

Scraping express

El arte de recuperar datos

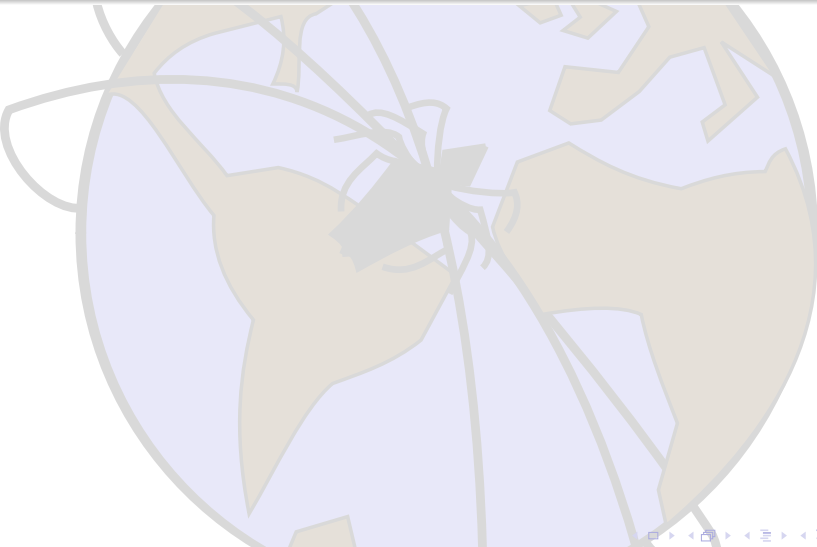
Serafín Vélez Barrera
serafa120000@gmail.com – @seravb




oficina de
software
libre



Índice



¿Qué eso del scraping?

A stylized background graphic featuring a globe with continents in light beige and oceans in light blue. Overlaid on the globe is a network of thin, grey lines that connect various points across the globe, suggesting a global web or data network.

El scraping es un técnica que se usa para recuperar **datos** de una web o documento básicamente.

¿Cómo se hace?

Existen varios métodos, por ejemplo:

Para una web
Tablas de PDF

Algún framework
Algunas web

Scrapy, FastCrawl..
Tabula

Instalación de Scrapy

Podremos instalar Scrapy de varias maneras:

- Descarga de la web oficial de Scrapy
- Línea de comandos:
 - `easy_install -U Scrapy`
 - `pip install Scrapy`
- Centro de software

Conociendo a Scrapy

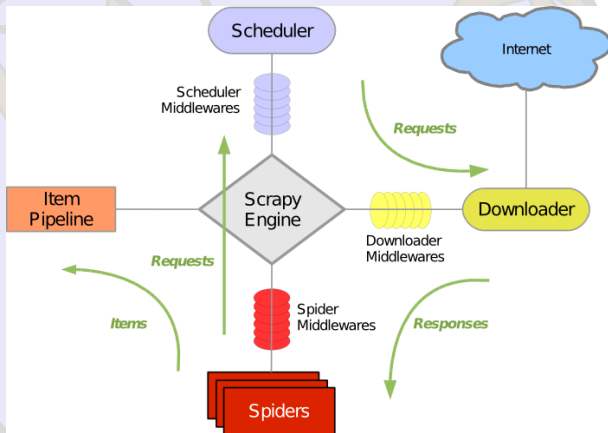
Cuando usamos Scrapy tenemos que crear un proyecto, y cada proyecto se compone de:

Items Definimos los elementos a extraer.

Spiders Es el corazón del proyecto, aquí definimos el procedimiento de extracción.

Pipelines Son los elementos para analizar lo obtenido: validación de datos, limpieza del código html...

Internamente Scrapy

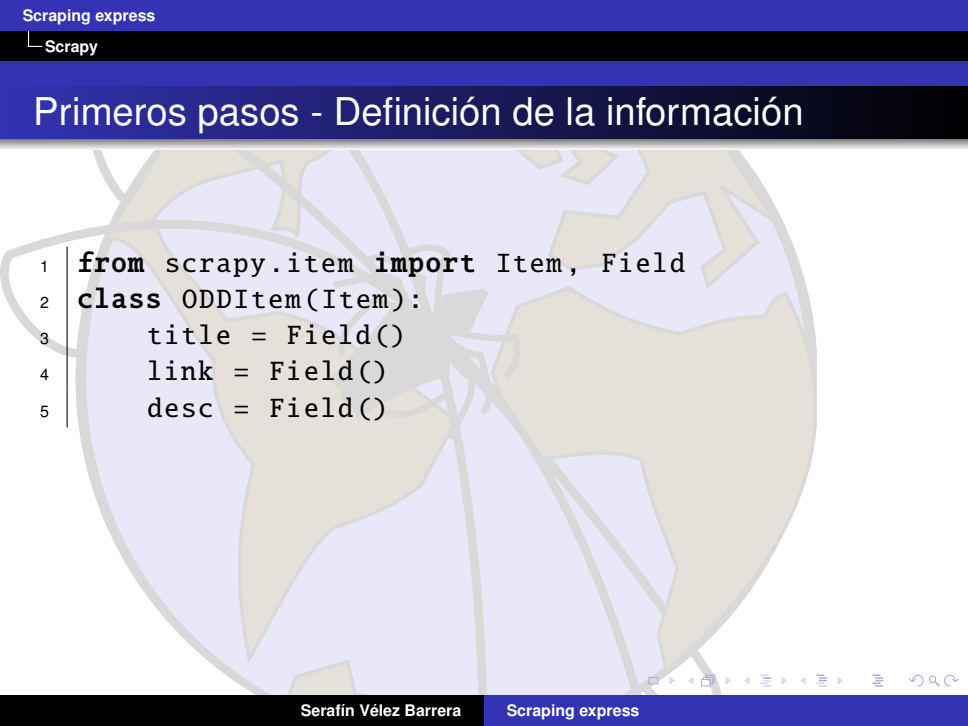


Primeros pasos - Crear un proyecto



```
1 | scrapy startproject OpenDataDayProject
```


Primeros pasos - Definición de la información



```
1 from scrapy.item import Item, Field
2 class ODDItem(Item):
3     title = Field()
4     link = Field()
5     desc = Field()
```

Primeros pasos - Programación de los Spiders

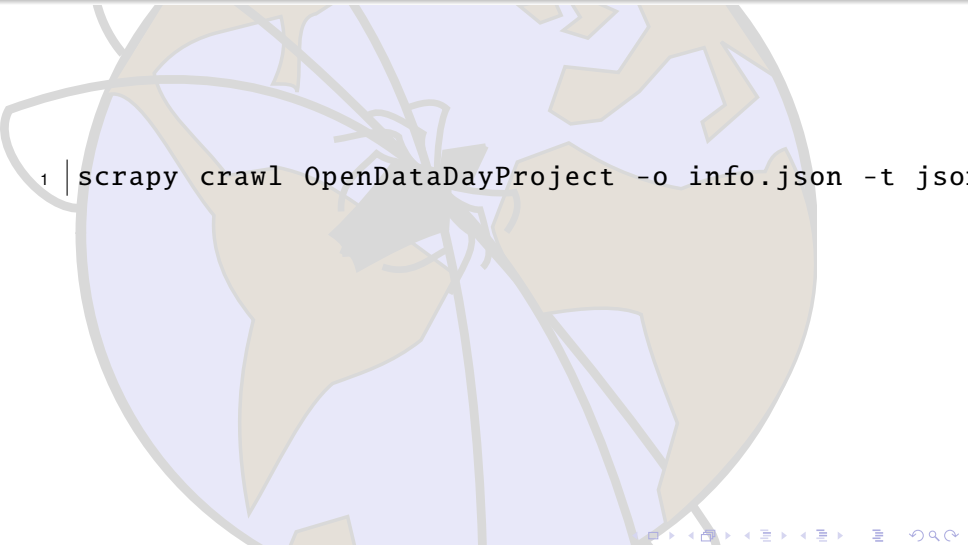
```
1 from scrapy.spider import BaseSpider
2 class ODDSpider(BaseSpider):
3     name = "odd"
4     allowed_domains = ["ugr.es"]
5     start_urls = [
6         "http://www.ugr.es"
7     ]
8     def parse(self, response):
9         filename = response.url.split("/")[-2]
10        open(filename, 'wb').write(response.body)
```

Primeros pasos - Ejecutando el proyecto



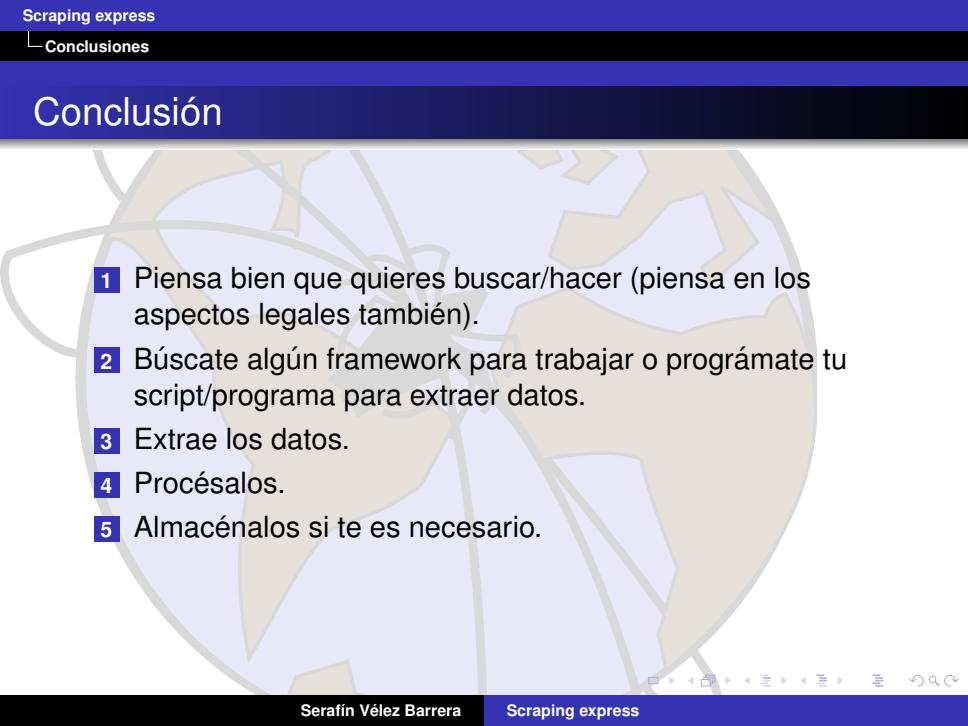
```
1 | scrapy crawl OpenDataDayProject
```

Primeros pasos - Salvando lo obtenido



```
1 | scrapy crawl OpenDataDayProject -o info.json -t json
```

Conclusión

- 
- 1 Piensa bien que quieres buscar/hacer (piensa en los aspectos legales también).
 - 2 Búscate algún framework para trabajar o prográmate tu script/programa para extraer datos.
 - 3 Extrae los datos.
 - 4 Procésalos.
 - 5 Almacénalos si te es necesario.


¿PREGUNTAS?

NOTAS

MÁS ACLARACIONES

¿ALGO OLVIDADO?

Bibliografía

- 
- A stylized background graphic featuring a light blue and tan globe with a network of grey lines and nodes overlaid, suggesting a global web or data network.
- Web oficial de Scrapy
 - Scrapy en un vistazo
 - Tutorial de Scrapy
 - Ejemplo en Github
 - Tabula

Licencia



Scraping express - El arte de recuperar datos
by **Serafín Vélez Barrera** is licensed under a
**Creative Commons Reconocimiento-
NoComercial-CompartirIgual 3.0 Unported
License.**