

Introducción a la minería de textos y procesamiento de lenguaje natural para ciencias sociales

Clase 3. Introducción a web scraping + APIs



Dinámica de clases

- Bloques de 50-55 minutos
- Cortes de 15 minutos
- Actividades independientes

Herramientas



Hoja de ruta

- Conceptos básicos (http, HTML, json)
- Práctica guiada 1 - primera parte-
- APIs
- Práctica guiada 1 -segunda parte- práctica guiada 2

¿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. Permite describir la estructura de una página web.

Es una forma en que cada computadora habla entre sí y define la forma en que debe ser procesado el texto. De esta forma, un intérprete de html, como por ejemplo un navegador, va a saber cómo mostrarnos la información.

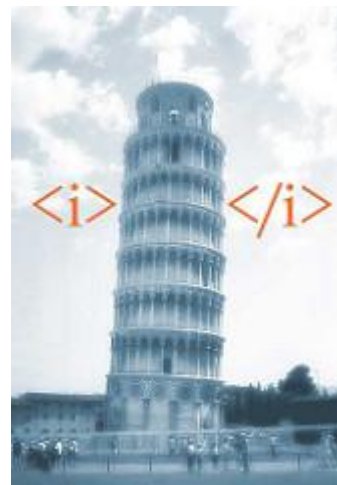


Image by [Jesper Rønn-Jensen](#)

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

Image by [Michaelbrabec](#)

¿Qué es?

- Scraping
- **HTML**

HyperText Markup Language

Cada elemento etiqueta con un determinado nombre algún contenido determinado. El nombre de esta etiqueta (**tag**) y el **atributo** de la misma nos permiten encontrar el contenido buscado.

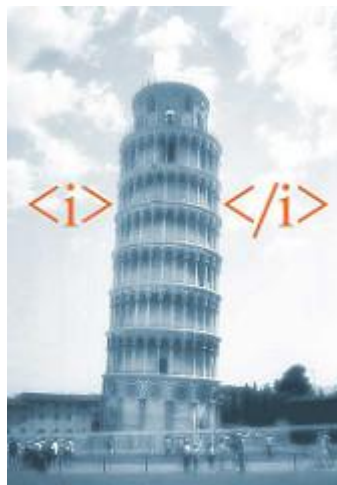


Image by [Jesper Rønn-Jensen](#)

```
<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.

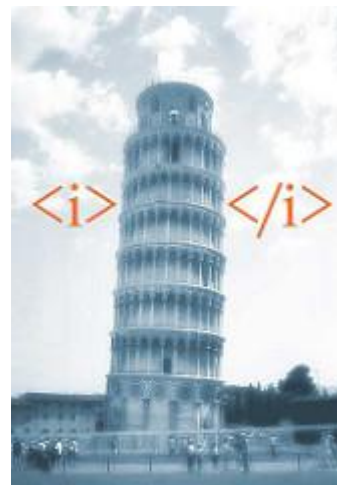


Image by [Jesper Rønn-Jensen](#)

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

Image by [Michaelbrabec](#)

¿Qué es?

- Scraping
- **HTML**

HyperText Markup Language

<div> marca secciones en el código

<p>

<a>

<h>

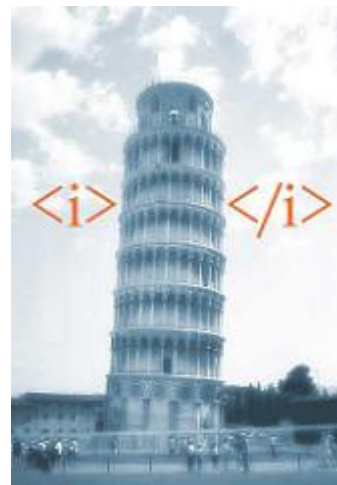


Image by [Jesper Rønn-Jensen](#)

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

Image by [Michaelbrabec](#)

¿Qué es?

- Scraping
- **HTML**

HyperText Markup Language

`<div>` marca secciones en el código

`<p>` marca párrafos

`<a>`

`<h>`

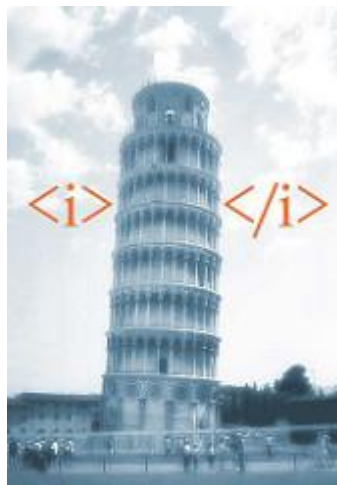


Image by [Jesper Rønn-Jensen](#)

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

Image by [Michaelbrabec](#)

¿Qué es?

- Scraping
- **HTML**

HyperText Markup Language

`<div>` marca secciones en el código

`<p>` marca párrafos

`<a>` introduce links

`<h>`

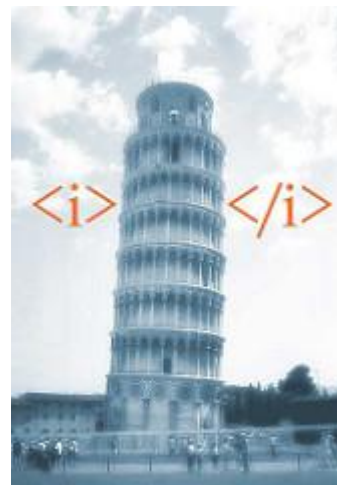


Image by [Jesper Rønn-Jensen](#)

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

Image by [Michaelbrabec](#)

¿Qué es?

- Scraping
- **HTML**

HyperText Markup Language

`<div>` marca secciones en el código

`<p>` marca párrafos

`<a>` introduce links

`<h>` **h1..h6**, marca encabezados

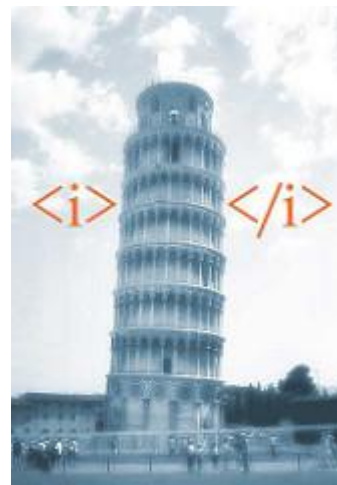


Image by [Jesper Rønn-Jensen](#)

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

Image by [Michaelbrabec](#)

¿Qué es?

- Scraping
- **HTML**

HyperText Markup Language

Los **atributos** contienen mayor información. Toma la forma de un tag abierto y se coloca información adicional dentro. Por ejemplo:

```

```

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

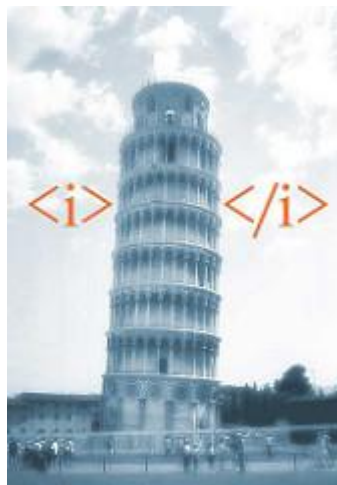


Image by [Jesper Rønn-Jensen](#)

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

Image by [Michaelbrabec](#)

¿Qué es?

- Scraping
- **HTML**

HyperText Markup Language

<class> define la clase de un tag, lo que permite mapearlo con un estilo dado un css

<href>



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

`<class>` define la clase de un tag, lo que permite mapearlo con un estilo dado un CSS

`<href>` define el link al cual efectivamente el tag nos referencia (en general dentro de un tag `<a>`)

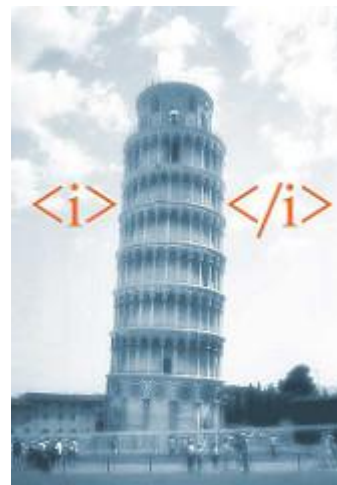


Image by [Jesper Rønn-Jensen](#)

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

Image by [Michaelbrabec](#)

¿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

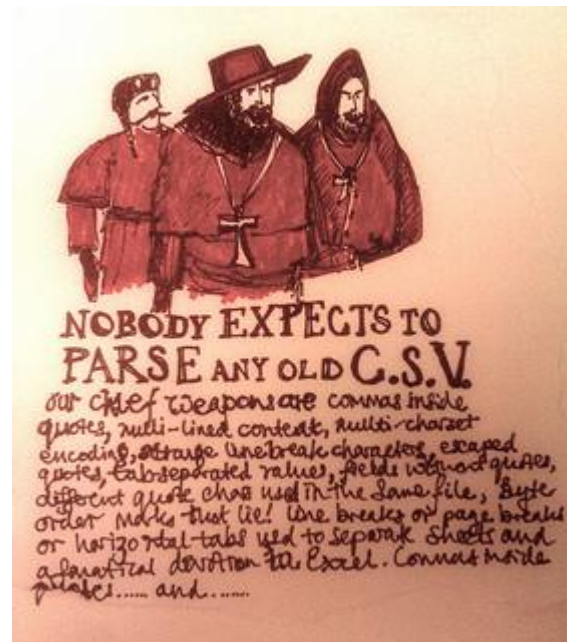


Image by [Paul Downey](#)

¿Qué es?

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

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



Image by [Dave Gingrich](#)

Vamos al Notebook

¿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.

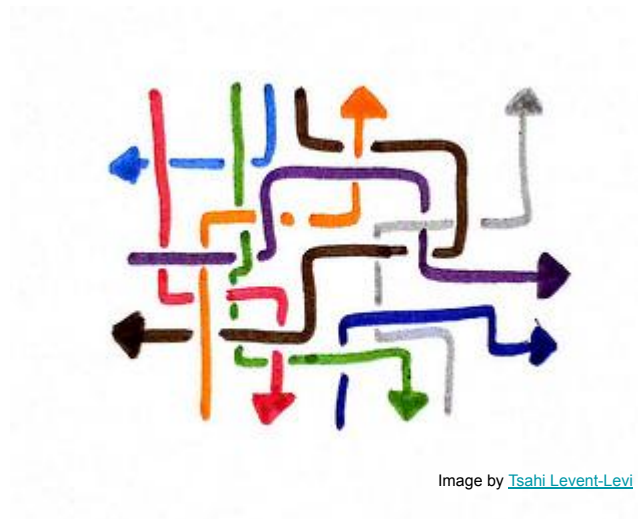


Image by [Tsahi Levent-Levi](#)

Vamos al Notebook