Clase 3: Intro a Web Scraping

Rvest y APIs

factor~data
IDAES_UNSAM

¿Qué es?

- **Scraping**
  “Rascar la olla”

  Gran variedad de herramientas, tanto basadas en lenguajes como en interfaces gráficas

¿Qué es?

- Scraping
- **HTML**
  HyperText Markup Language

  El lenguaje “estándar” en internet

  Es un la forma en que cada computadora habla entre sí y definen la forma en que debe ser procesado el texto.

  Para esto, el HTML usa dos elementos:

  tags y atributos

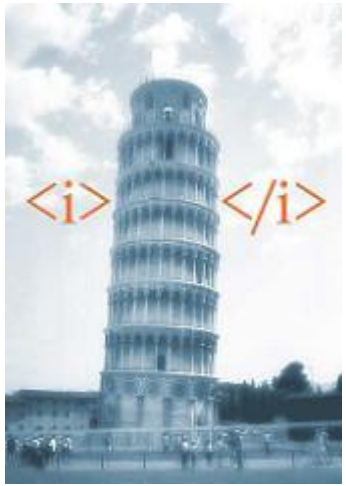


Image by Jesper Rønn-Jensen

```
<html>
  <head>
    <title>N&aacute;</title>
  </head>
  <body>
    Text stránky
  </body>
</html>
```

Image by Michaelbrabec

¿Qué es?

- Scraping
- **HTML**
  HyperText Markup Language

  Los **tags** se usan para marcar el inicio de un elemento HTML y se enmarcan, generalmente, en corchetes angulares. Un ejemplo: <h1>.

  La mayoría de los tags deben ser abiertos <h1> y cerrados </h1> para que funcionen.

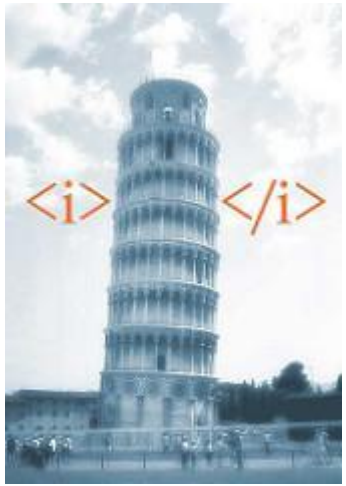


Image by Jesper Rønn-Jensen

```
<html>
  <head>
    <title>N&aacute;</title>
  </head>
  <body>
    Text stránky
  </body>
</html>
```

Image by Michaelbrabec

¿Qué es?

- Scraping
- **HTML**
  HyperText Markup Language

  Los **atributos** “estilizan” la página mediante CSS. Toma la forma de un tag abierto y se coloca información adicional dentro. Por ejemplo:

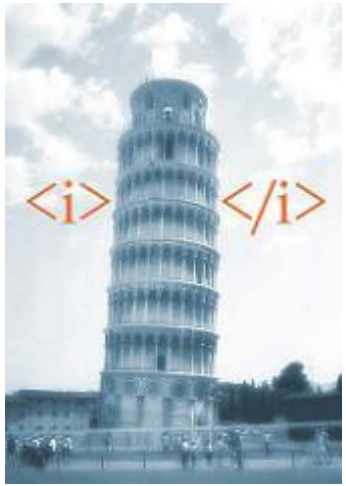


Image by Jesper Rønn-Jensen

```
<html>
  <head>
    <title>N&aacute;</title>
  </head>
  <body>
    Text stránky
  </body>
</html>
```

Image by Michaelbrabec

<img src="mydog.jpg" alt="A photo of my dog.">

Aquí, (src) y (alt) son atributos del tag <img>.

¿Qué es?

- Scraping
- HTML
- **Parseo**
  Proceso de analizar una secuencia de símbolos a fin de determinar su estructura gramatical con respecto a una gramática formal dada.

  En este caso, parsearemos código HTML para detectar la data que queremos extraer de un sitio

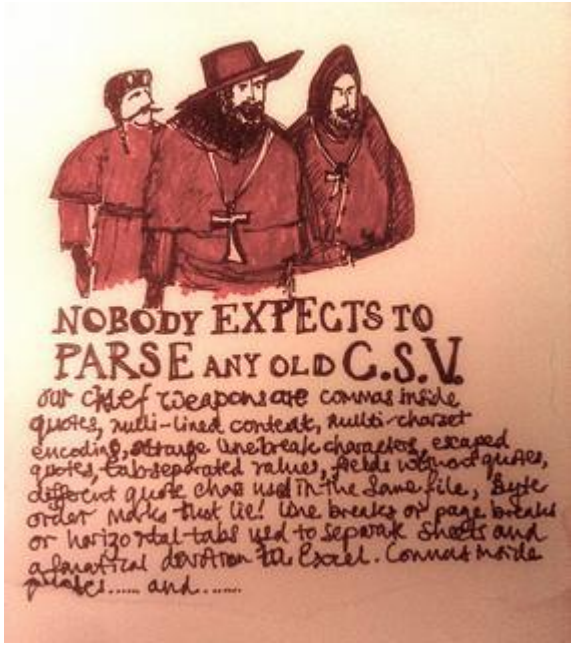


Image by Paul Downey

¿Qué es?

- Scraping
- HTML
- Parsing
- **Crawling**

Moverse a lo largo o a lo ancho de un sitio web para obtener y extreaer data de una o más URLs

Image by Dave Gingrich

Vamos al Notebook…

¿Qué es?

- Scraping
- HTML
- Parsing
- Crawling
- **JSON**
  Javascript Open Notation

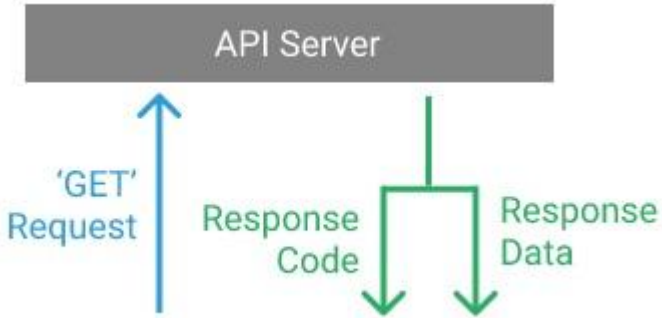
  Formato de texto usado para transmitir objetos de datos que consisten en pares "atributo-valor" -- Wikipedia


```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

¿Qué es?

- Scraping
- HTTP
- HTML
- Parsing
- JSON
- Crawling
- **API**
  Application Programming Interface

Un set de reglas y protocolos para construir aplicaciones. En el contexto del web scraping una API es un método para obtener data de forma limpia y estructurada de un sitio web.



Volvemos al Notebook…