Temas Selectos De Ciencia De Datos

Lic. Alondra Vanianinetl Berzunza Rodríguez

EVALUACIÓN

alondraberzunza09@gmail.com

Actividad	Porcentaje
Exámenes Parciales	25 %
Exposiciones	15 %
Participaciones y tareas	10 %
Portafolio (OBLIGATORIO)	50 %
Total	**100 %**
Examen Final	100 %

Calendario

- Última clase 22 de mayo
- 26 de Mayo 1er final
- 2 de Junio 2do final

Las tareas se revisarán a la siguiente clase

Los exámenes se realizarán al término de cada unidad

Las prácticas y exposiciones se realizarán en equipos*

TEMARIO

1. Web scraping.
2. Aprendizaje por refuerzo.
3. Ciberseguridad y protección de datos.
4. Aprendizaje profundo.
5. Algoritmos genéticos.

UNIDAD I - Web Scraping

1.1. Conceptos básicos.

1.2. Aplicaciones del web scraping y herramientas.

1.3. Python y web scraping.

1.4. Diseño de web scraper.

1.1. Conceptos básicos

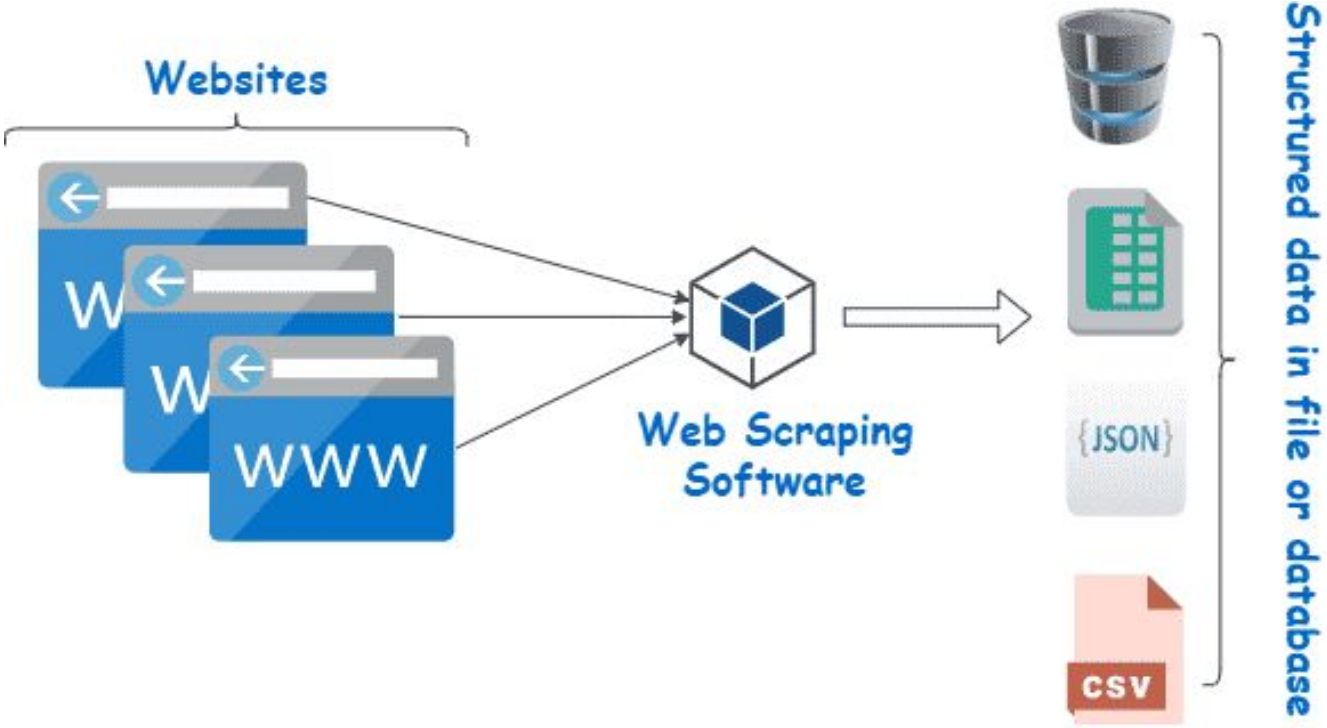
1.1. Conceptos básicos.

- Proceso de extracción de contenidos y datos de sitios web mediante software.
- Es un conjunto de prácticas utilizadas para extraer automáticamente o “scrapear” datos de la web.
- Es una HERRAMIENTA útil para la recopilación de datos online.

Google

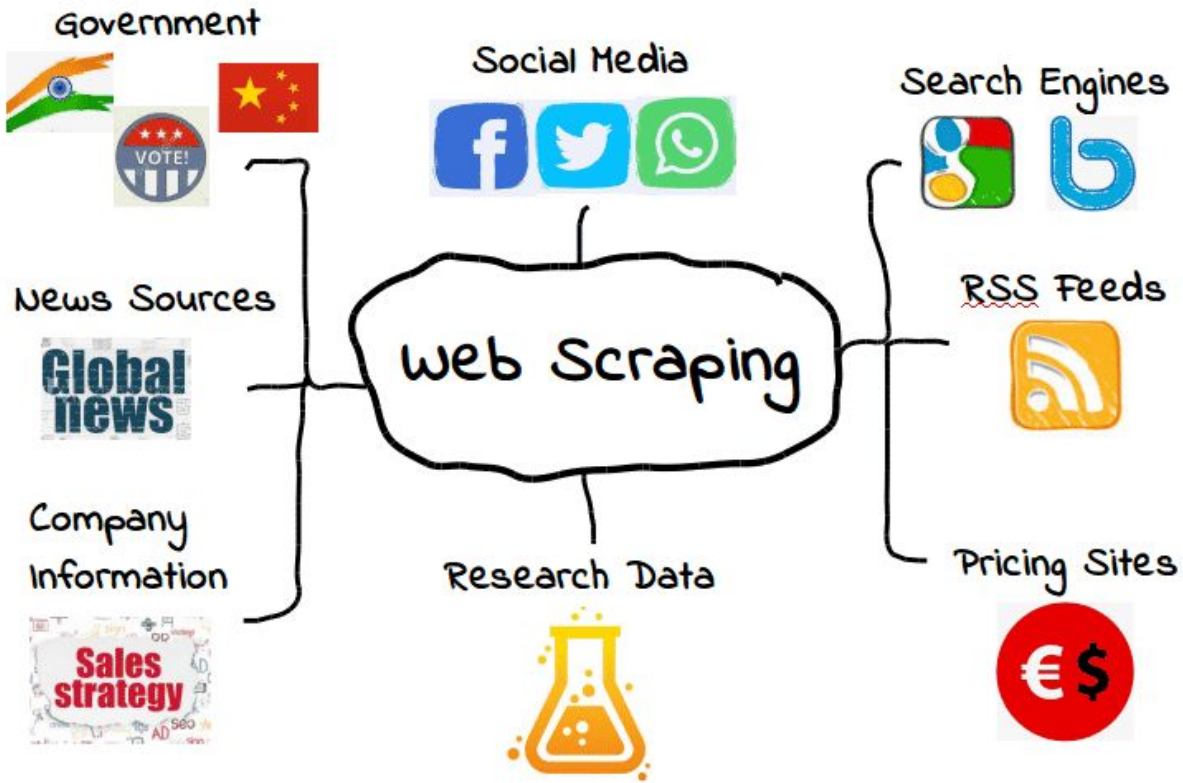
Scraping de contenidos
Scraping de datos

1.1. Conceptos básicos.



1.1. Conceptos básicos.

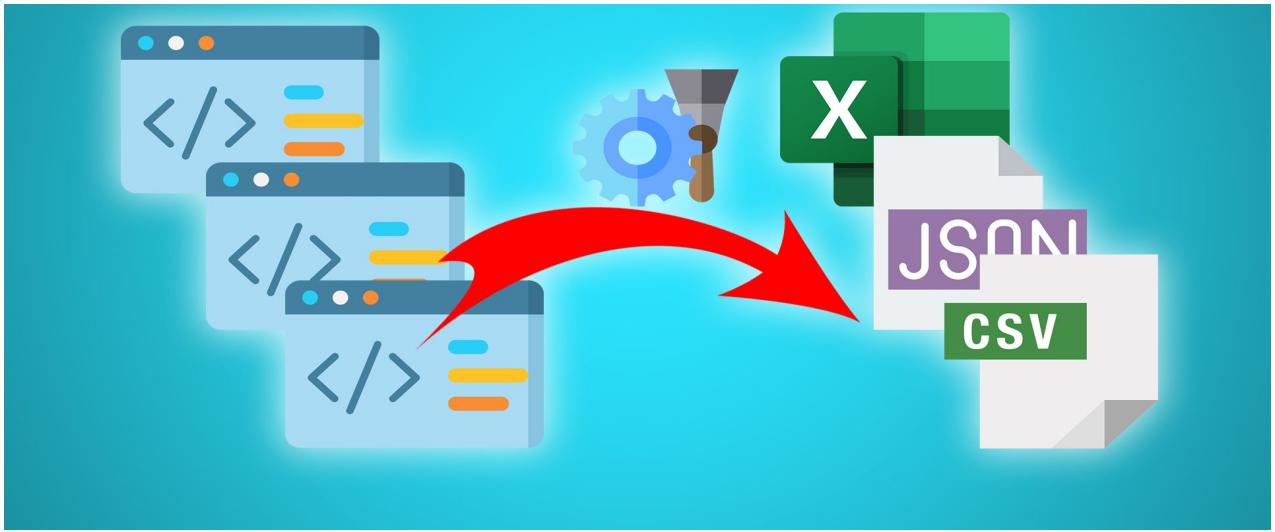
Es posible scrapear todo tipo de datos de la web, desde los motores de búsqueda y los feeds RSS hasta la información gubernamental, la mayoría de los sitios web ponen sus datos a disposición de los scrapers, crawlers y otras formas de recopilación automática de datos.



1.1. Conceptos básicos.

¡Los datos no siempre están dsponibles!

El web scraping se usa solo cuando los datos no están disponibles o cuando no están en un formato adecuado.



1.1. Conceptos básicos.

¿Es Legal el Web Scraping?

No hay nada intrínsecamente ilegal en el web scraping. Cuando se publican datos en la web normalmente están disponibles al público.

No todos los datos de la web están diseñados para el público.

WEB SCRAPING MALICIOSO

AVISO DE RETIRADA DE LA DMCA

1.1. Conceptos básicos.

WEB SCRAPING MALICIOSO

- Es el web scraping que el editor no pretendía o no consintió compartir.

- Puede aplicarse a cualquier cosa que no esté destinada al público, aunque no sean datos personales o de propiedad intelectual.

- Aunque los datos no estén protegidos de forma legal (GDPR), no signiica que sea legal su scrapeado.

1.1. Conceptos básicos.

Aunque el web scraping es definitivamente legal, puede utilizarse fácilmente con fines maliciosos o poco éticos.

A muchos proveedores de servicios web no les gusta que sus datos sean scrapeados, independientemente de que sea legal.

ZONA GRIS

Supongamos que un alojamiento web pone «*accidentalmente*» a disposición del público la información de sus usuarios. Eso podría incluir una lista completa de nombres, correos electrónicos y otra información que es técnicamente pública, pero que tal vez no estaba destinada a ser compartida.

Aunque también sería *técnicamente* legal scrapear estos datos, probablemente no sea la mejor idea. El hecho de que los datos sean públicos no significa necesariamente que el administrador de la web haya consentido que se hayan scrapeado, aunque su falta de supervisión los haya hecho públicos.

1.1. Conceptos básicos.

Tipos de WEB SCRAPING MALICIOSO

- OVER SCRAPING
  - Se envían excesivas solicitudes en poco tiempo para sobrecargar los servidores del sitio web y así afectar al rendimiento y experiencia del usuario.

- SCRAPERS PING
  - Perjudican a las visitas web degradando el rendimiento de sitio para que sus consumidores abandonen.

1.1. Conceptos básicos.

PREVENCIÓN DEL WEB SCRAPING

Existen productos en el mercado capaces de detectar y mitigar el web scraping evasivo. Unos productos que incluyen características como:

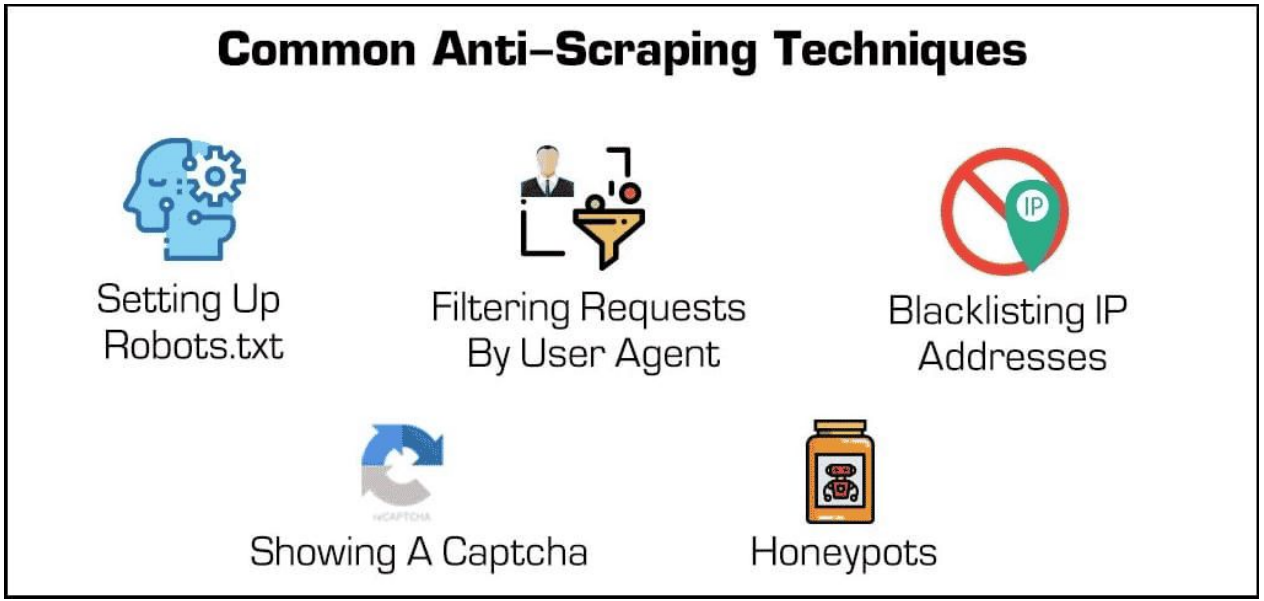
- evaluación a nivel de protocolo
- evaluación a nivel de aplicación
- evaluación de la interacción y comportamiento del usuario
- clasificación de riesgo.

1.1. Conceptos básicos.

MEDIDAS ADICIONALES

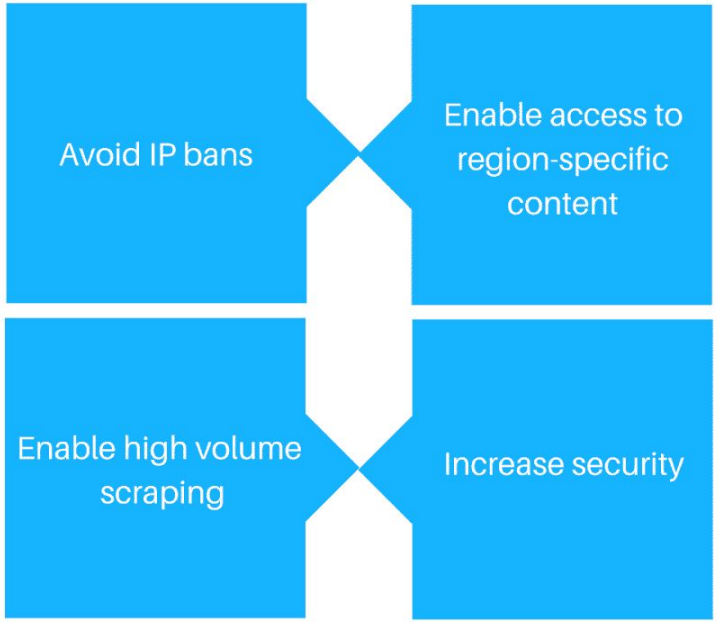
- Usar una red privada virtual (VPN)
- Evitar compartir datos personales en línea
- Evitar malas prácticas

(capa 3)



1.1. Conceptos básicos.

EVITAR LA PREVENCIÓN DEL WEB SCRAPING

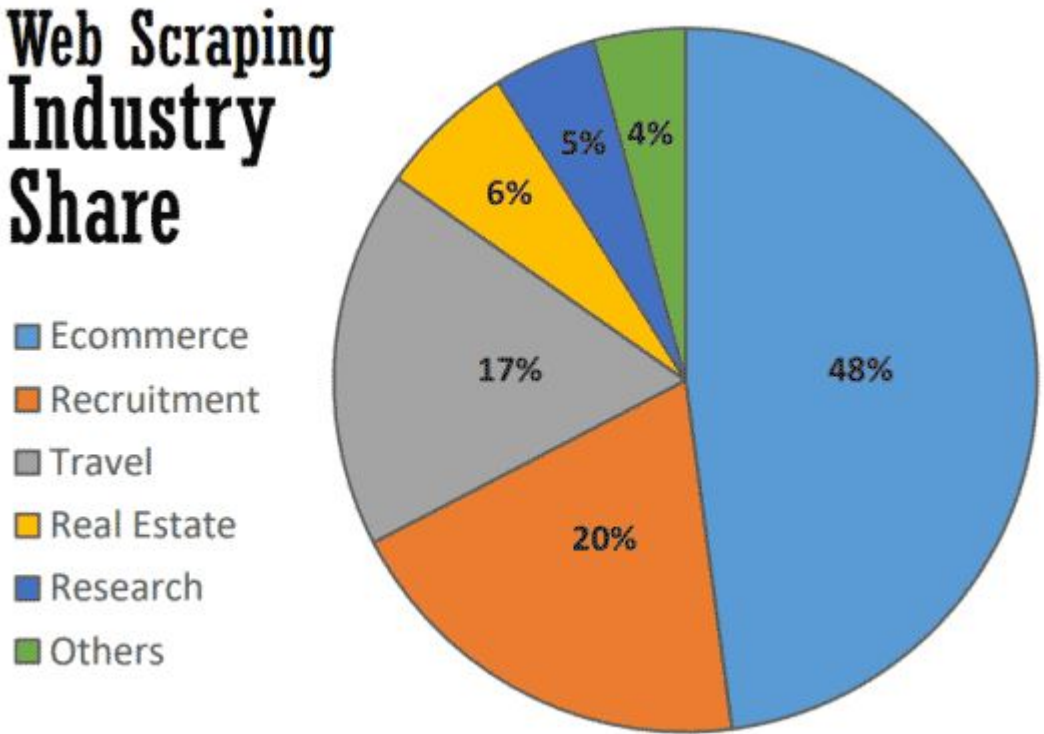


1.2 Aplicaciones y Herramientas

Aplicaciones

1.2 Aplicaciones y Herramientas

Casi la mitad del web scraping se utiliza para reforzar las estrategias de comercio electrónico.



1.2 Aplicaciones y Herramientas

ESTUDIO DE MERCADO

El web scraping se ha convertido en una herramienta inestimable para los equipos de marketing que buscan vigilar su mercado sin tener que realizar una investigación manual que requiere mucho tiempo.

¿Qué hacen tus clientes? ¿Y tus clientes potenciales? ¿Cómo son los precios de tus competidores en comparación con los tuyos? ¿Tienes información para crear una campaña exitosa?

1.2 Aplicaciones y Herramientas

AUTOMATIZACIÓN DE NEGOCIO

Cuando muchas tareas de automatización empresarial requieren la recopilación y el procesamiento de grandes cantidades de datos, el web scraping puede ser muy valioso.

Por ejemplo, supongamos que se necesita reunir datos de diez sitios web diferentes. Aunque se extraiga el mismo tipo de datos de cada uno, cada sitio web puede requerir un método de extracción diferente.

1.2 Aplicaciones y Herramientas

GENERACIÓN DE LEADS

El web scraping también puede generar valiosas listas de clientes potenciales con poco esfuerzo.

Aunque es necesario establecer tus objetivos con cierta precisión, puedes utilizar el web scraping para generar suficientes datos de usuarios para crear listas de leads estructuradas.

1.2 Aplicaciones y Herramientas

SEGUIMIENTO DE PRECIOS

La extracción de precios — también conocida como scraping de precios — es una de las aplicaciones más comunes del web scraping.

Ej. Camelcamelcamel. La aplicación extrae regularmente los precios de los productos y luego los compara en un gráfico a lo largo del tiempo.

Amazon Price History

Date Range

1m 3m 6m 1y All

Price Type

- [x] Amazon
- [x] 3rd Party New
- [ ] 3rd Party Used

Chart Options

- [x] Close-up View
- [ ] Remove Extreme Values

1.2 Aplicaciones y Herramientas

Los precios pueden fluctuar mucho, incluso a diario. Con el acceso a las tendencias históricas de los precios, los usuarios pueden comprobar si el precio que están pagando es el ideal.

A pesar de su utilidad, el scraping de precios conlleva cierta controversia. Como mucha gente quiere actualizaciones de precios en tiempo real, algunas aplicaciones de seguimiento de precios se convierten rápidamente en maliciosas al sobrecargar ciertos sitios web con peticiones al servidor.

Como resultado, muchos sitios web de comercio electrónico han empezado a tomar medidas adicionales para bloquear totalmente a los web scraping.

1.2 Aplicaciones y Herramientas

NOTICIAS Y CONTENIDOS

El web scraping es una valiosa herramienta para mantenerse informado.

Aunque algunos sitios web de noticias y blogs ya ofrecen canales RSS y otras interfaces sencillas, no siempre son la norma — ni son tan comunes como antes.

1.2 Aplicaciones y Herramientas

MONITORIZACIÓN DE MARCAS

En el caso de las marcas que reciben mucha cobertura informativa, el web scraping es una herramienta inestimable para estar al día sin tener que revisar innumerables artículos y sitios de noticias.

El web scraping también es útil para comprobar el precio mínimo disponible de un producto o servicio de una marca (MAP).

1.2 Aplicaciones y Herramientas

SECTOR INMOBILIARIO

Con miles de anuncios dispersos en múltiples sitios web inmobiliarios, puede ser difícil encontrar exactamente lo que buscas.

1.2 Aplicaciones y Herramientas

Muchos sitios web utilizan el «web scraping» para agregar listados inmobiliarios en una única base de datos para facilitar el proceso.

Sin embargo, la agregación de listados no es el único uso del web scraping en el sector inmobiliario. Por ejemplo, los agentes inmobiliarios pueden utilizar las aplicaciones de scraping para estar al tanto de los precios medios de alquiler y venta, los tipos de propiedades que se venden y otras tendencias valiosas.

Herramientas

1.2 Aplicaciones y Herramientas

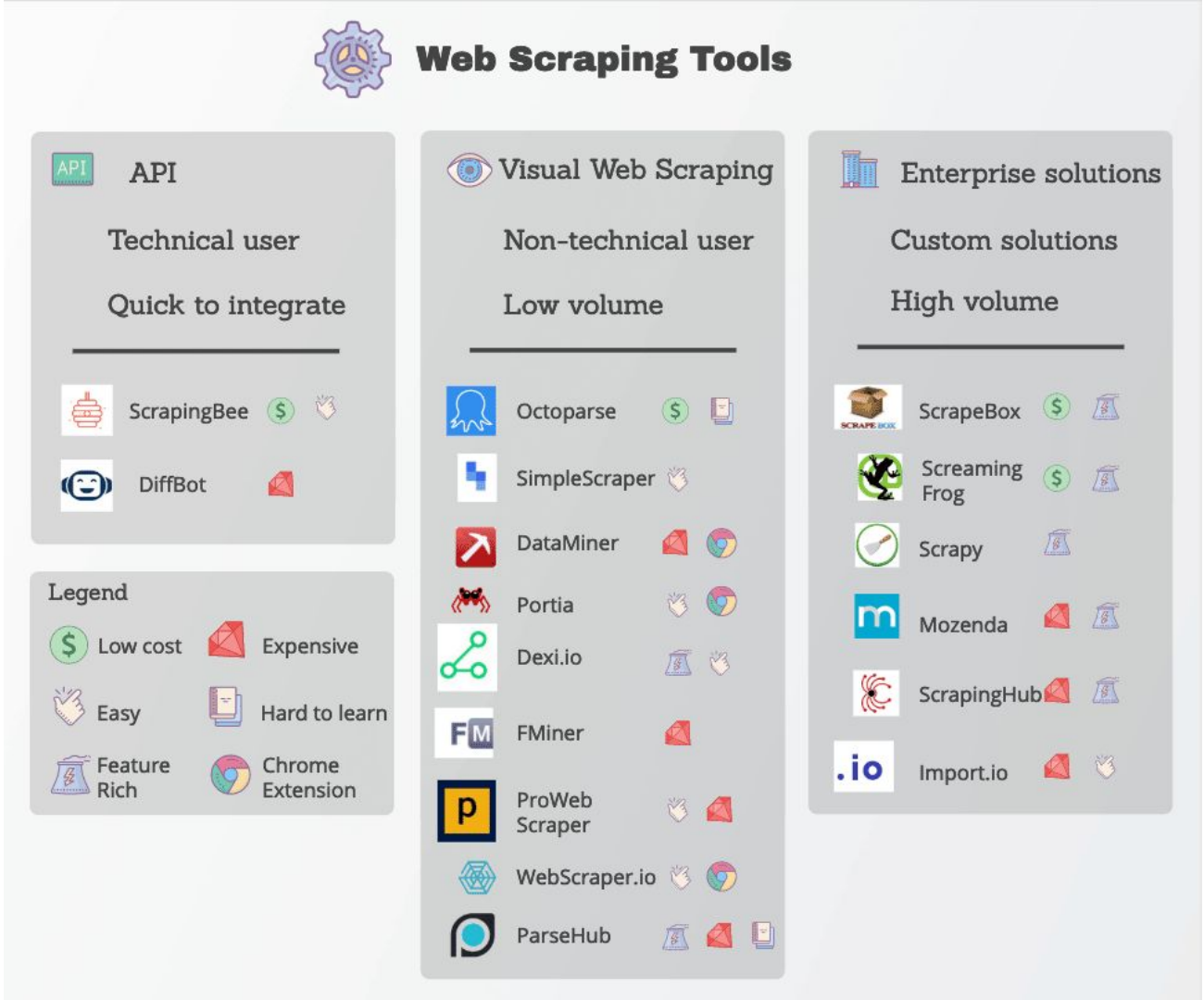
Muchas funciones de web scraping están disponibles en forma de herramientas de web scraping. Aunque hay muchas herramientas disponibles, varían mucho en cuanto a calidad, precio y ética.

Un buen web scraper será capaz de extraer de forma fiable los datos que necesitas sin toparse con demasiadas medidas anti-scraping.

1.2 Aplicaciones y Herramientas

Algunas características a considerar:

- Localizadores precisos.
- Calidad de los datos.
- Entrega de datos.
- Manejo del anti-scraping.
- Precios transparentes.
- Asistencia al cliente.



Octoparse
Import.io
ParseHub

1.3. Python y web scraping.

1.3. Python y web scraping.

¿Cómo Funciona el Scraping Web?

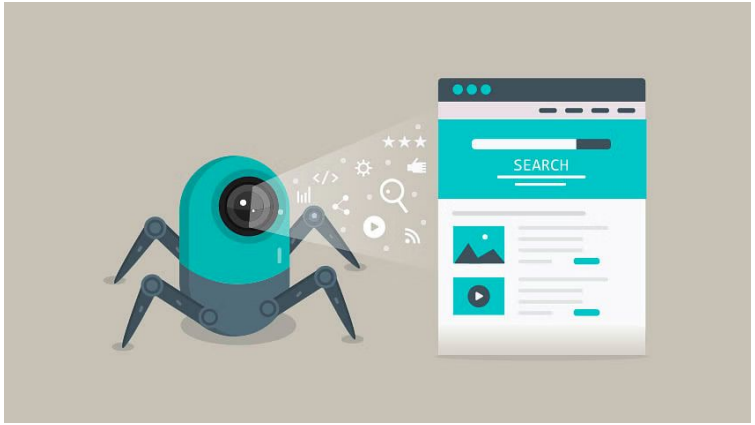
El scraping web puede parecer complicado, pero en realidad es muy sencillo.

Aunque los métodos y las herramientas pueden variar, todo lo que se tiene que hacer es encontrar una manera de

1. Navegar automáticamente por el/los sitio/sitios web de destino.
2. Extraer los datos una vez en el sitio web.

Normalmente, estos pasos se realizan con scrapers y crawlers.

1.3. Python y web scraping.



Crawlers

A veces también conocidos como arañas.

Son programas básicos que navegan por la web buscando e indexando contenidos. Aunque los crawlers (rastreadores) guían a los web scrapers, no se utilizan exclusivamente para este fin.

Por ejemplo, los motores de búsqueda como Google utilizan rastreadores para actualizar los índices y las clasificaciones de los sitios web.

Los rastreadores suelen estar disponibles como herramientas preconstruidas que permiten especificar un determinado sitio web o término de búsqueda.

1.3. Python y web scraping.

Scrapers

Hacen el trabajo sucio de extraer rápidamente la información relevante de los sitios web. Dado que los sitios web están estructurados en HTML, los scrapers utilizan expresiones regulares (regex), XPath, selectores CSS y otros localizadores para encontrar y extraer rápidamente determinados contenidos.

Por ejemplo, se puede dar a un web scraper una expresión regular que especifique el nombre de una marca o una palabra clave.

1.3. Python y web scraping.

En su nivel más básico, el web scraping se reduce a unos simples pasos:

1. Especifica las URLs de los sitios web y las páginas que quieres scrapear.
2. Haz una petición HTML a las URL (es decir, «visita» los sitios web).
3. Utiliza localizadores como expresiones regulares para extraer la información deseada del HTML.
4. Guarda los datos en un formato estructurado (como CSV o JSON).

Uno de los mayores retos del web scraping es mantener tu scraper actualizado a medida que los sitios web cambian de diseño o adoptan medidas anti-scraping. Aunque esto no es demasiado difícil si sólo se scrapean unos pocos sitios web a la vez, scrapear más puede convertirse rápidamente en una complicación.