Taller Web Scrapping con Python

ÀREA HACKERS CÍVICS

Vicent Blanes

01/05/2017

¿Por qué Python?

- Python es un lenguaje **interpretado** y con una gran expresividad.
- Dispone de librerías potentes que nos ahorran mucho trabajo.
- Además es muy limpio y fácil de leer, un lenguaje ideal para aprender a programar.

Herramientas para Web Scrapping

- Utilizaremos la librería **urllib** para conectarnos a una dirección de Internet y obtener su contenido.
- Después utilizaremos **beautifulsoup4** para *parsear* el contenido y poder extraer información de forma sencilla.
- También utilizaremos la biblioteca **iPython** para trabajar con notebooks, que nos permiten ejecutar fragmentos de código de forma independiente y resultan mucho más interactivos.(?)

Anaconda

- Anaconda es una distribución de python que tiene pre-instaladas montones de librerías útiles
- Su instalación es muy sencilla. Podéis encontrarla en `https://www.continuum.io/downloads`.
- Además posee su propio gestor de paquetes. Si te falta alguna librería puedes usar **conda install** y el nombre de esta.
- Para el uso de los *notebooks* tenemos que ejecutar **conda install ipython**.

Material adicional

Junto a esta breve presentación y los notebooks que vamos a utilizar encontraréis también un anexo explicativo donde muestran algunas de las funciones python más importantes que usaremos durante el taller. Podéis descargar los ficheros del repositorio `https://github.com/vblanes/tallerWebScrappingAHC`.

Muchas opciones

- Para este taller usaremos **urllib** y **beautifulsoup4** pero existen más alternativas cuando hablamos de descargar el contenido de una web y tratarlo para conseguir la información deseada.
- Python dispone de herramientas para hacer Web Scrapping a una escala más grande. Por ejemplo la librería **scrappy** nos permite construir una *araña* para ir explorando los enlaces de la web, como si recorriéramos un grafo.

Contacto

-  https://t.me/vblanes
-  viblasel@gmail.com
-  https://www.facebook.com/viblasel