Scrapy

Artículo original de https://gabrielperales.wordpress.com/2014/09/01/web-scraping-con-scrapy/

En este pequeño tutorial vamos a extraer datos de una página de anuncios clasificados, de cada item o artículo nos interesará solo tres cosas, el nombre, la descripción y el precio.

Instalación y configuración

Primero instalamos python

```
sudo apt-get install python-pip python-dev libxml2-dev libxslt1-dev
```

Seguidamente vamos a crear un entorno virtual con virtualenv de python para nuestro proyecto:

```
virtualenv venv
```

Esto creará una carpeta venv donde contendrá un entorno virtual para python y donde instalaremos Scrapy. Ahora activamos el entorno virtual con el siguiente comando:

```
source venv/bin/activate
```

Una vez activado el entorno virtual instalamos Scrapy con pip install scrapy

```
pip install scrapy
```



Hay que tener instalada la librería libxml2 y libxslt para que no de error

Una vez instalado vamos a crear nuestra primera araña.



utilizaremos la plantilla crawl, que es igual que la araña basic pero esta nos permite seguir enlaces.

Para generar una araña usando esta plantilla debemos escribir en la consola scrapy genspider -t crawl. En nuestro caso a la araña la llamaremos segundamano y el dominio será el de la segundamano.es

```
scrapy genspider -t crawl segundamano segundamano.es
```

Al ejecutar el comando se genera dentro un fichero segundamano.py dentro del **directorio del**

proyecto/spiders

```
├── tutorial
│   ├── __init__.py
│   ├── items.py
│   ├── pipelines.py
│   ├── settings.py
│   └── spiders
│       ├── __init__.py
│       └── segundamano.py
└── scrapy.cfg
```

Referencias

- http://teclp.com/2015/03/install-scrapy-on-ubuntu-using-python-pip/

From:
https://wiki.intrusos.info/ - **LCWIKI**

Permanent link:
https://wiki.intrusos.info/doku.php?id=programacion:python:scrapy

Last update: **182023/01/ 13:36**