TRABAJO WEB CRAWLER (ARAÑAS WEB)

Los web crawlers o arañas web son programas o script automatizados utilizados para explorar y extraer información de páginas web en Internet. Su función principal es recorrer sitios web de manera sistemática, siguiendo los enlaces de una página a otra, para recopilar datos o indexar contenido.
El primer web crawler se llamaba **World Wide Web Wanderer** y se basaba en el lenguaje de programación PERL.

¿Cómo funciona un web crawler?

Los web crawlers están compuestos por un código de algoritmos y scripts que establece unos comandos e instrucciones claros. El web crawler repite las funciones indicadas en el código de forma automática y continua.

Los web crawlers se mueven por Internet a través de los hipervínculos que aparecen en los sitios web existentes. Evalúan palabras clave y hashtags, indexan el contenido y las URL de cada sitio web, copian páginas web para analizar sitios web.

Características principales

- **Automatización:** Los web crawlers funcionan de manera autónoma a la hora de visitar múltiples páginas web.
- **Exploración de enlaces:**Un web crawler es capaz de seguir los enlaces de una página a otra y de ese modo descubrir nuevas páginas dentro de la estructura de un sitio web.
- **Extracción de datos:** Es capaz de extraer información relevante de las páginas web (como texto, imágenes, enlaces, metadatos…).
- **Velocidad:** Están diseñados para rastrear una gran cantidad de sitios web en un corto espacio de tiempo.
- **Uso de algoritmos:** Utiliza algoritmos de búsqueda para decidir cómo organizar la información que extrae y cómo almacenar los resultados.

Tipos de web crawler

1. Motores de búsqueda:
    a. GoogleBot (Google)
    b. Bingbot (Bing)
    c. DuckDuckBot (DuckDuckGo)

2. Arañas web personalizadas: presentan una funcionalidad muy simple y son utilizados por las empresas para realizar tareas concretas. Por ejemplo, monitorean la frecuencia de determinados términos de búsqueda o la disponibilidad de ciertos URL

3. Arañas web comerciales: se trata de soluciones de software complejas desarrolladas por proveedores que las comercializan como herramienta.

4. Arañas web de escritorio: puedes ejecutar pequeños web crawlers en tu propio PC u ordenador portátil. Estas arañas web son económicas, pero tienen un uso muy limitado y, por lo general, solo pueden evaluar pequeñas cantidades de datos y sitios web.

5. Arañas web en la nube: también hay arañas web que no almacenan los datos en servidores locales, sino en una nube

6. Crawlers de monitoreo:
    a. Su objetivo es seguir el contenido de una web para monitorear cambios, como actualizaciones de noticias, precios, o la disponibilidad de productos.
    b. Ejemplo: Crawlers utilizados por empresas de monitoreo de precios o de seguimiento de contenido.

7. Crawlers de validación:
    a. Se utilizan para verificar enlaces rotos en sitios web o comprobar la validez de la información contenida en ellos.
    b. Ejemplo: W3C Link Checker para detectar enlaces rotos.

Proceso de Rastreo

1. El web crawler comienza con una lista de URLs conocidas también llamadas **semillas** o **seeds**.
2. Accede a la URL de la lista y descarga el código HTML de la página.
3. Analiza el código HTML para identificar todos los enlaces dentro de la página. Estos enlaces pueden ser internos (a otras páginas del mismo sitio) o externos (a otros sitios).
4. Filtra los enlaces para evitar rastrear páginas duplicadas, no relevantes o que están bloqueadas mediante archivos **robots.txt**.
5. Los enlaces se priorizan según ciertos criterios (frecuencia de actualización, relevancia…).
6. El crawler repite este proceso con los enlaces encontrados, siguiendo su camino por la web de manera recursiva.
7. La información recolectada se almacena en una base de datos o índice para su posterior análisis o acceso.