Webscraping

David Arroyo Menéndez

September 30, 2019

Conceptos Básicos

El webscraping se trata de extraer el contenido que aparece en una página web escrito en HTML generalmente y mostrarlo procesado.

Requests (I)

La librería estándar de Python para acceder a contenido a través de http es requests. En caso de duda se recomienda esta opción

Requests (II)

Descarga un fichero html

```
$ python3 requests-example.py http://www.gnu.org > gnu.html
```

Descarga un fichero json

```
$ python3 requests-example2.py > myrepos.json
```

Requests (III)

Ejercicio de sacar compradores y precios en una web

```
$ python3 requests-example3.py
Buyers: ['Carson Busses', 'Earl E. Byrd', 'Patty Cakes', 'Derri
Prices: ['$29.95', '$8.37', '$15.26', '$19.25', '$19.25', '$13.
```

Ejercicio de post, put, delete

```
$ python3 requests-example4.py # post, put, delete
```

Requests (IV)

Saca información de diferentes maneras desde el título

```
$ python3 titles.py
title tag: title
title text: David Arroyo Menéndez
title html: b'<title>David Arroyo Men&#233;ndez</title>\n    '
title tag: title
title's parent's tag: head
```

Request (V)

Detecta enlaces rotos en una web.

```
$ python3 urls.py
[...]
http://www.davidam.com/docu/aplic-ia/aplic-ia.sgml
200
http://www.davidam.com/docu/bibdigwebsem.html
200
http://www.davidam.com/docu/aumenta-tu-productividad-con-gnu-em
200
http://www.davidam.com/docu/lisp/lisp1.pdf
200
http://www.davidam.com/docu/un-lenguaje-en-diez-minutos.html
200
http://www.davidam.com/docu/lisp/lisp2.pdf
200
Urls with troubles:
[]
```

Beatiful Soup

Esta librería tiene una sintaxis algo más sencilla para el acceso a etiquetas html

```
$ python3 beatifulsoup-quickstart.py
```

Ejercicio de Webscraping con soporte multihilo

```
./app/urlthread.py
./app/crawler.pyc
./app/__init__.py
./app/formatter.py
./app/crawler.py
./app/damcrawler.py
./app/elmundo.py
./app/timeout.py
./test/test_crawler.py
./test/test_timeout.py
./test/test_urlthread.py

nosetest3 test
```