Posibles problemas técnicos del Scraping

El raspado web a menudo puede parecer una búsqueda del tesoro en la que se explora la web en busca de información oculta que no proporcionan las API. Y como en toda buena búsqueda del tesoro, hay retos que superar.

Un obstáculo notable es encontrarse con bloqueos de acceso impuestos por el sitio web de destino. Estos bloqueos pueden surgir por varias razones, como políticas de raspado estrictas, preocupaciones relacionadas con el abuso de recursos, problemas de reputación de la IP de origen o la detección de agentes de usuario (falsos)

1. Entender las políticas y condiciones de servicio de su objetivo

Cuando empiece a raspar un sitio nuevo, deberá familiarizarse con él más allá de aprender la estructura HTML de la página. La familiarización también debe incluir la comprensión de las políticas y términos de servicio del sitio que pretende raspar. Esto suele implicar cuál es la postura del sitio con respecto al raspado web, si permite el raspado y qué páginas específicas se le permite raspa

2. Adherirse a las normas éticas de raspado

- **No bombardee los servidores con peticiones incesantes**: Deje un intervalo de tiempo suficiente entre las solicitudes. Algunos sitios web pueden detectar y bloquear a los raspadores web que extraen grandes cantidades de datos rápidamente porque no parece un comportamiento humano.
- **No raspe datos personales sin consentimiento**: No se trata sólo de una cuestión ética, sino también legal. Asegúrate siempre de contar con los permisos necesarios antes de raspar datos personales.
- **Respeto de los datos que obtenga**: Utilice los datos que raspe de forma responsable y legal. Procure que el uso que haga de los datos se ajuste a todas las leyes y normativas aplicables, como las leyes de derechos de autor y el Reglamento General de Protección de Datos (GDPR).

3. Utilizar proxy (rotativos)

Una herramienta útil en su kit de herramientas de raspado web es el proxy Una de las formas más fáciles de desbloquear sitios web es usar un proxy web público. Puede que no sea tan rápido ni tan seguro como una VPN, pero un proxy web público es una buena opción cuando utiliza equipos públicos donde no tiene permiso para instalar una VPN. Los proxy ocultan su dirección IP y redirigen su tráfico de Internet a través de distintos servidores y direcciones públicas diferentes.

4. Utilice las cabeceras y los agentes de usuario adecuados

Los sitios web suelen utilizar cabeceras y agentes de usuario para detectar bots. Un User-Agent es una cabecera que su navegador envía al servidor, proporcionando detalles sobre el software y el sistema que inicia las peticiones. Suele incluir el tipo de aplicación, el sistema operativo, el proveedor y la versión del software. Esta información ayuda al servidor a ofrecer contenidos adecuados para su navegador y sistema específicos.

Cuando se raspa la web, es crucial emplear cadenas de agente de usuario legítimas. Al imitar a un usuario real, es posible eludir eficazmente los mecanismos de detección y reducir la probabilidad de ser bloqueado.

5. Manejar trampas y errores de Honeypot

Honeypots son enlaces invisibles para los visitantes normales, pero están en el código HTML y pueden ser encontrados por los raspadores web(web scraping). Son como trampas para detectar el raspador dirigiéndolos a páginas en blanco. Una vez que un visitante particular navega por una página honeypot, el sitio web puede estar relativamente seguro de que no es un visitante humano y comienza a limitar o bloquear todas las solicitudes de ese cliente.

6. Utilice un servicio de resolución de CAPTCHA

La Prueba Pública de Turing Completamente Automatizada para Distinguir a los Ordenadores de los Humanos (CAPTCHA) es una medida de seguridad implementada por muchos sitios web para prevenir actividades automatizadas de bots, incluyendo el raspado web. Están diseñados para que sean fáciles de resolver para los humanos, pero un reto para las máquinas, de ahí su nombre.

7. Supervise los límites de velocidad y las denegaciones de acceso

La mayoría de las actividades de web scraping tienen como objetivo obtener datos lo más rápido posible. Sin embargo, cuando un humano visita un sitio, la navegación será mucho más lenta en comparación con lo que sucede con el **web scraping**. Por lo tanto, es realmente fácil para el sitio captuarlo al rastrear su velocidad de acceso. Una vez que descubre que estás pasando por las páginas demasiado rápido, sospechará que no eres humano y te bloqueará de forma natural.

8. Raspado desde la caché de Google

En el caso de sitios web difíciles de raspar o de datos no sensibles al tiempo, un método alternativo consiste en raspar datos de la copia en caché de Google,

(mantiene copias locales de los sitios web para que se carguen más rápidamente cuando los visita. A esta técnica se la denomina «guardar en la memoria caché». Aunque tenga bloqueado el acceso a la versión original de un sitio, tal vez pueda conectarse a su versión en caché. Así es como se puede encontrar la versión en caché de un sitio o página en Chrome)
Esta técnica puede resultar especialmente útil cuando se trata de sitios web extremadamente difíciles que bloquean activamente los raspadores web. Estas páginas almacenadas en caché se pueden raspar en lugar de las páginas web originales para evitar que se activen los mecanismos antiraspado. Tenga en cuenta que este método puede no ser infalible, ya que algunos sitios web dan instrucciones a Google para que no almacene su contenido en caché. Además, es posible que los datos de la caché de Google no estén actualizados.

9. Utilizar proxies y servicios de raspado de terceros

A medida que se intensifica el juego del gato y el ratón entre los raspadores web y los administradores de sitios web, aumenta la complejidad de mantener una configuración de raspado web eficaz y sigilosa. Los sitios web siempre están ideando nuevas formas de detectar, ralentizar o bloquear los raspadores web, lo que requiere un enfoque dinámico para superar estas defensas.

A veces, lo mejor es dejar que los expertos se encarguen de las partes difíciles. Aquí es donde sobresalen los proxies de terceros y los servicios de raspado como Bright Data. Bright Data está constantemente a la vanguardia de las tecnologías anti-scraping, adaptando rápidamente sus estrategias para superar los nuevos obstáculos.

10. pruebe un acortador de URL

Tal vez sea capaz de superar bloqueadores de sitios poco sofisticados utilizando un servicio acortador de URL como Bitly, TinyURL o ls.gd. Estos servicios reemplazan la dirección URL de un sitio web con un nombre de dominio más corto. Si su centro educativo o su empresa bloquean YouTube, la versión acortada de Bitly tal vez le permita desbloquear el vídeo que desea ver.

Referencias:

https://brightdata.es/blog/datos-web/web-scraping-without-getting-blocked

https://www.octoparse.es/blog/scrape-websites-sin-ser-bloqueado

https://www.avast.com/es-es/c-how-to-unblock-websites-with-vpn