

# THE BRIDGE

Web Scraping

# La Web

La World Wide Web (WWW) o red informática mundial es un sistema de distribución de documentos de hipertexto o hipermedios interconectados y accesibles vía Internet.

Con un navegador web, un usuario visualiza sitios web compuestos de páginas web que pueden contener textos, imágenes, vídeos u otros contenidos multimedia, y navega a través de esas páginas usando hiperenlaces.



# Web 1.0

La Web 1.0 es la original, el principio, el primer contacto que tuvimos con un entramado de páginas web, en las que básicamente nos limitábamos a consumir contenido sin más actualización o interacción.



- Primera página web:  
<http://info.cern.ch/hypertext/WWW/TheProject.html>

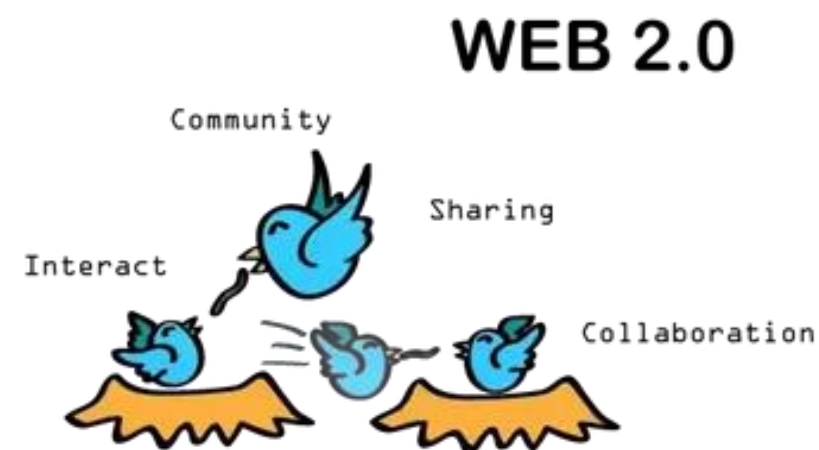
# Web 2.0

La Web 2.0 fue la primera gran evolución. La conocida web social permitió intercambiar información entre usuarios a través de blogs o las populares redes sociales que hoy usan millones de usuarios en todo el mundo



# Web 3.0

- En la Web 3.0, o web semántica, la clave y principal factor diferencial es el cómo accedemos a la información, de una forma más flexible y versátil.
- Los buscadores permiten hacer uso de un lenguaje más natural, de forma que obtenemos una información más personalizada.
- Se utilizan técnicas de machine learning e inteligencia artificial.





# Web 4.0

- La Web 4.0 es el próximo gran avance y se centrará en ofrecer un comportamiento más inteligente y predictivo
- Ofrecer soluciones a partir de toda la información que damos y existe en la Web. Para lograrlo, se fundamentará en tres pilares:
  1. La comprensión del lenguaje natural y tecnologías speech to text
  2. Nuevos sistemas de comunicación máquina a máquina (M2M)
  3. Uso de la información de contexto. Por ejemplo, ubicación que aporta el GPS, ritmo cardiaco que tu smartwatch registra, etc





“

La web 3.0 nunca podrá  
responder a consultas del tipo:  
“Quiero que un taxi venga a  
buscarme”

# La Web

Cada minuto en la web...

## 2020 *This Is What Happens In An Internet Minute*





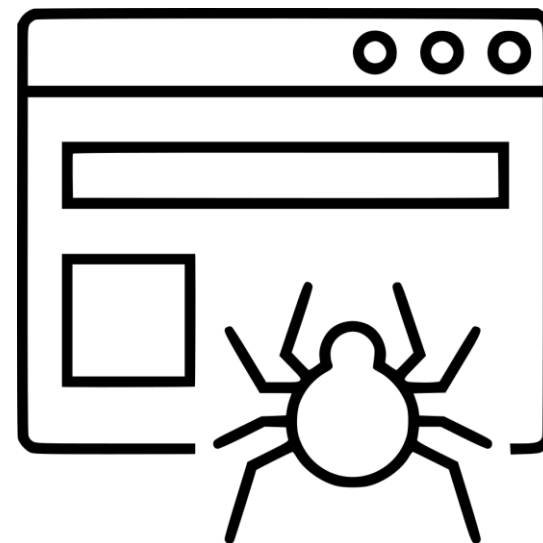
“

En internet hay más palabras que todas las que ha pronunciado la humanidad a lo largo de su historia

# Obtención de Datos web

## Scraping, Crawling 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.
- 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 eCommerce
- 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.

# Arañas web o crawlers

Una araña web es un programa que inspecciona las páginas 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.

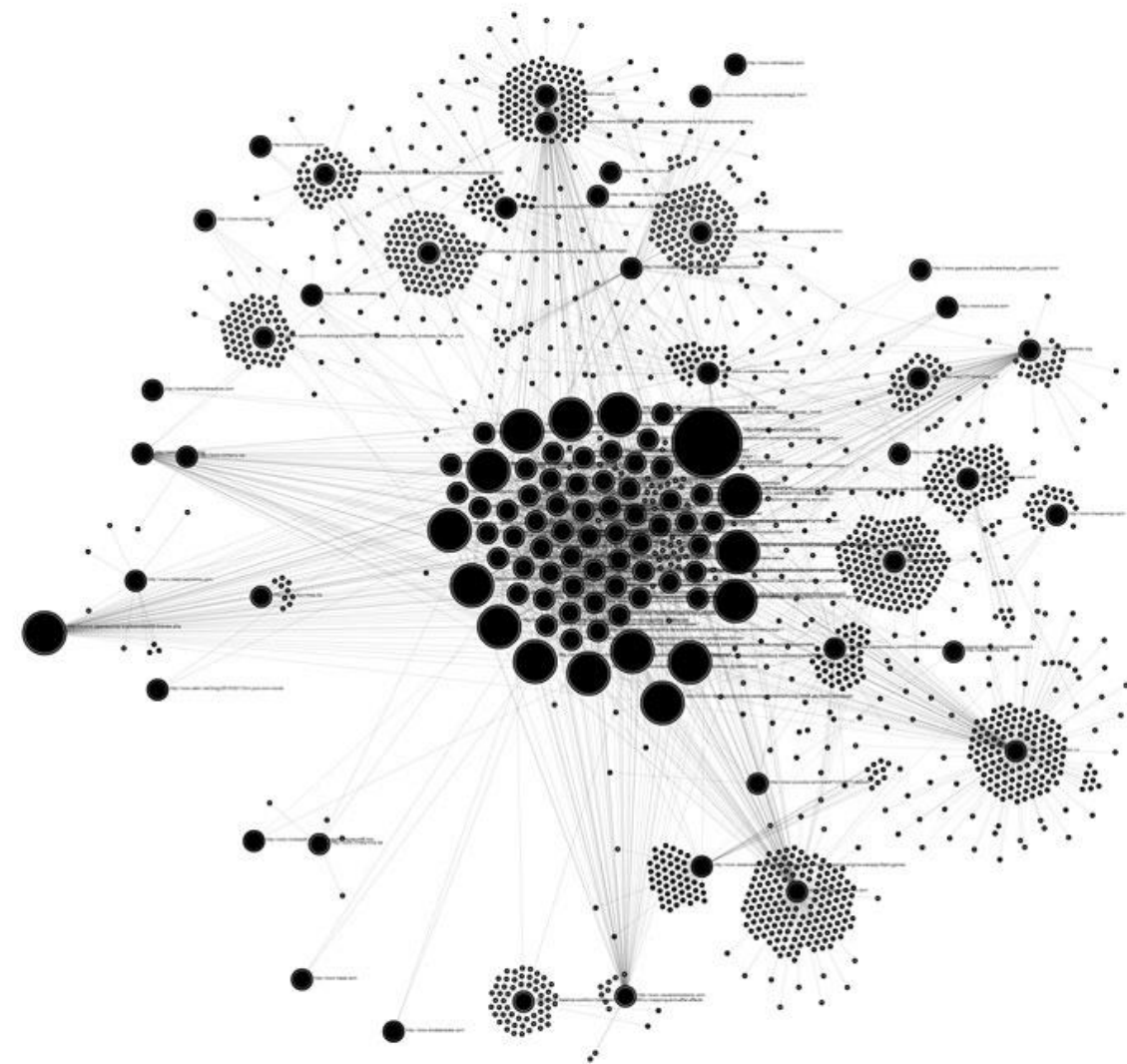
Las arañas web suelen ser bots.



# Arañas web o crawlers

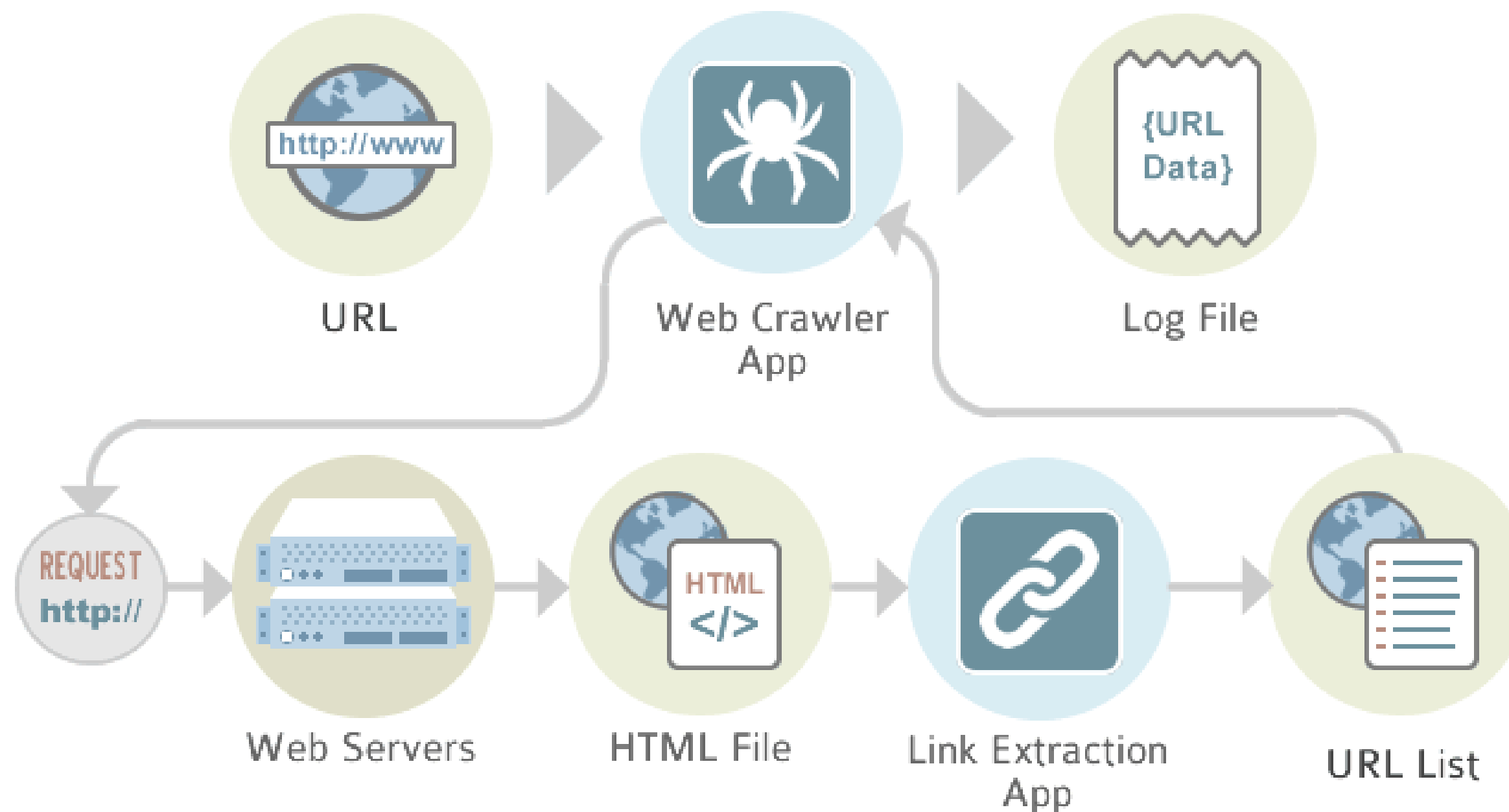
## Funcionamiento

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.



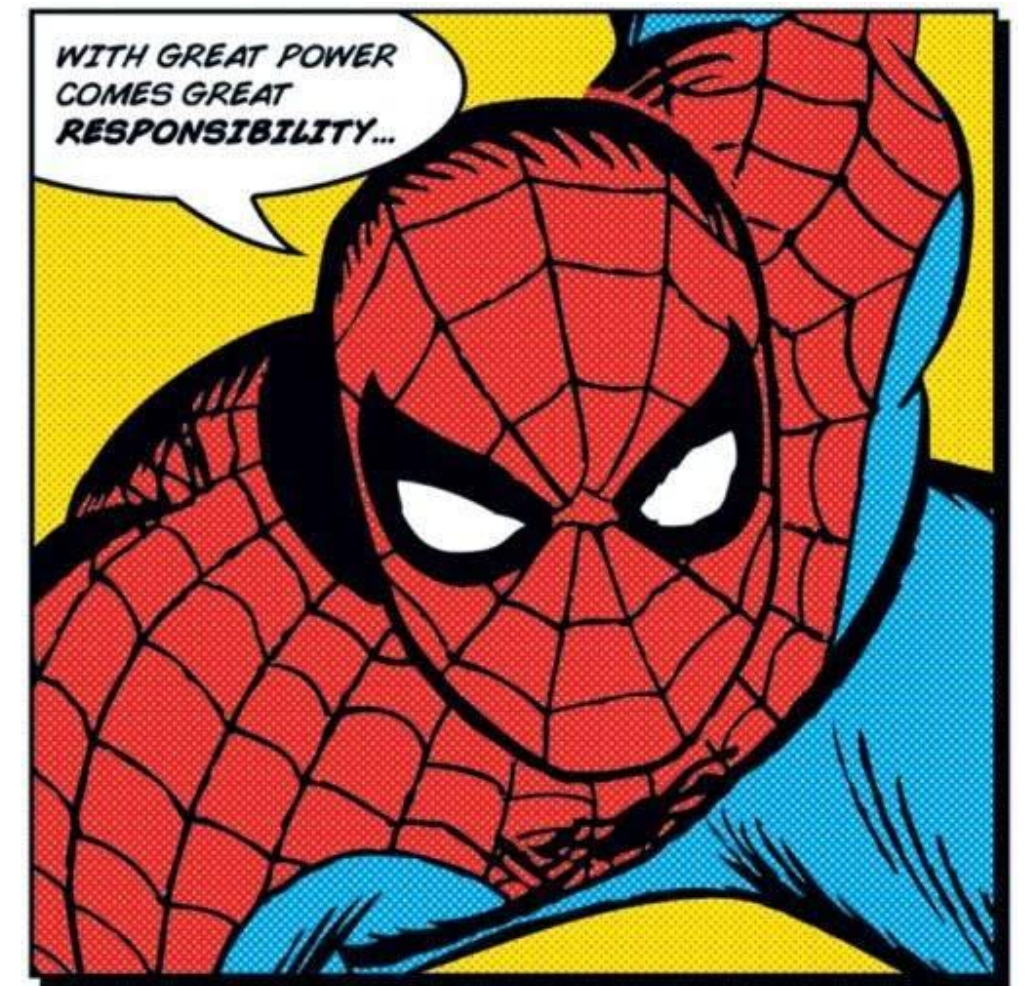
# Arañas web o crawlers

## Funcionamiento



# Problemas al extraer datos web

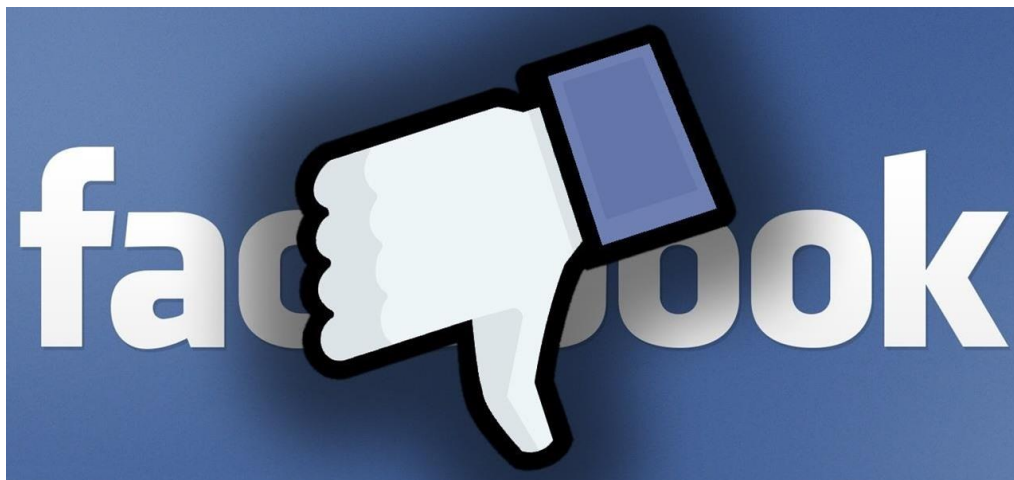
- Existe cierta controversia sobre el scraping y algunas webs
- Cuanto más interesantes sean los datos proporcionados por una web, intentarán protegerlos y evitar las técnicas de web scraping o crawling
- Los accesos a una web que no se corresponden con “acciones humanas” (por ejemplo, el número de páginas 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

<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;

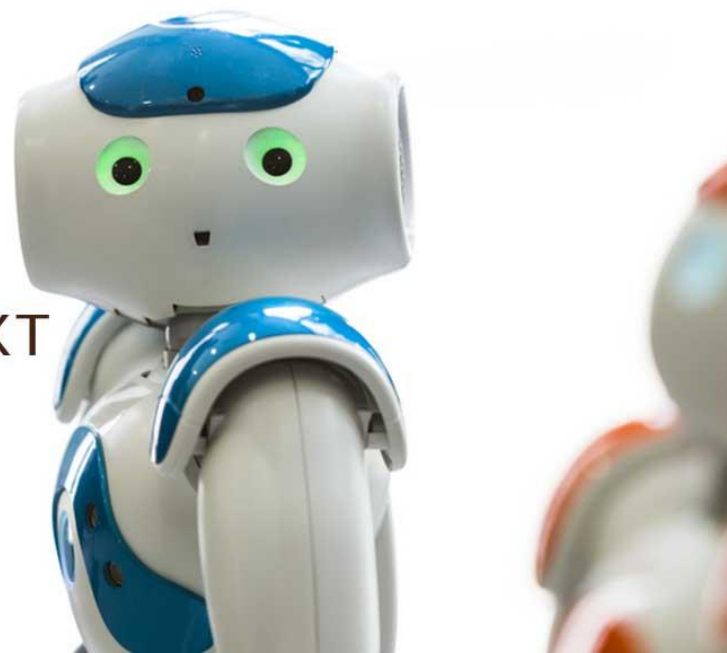




# Problemas al extraer datos web

- En la mayoría de las webs, existe un archivo llamado robots.txt que especifica a los bots como pueden indexar la página web, limitando a los web spiders
- Ejemplo: <https://facebook.com/robots.txt>
- Más información: <https://www.robotstxt.org/robotstxt.html>

ROBOTS.TXT




# Problemas al extraer datos web

Business

## LinkedIn sues 100 information scrapers after technical safeguard fail

Botnet harvests user data for spam and profit

By Iain Thomson in San Francisco 16 Aug 2016 at 20:36

55  SHARE ▼



# Consejos para extraer datos web

- 1) Si la web comparte sus datos a través de APIs, utilízalas en lugar de hacer scraping
- 2) Respeta los términos y servicios
- 3) Respeta las normas de robots.txt
- 4) Usa una tasa de peticiones razonables (1 petición por cada 10 segundos)
- 5) Si los términos y condiciones o robots.txt impiden acceder a una web, puedes intentar pedir permiso al propietario
- 6) No publiques tu código con crawlers o scraping
- 7) No bases tu negocio únicamente en web scraping
- 8) Sospecha de los consejos que encuentres en Internet



# Navegadores web y HTTP

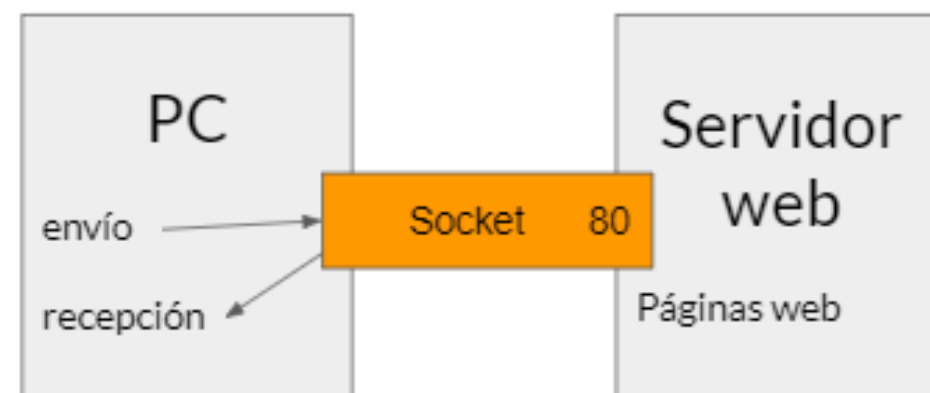




# Navegadores web y HTTP

## Protocolos

- Un **protocolo** es un conjunto de reglas precisas que permiten que dos o más entidades de un sistema de comunicación se comuniquen entre ellas para transmitir información.
- El **HTTP** ("Hypertext Transfer Protocol") es un protocolo mediante el cual se permite la transferencia de información entre diferentes servicios y los clientes que utilizan páginas web



Gracias

---

Rafa Zambrano

[rafael@thebridgeschool.es](mailto:rafael@thebridgeschool.es)