Crawling & Scraping Web

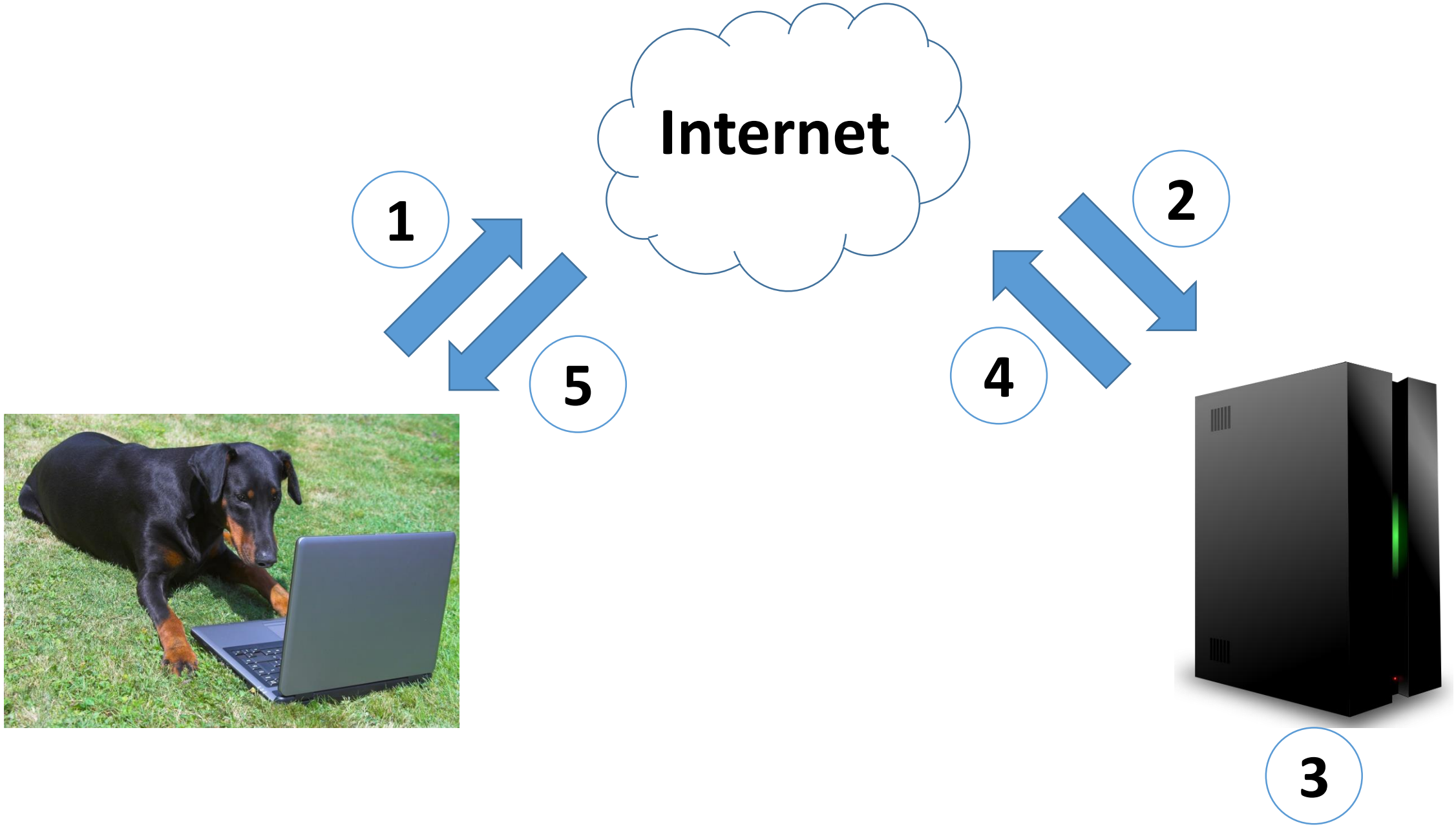
Por Ricardo Azpeitia Pimentel

¿Qué es HTTP?

Es un protocolo para realizar transacciones, basado en el esquema peticion-respuesta.

Para cada petición necesita al menos 2 cosas, un verbo (método GET, POST, etc) y un sustantivo (una URL)

¿Cómo funciona una petición HTTP?


Internet
1
2
3
4
5

HTML en 5 minutos

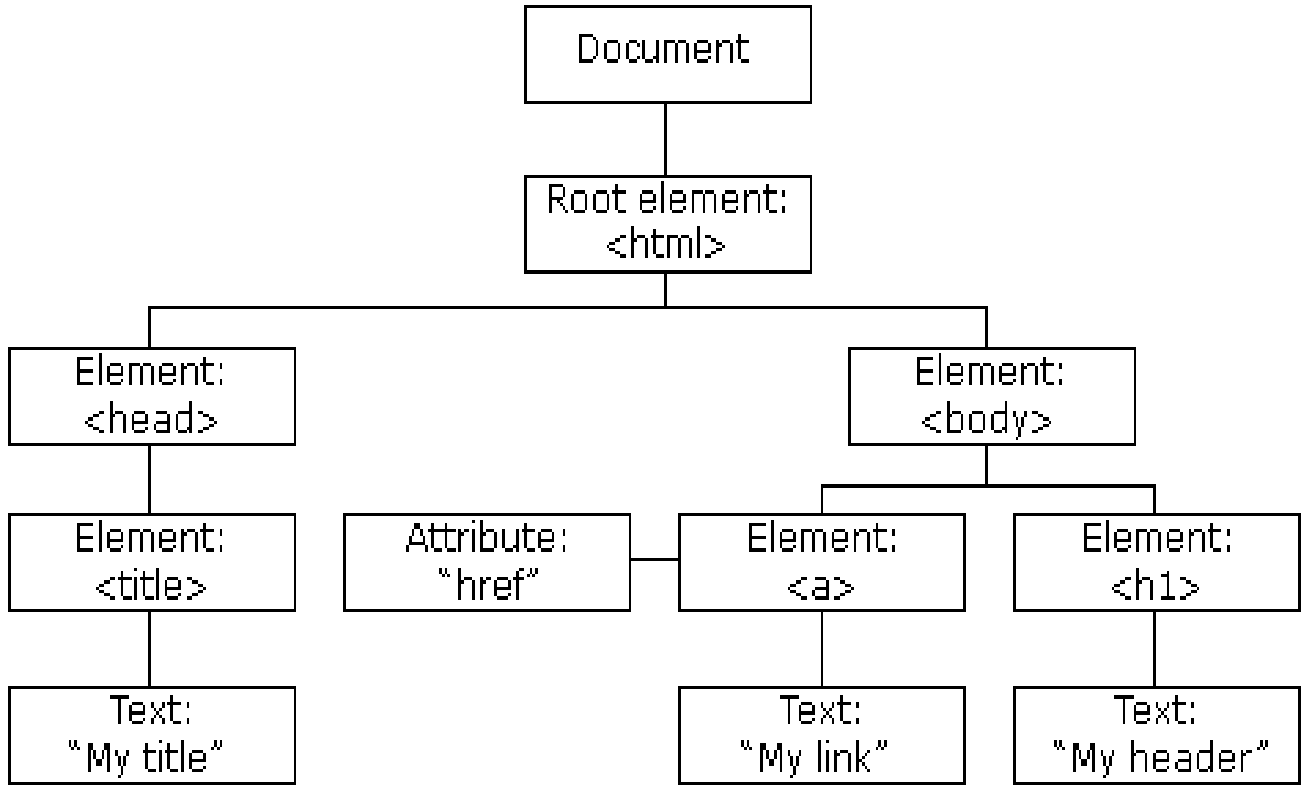
Básicamente solo es un lenguaje para estructurar como se presenta la información.

```
<html>
 <head>
  <title>My title</title>
 </head>

 <body>
  <a href="">My link</a>
  <h1>My header</h1>
 </body>
</html>
```

```
<html>
 <head>
  <title>My title</title>
 </head>

 <body>
  <a href="">My link</a>
  <h1>My header</h1>
 </body>
</html>
```



XPATH

Ejemplos XPath

- /html/body/p
- //a
- //a/@href
- /html/body/table/tr[3]/td/text()

¿Qué es crawling?

Un bot (o programa) que recorre paginas web.

¿Qué es scraping?

Obtener información a partir de los datos obtenidos por Internet

***Nota: No toda la información es HTML**

¿Cuándo es necesario hacer scraping?

A menos que sea la última alternativa

Urllib2 + regex

~~Urllib2 + regex~~

Herramientas

- Urllib2 + regex
- Requests
- Scrapy
- Mechanize
- Beatifulsoup (deprecado)
- Htmllib5

¿Cómo empezar un proyecto de scraping?

Prerequisitos para empezar

- Asegurarse que no existen APIs o Web Services que entreguen el contenido deseado, en formatos como JSON, XML, YAML, etc..

Pasos para empezar un proyecto de scraping

- Encontrar una fuente
- Encontrar los datos a extraer
- Obtener una llave primaria
- Code Code Code
- Probar
- Reparar y probar (indefinidamente)

Flujo de trabajo

Flujo de trabajo

- Visitar las URLs iniciales
- ¿Hacer scraping?
- Obtener las siguientes URLs
- ¿Más scraping?
- Repetir indefinidamente

Ejercicios