Introducción al web scraping, crawling y parsing

1.

Conceptos básicos

Crawling, *Scraping* y *Parsing*

- El **web *scraping*** (“raspado” de páginas web) consiste en la extracción de los datos significativos de una o varias páginas web determinadas, para una manipulación o análisis posterior

- Los conceptos de **web *crawling*** o **web *spider***, se refieren concretamente a que para obtener las páginas web que nos interesan hemos de rastrear sus enlaces web, realizando una exploración recursiva de todos sus enlaces

- Normalmente, hay que **parsear** los datos para extraer las partes que nos interesan

Web *scraping* y *crawling*

Estas técnicas permiten extraer datos web y analizarlos, para diversas aplicaciones:

- Alimentar una base de datos
- Hacer una migración de un sitio web
- Recopilar y ofrecer datos dispersos por varias webs
- Generar alertas
- Monitorización de precios de la competencia
- Localización de ítems o stock en *eCommerces*
- Recolección de fichas de productos
- Detección de cambios en sitios web
- Registrar lanzamientos y novedades
- Analizar los enlaces de un sitio para buscar links rotos
- Etc.

2.

Arañas web o *crawlers*

Arañas web o *crawlers*

▸Una araña web es un programa que **inspecciona** las páginas del World Wide Web de forma **metódica y automatizada**. Su uso más frecuente se centra en:

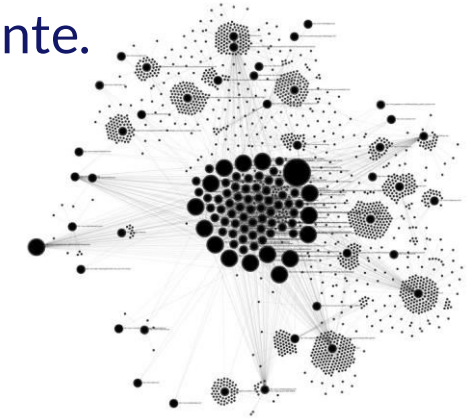
- Crear una copia de todas las páginas web visitadas
- Procesado posterior por un motor de búsqueda que indexa las páginas.
- Sistema de búsquedas rápido.

▸Las arañas web suelen ser bots.

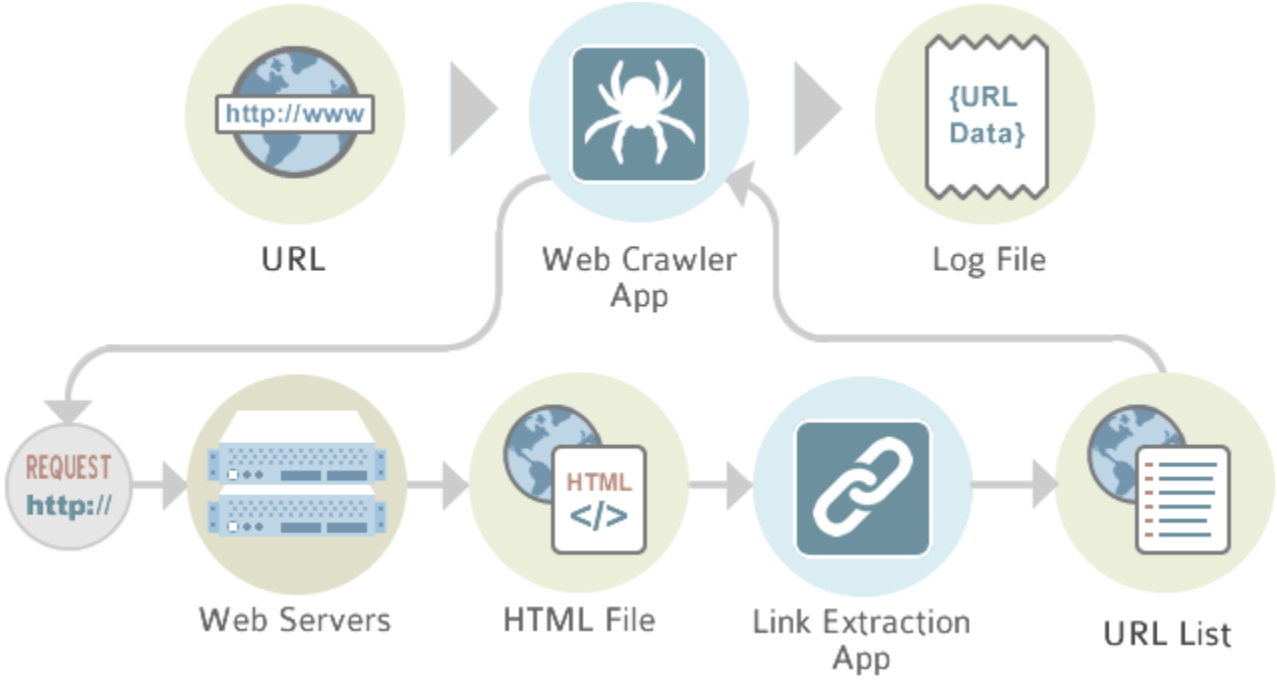
Arañas web o *crawlers*

Funcionamiento:

1. Las arañas visitan una lista de URLs.
2. Se descargan las páginas.
3. Identifica los hiperenlaces.
4. Los añade a la lista a visitar recurrentemente.
5. Luego descarga estas páginas nuevas.
6. Analiza sus enlaces.
7. Así sucesivamente.



Arañas web o *crawlers*



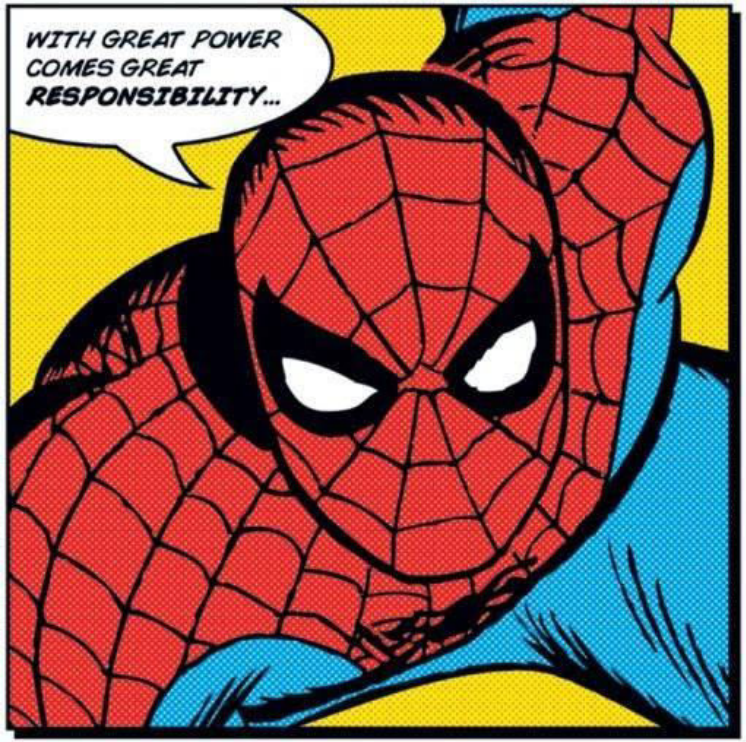
2.

Problemas al extraer datos web

Problemas al extraer datos web

- Existe cierta controversia sobre el *scraping* y algunas webs
- Cuanto más interesantes sean los datos proporcionados por una web, intentarán protegerlos y evitar las técnicas de web *scraping* o *crawling*
- Los accesos a una web que no se corresponden con "acciones humanas" (por ejemplo, el número de páginas solicitadas por minuto), pueden provocar el bloqueo de la IP
- Es conveniente mirar atentamente los términos legales de la web y tener en consideración los aspectos legales por la utilización de los datos obtenidos mediante web *scraping*

Problemas al extraer datos web



Problemas al extraer datos web

- http://www.facebook.com/terms.php

User Conduct

You understand that except for advertising programs offered by us on the Site (e.g., Facebook Flyers, Facebook Marketplace), the Service and the Site are available for your personal, non-commercial use only. You represent, warrant and agree that no materials of any kind submitted through your account or otherwise posted, transmitted, or shared by you on or through the Service will violate or infringe upon the rights of any third party, including copyright, trademark, privacy, publicity or other personal or proprietary rights; or contain libelous, defamatory or otherwise unlawful material.

In addition, you agree not to use the Service or the Site to:

- harvest or collect email addresses or other contact information of other users from the Service or the Site by electronic or other means for the purposes of sending unsolicited emails or other unsolicited communications;
- use the Service or the Site in any unlawful manner or in any other manner that could damage, disable, overburden or impair the Site;
- use automated scripts to collect information from or otherwise interact with the Service or the Site;

¡GRACIAS!

¿Preguntas?