudes para evitar problemas con los servidores.

### Conclusión

El Web Scraping es una herramienta indispensable en el ámbito de la Inteligencia Artificial debido a su capacidad para recopilar grandes volúmenes de datos estructurados de manera eficiente. Estos datos son fundamentales para el desarrollo de modelos más precisos y funcionales, mejorando el rendimiento en tareas como el análisis de sentimientos, la visión artificial y el procesamiento del lenguaje natural.

A pesar de sus numerosas ventajas, el uso del Web Scraping requiere un enfoque responsable. Es esencial equilibrar los objetivos técnicos con la ética y el cumplimiento legal, respetando la privacidad de los usuarios y las normativas de propiedad intelectual. Las malas prácticas, como el scraping excesivo o la recopilación de datos sensibles sin consentimiento, no solo pueden ser ilegales, sino también perjudicar la reputación del proyecto o la organización.

El futuro del Web Scraping promete ser aún más relevante a medida que crecen las necesidades de datos para entrenar sistemas avanzados de IA. Las herramientas y tecnologías continuarán evolucionando para facilitar procesos más eficientes, seguros y éticos. Sin embargo, es importante que los desarrolladores consideren alternativas oficiales, como APIs, cuando estén disponibles, y trabajen en armonía con las plataformas de las cuales obtienen información.

En resumen, el Web Scraping, cuando se emplea con ética y responsabilidad, se convierte en una poderosa técnica para potenciar la innovación y el desarrollo tecnológico. Es crucial educar a los profesionales y organizaciones sobre las mejores prácticas, fomentando un uso que sea tanto beneficioso como sostenible en el tiempo.

### Bibliografía

Web Scraping: Challenges and Techniques (ResearchGate): <https://www.researchgate.net/publication/334165092_Web_Scraping_Challenges_and_Techniques>

Ethical Implications of Web Scraping (SpringerLink): <https://link.springer.com/article/10.1007/s10676-019-09517-5>

Web Scraping Techniques and Its Applications: A Review:

<https://ieeexplore.ieee.org/document/10351298>

Documentacion librerias de Web Scraping:

Beautiful Soup:

<https://www.crummy.com/software/BeautifulSoup/>

Scrapy:

<https://scrapy.org/>

Selenium:

<https://www.selenium.dev/>