Librerías para determinar el URL requesting y el HTML parsing

Fecha: Viernes, 04/09/2019

Autor: Jonathan Bandes

Populares módulos de python para web scraping

1. Requests
    a. La biblioteca Requests es vital para agregar al kit de herramientas para la recolección de datos. Es una biblioteca HTTP simple pero potente, lo que significa que puede usarse para acceder a páginas web.
    b. Su simplicidad es definitivamente su mayor fortaleza. Es tan fácil de usar que puedes saltar directamente sin leer la documentación.
    c. **Requests Quickstart Guide** - Documentación Oficial.

2. Mechanize
    a. Es un popular módulo de Python que permite la creación de una instancia de navegador. También mantiene sesiones que ayudan como un kit de herramientas para obtener tareas como inicio de sesión, automatización de registro, etc.
    b. **Mechanize** - Documentación Oficial

3. BeautifulSoup
    a. Es otro módulo de Python que ayuda al scraping de los datos requeridos de HTML y XML a través de etiquetas. Con Beautiful se puede hacer scraping de

casi todo porque ayuda con diferentes métodos, como buscar a través de etiquetas, encontrar todos los enlaces, etc.

b. Una de las ventajas de BS4 es su capacidad para detectar automáticamente las codificaciones. Esto le permite manejar con gracia los documentos HTML con caracteres especiales.

c. **Beautiful Soup Documentation** - Incluye una rápida guía

d. **Really Short Example** - Incluye un ejemplo usando BeautifulSoap y Requests, juntos

4. Lxml

a. lxml es otra biblioteca maravillosa para analizar xml / htmls. La funcionalidad es similar a la de BeautifulSoup pero éste lo supera ligeramente. Por lo que se podría usar cualquiera de los dos módulos, ya que hacen prácticamente lo mismo.

b. **lxml Documentation** - Guía Oficial

c. **HTML Scraping with lxml and Requests** - Tutorial más breve que la guía oficial

5. BeautifulSoap vs lxml

Históricamente, la regla de oro era:

Si necesita velocidad, se usa lxml.

Si necesitas manejar documentos “sucios”, elige Beautiful Soup.

Sin embargo, esta distinción ya no se mantiene. Beautiful Soup ahora admite el uso del analizador lxml y viceversa. También es bastante fácil aprender el otro una vez que hayas aprendido uno.

La recomendación es usar ambos y probar cuál resulta más eficiente para lo que se quiere. Para este caso, de momento, después de hacer ambas pruebas, se prefirió BeautifulSoap.