Analizadores de páginas web

(Analizadores HTML)

Analizadores de páginas web (Analizadores Html) Todos los sitios web y blogs modernos generan sus páginas usando JavaScript (como con AJAX, jQuery y otras técnicas similares). Por lo tanto, el análisis de páginas web a veces es útil para determinar la ubicación de un sitio y sus objetos. Una página web adecuada o un analizador HTML es capaz de descargar el contenido y los códigos HTML y puede llevar a cabo múltiples tareas de minería de datos a la vez.

Un analizador de HTML solicita una página web desde un servidor Web, tal y como se escribe las direcciones Web en la barra de direcciones del navegador. El servidor envía el código HTML para el analizador, que luego se explora a través de la página, en busca de etiquetas y texto. Se comprueba el archivo para asegurarse de que tiene las etiquetas HTML y en el orden correcto; de lo contrario puede ser un archivo Acrobat o algún otro tipo de documento. Si el autor preparó la página a mano, que puede haber cometido errores en el código HTML, lo que lleva al analizador para rechazarla. Si la página comprueba hacia fuera, el analizador recoge lo distinguen de acuerdo a las reglas de HTML. El analizador continuación, organiza, pantallas y recolecta información de la página web.

Los navegadores y web

La intención original era de la Web para hacer páginas legibles por humanos, y que es lo que hace un navegador. Se analiza el código HTML y crea una página visible, con el formato de su contenido. El navegador sabe cuándo hacer un poco de texto más grandes que otros, cómo mostrar enlaces web y cómo mostrar imágenes. Cuando el navegador haya acabado de crear la página, se espera a que el usuario haga clic en el ratón, escriba el texto o realizar alguna otra acción. Si el usuario hace clic en un vínculo o una dirección de tipos, el navegador va a buscar otra página Web.

rastreo web

sitios Web de búsqueda como Google, Bing y Ask tienen programas que escanean automáticamente toda la web, buscando información fresca. Estos llamados "rastreadores Web" leer una página web, catálogo de su texto y examinarla en busca de enlaces a otras páginas. Para encontrar los enlaces y otras informaciones importantes, los programas rastreadores web analizar el HMTL. A diferencia de un navegador, sin embargo, que no se muestran las páginas en una pantalla.

Captura de imágenes y spam

Los programadores escriben analizadores HTML para una variedad de propósitos. Algunos extraer automáticamente los datos del informe de tabla de páginas web, otros se reúnen imágenes. Los programadores llaman a esto "captura de imágenes", como los extractos de programas, o "raspaduras" los datos de la página Web y la recoja. Una práctica ilegal llamada "correo basura" implica un analizador automático que examina Web páginas para direcciones de correo electrónico, que el analizador puede identificar fácilmente. Una vez que el analizador sintáctico extrae la dirección, se añade a la base de datos del spammer. Otros programas a continuación, enviar automáticamente mensajes de venta por correo electrónico, o "spam", a la dirección. Para evitar que su buzón de correo electrónico inundado con mensajes de venta, no incluya su dirección de correo electrónico en páginas web accesibles al público.

Ventajas y desventajas

Ventajas: Trae consigo una variedad extensa de métodos para analizar diferentes aspectos, que se pueden complementar con diferentes funciones para complementarse ej: GitHub y ParseHub (son dos raspadores de páginas web más útiles que se pueden usar tanto para sitios básicos como dinámicos), pudiendo usar varias librerías de scripting para cumplir su objetivo.

Desventaja: No siempre se puede usar una librería de scraping por la protecciones que tiene la web esto hace mucho mas tardado el trabajo al tener que hacerlo manualmente y también algunas librerías tampoco puede complementarse entre si.

Referencia:

[Analizadores de páginas web o cómo obtener los datos que desea de la red | Semalt Q&A]

[La definición de un analizador de HTML / Seabrookewindows.com]