Práctica 1: Web scraping

Descripción

Esta práctica se ha realizado bajo el contexto de la asignatura Tipología y ciclo de vida de los datos, perteneciente al Máster en Ciencia de Datos de la Universitat Oberta de Catalunya. En ella, se aplican técnicas de web scraping mediante el lenguaje de programación Python para extraer así datos de la web QueLibroLeo para obtener información de un listado de libros recomendados y sus características, para luego usar la REST API de Google Books, cruzar los datos y añadir información adicional para el dataset final.

Para el web scraping, ya que en anteriores ocasiones había usado la librería BeautifulSoap, he utilizado la librería Scrapy de Python en conjunción con xpaths. Se obtiene un listado de libros recomendados para su lectura con los atributos Nombre, Género, Editorial, Año, ISBN e Idioma. A continuación se hace uso de la REST API de Google Books. La autenticación se realiza con un API key y se realizan búsquedas basadas en los ISBNs de los libros ya capturados en la web QueLibroLeo. Desde Google Books obtenemos el número de páginas del libro y la valoración media de los usuarios (ambas cuando están disponibles) y se añade al dataset creado desde el scraping.

El motivo de la elección del sitio web es el interés por la literatura. Dicha web proporciona recomendaciones de lectura muy interesantes, pero desafortunadamente la usabilidad y el rendimiento de la web no es la deseada. Es por eso que he pensado en crear un dataset con dicha información aprovechando el ejercicio de web scraping de esta práctica, y además he querido complementaria con información adicional haciendo uso de otro método de adquisición de datos (API).

Dataset (libros-recomendados)

El Dataset está formado por los siguientes campos:

Nombre del libro, Género, Editorial, Año, ISBN, Idioma, Número de páginas (obtenido desde Google Books) y Average Rating (obtenido desde Google Books)

Se puede ver una imagen de un extracto del dataset aquí:

https://github.com/josem-gomez/Web-scraping---REST-API-Libros/blob/master/Imagen_Dataset.png

	A	B	C	D	E	F	G	H
1	nombre_libro	genero	editorial	año	isbn	idioma	numero_paginas	average_rating
2	PAULA	Biografías, Memorias	DEBOLSILLO	2009	9788497593885	Español	45	3.5
3	CRÓNICA DE UNA MUERTE ANUNCI/	Literatura contemporánea	DEBOLSILLO	2003	9788497592437	Español	54	NULL
4	EL CORONEL NO TIENE QUIEN LE ES(	Narrativa	DEBOLSILLO	2003	9788497592352	Español	86	NULL
5	LOS SANTOS INOCENTES	Literatura contemporánea	CRÍTICA	2005	9788484325659	Español	96	NULL
6	LOS GIRASOLES CIEGOS	Narrativa	ANAGRAMA	2006	9788433968555	Español	109	4
7	YERMA	Poesía, teatro	AUSTRAL	2011	9788467033632	Español	111	NULL
8	MARIANELA	Literatura contemporánea	CÁTEDRA	2009	9788437625430	Español	127	NULL
9	LUCES DE BOHEMIA	Poesía, teatro	AUSTRAL	2010	9788467033274	Español	127	NULL
10	LA CASA DE BERNARDA ALBA	Poesía, teatro	AUSTRAL	2010	9788467033328	Español	127	NULL
11	PLATERO Y YO	Clásicos de la literatura	ESPASA	2011	9788467019766	Español	128	3
12	LA TREGUA	Poesía, teatro	ALIANZA	2011	9788420650852	Español	132	5
13	CAMPOS DE CASTILLA	Poesía, teatro	ALIANZA	2006	9788420660554	Español	140	NULL
14	EL TÚNEL	Narrativa	SEIX BARRAL	2007	9788432217531	Español	144	NULL
15	HISTORIAS DE CRONOPIOS Y DE FAN	Narrativa	EDHASA	2010	9788435018678	Español	149	NULL
16	LA SOMBRA DEL ÁGUILA	Histórica y aventuras	ALFAGUARA	2017	9788420474694	Español	154	NULL
17	LAS RATAS	Literatura contemporánea	DESTINO	2010	9788423343409	Español	164	NULL
18	SAN MANUEL BUENO, MÁRTIR	Clásicos de la literatura	ESPASA	2007	9788467021677	Español	175	4
19	RELATO DE UN NÁUFRAGO	Narrativa	TUSQUETS	1989	9788472230088	Español	176	5
20	HISTORIA DE UNA ESCALERA	Poesía, teatro	ESPASA	2010	9788467021455	Español	187	NULL
21	MANOLITO GAFOTAS	Infantil y juvenil	ALFAGUARA	2002	9788420464534	Español	188	5
22	SIN NOTICIAS DE GURB	Humor	SEIX BARRAL	2002	9788432221255	Español	188	3
23	BODAS DE SANGRE	Poesía, teatro	AUSTRAL	2010	9788467033397	Español	192	NULL
24	DEL AMOR Y OTROS DEMONIOS	Literatura contemporánea	DEBOLSILLO	2003	9788497592420	Español	192	4
25	PEDRO PÁRAMO	Literatura contemporánea	CÁTEDRA	2004	9788437604183	Español	198	NULL
26	EL CAMINO	Literatura contemporánea	ESPASA	2006	9788467023664	Español	200	NULL
27	EL ALEPH	Narrativa	ALIANZA	2003	9788420633114	Español	203	4
28	COMO AGUA PARA CHOCOLATE	Romántica, erótica	DEBOLSILLO	2016	978	Español	209	3.5
29	FICCIONES	Ficción literaria	ALIANZA	2006	9788420633121	Español	218	3
30	EL ÁRBOL DE LA CIENCIA	Literatura contemporánea	ALIANZA	2011	9788420658803	Español	250	NULL
31	NIEBLA	Clásicos de la literatura	ESPASA	2010	9788467033861	Español	254	5
32	LA SOMBRA DEL VIENTO	Narrativa	PLANETA	2006	9788408057932	Español	255	NULL
33	TUAREG	Histórica y aventuras	DEBOLSILLO	2003	9788497592796	Español	256	1
34	CIENFUEGOS	Histórica y aventuras	DEBOLSILLO	2005	9788497595766	Español	256	5
35	LA FAMILIA DE PASCUAL DUARTE	Literatura contemporánea	DESTINO	2003	9788423334797	Español	260	5
36	RÉQUIEM POR UN CAMPESINO ESP/	Literatura contemporánea	DESTINO	2002	9788423308606	Español	272	NULL
37	EL ALQUIMISTA IMPACIENTE	Novela negra, intriga, terror	DESTINO	2011	9788423344741	Español	281	NULL
38	LA CIUDAD Y LOS PERROS	Narrativa	ALFAGUARA	2012	9788420412337	Español	288	NULL

Click to see details of this meeting

Y se puede acceder al dataset desde el repositorio de Github o desde el enlace de Zenodo:

https://zenodo.org/api/files/a4eb9576-c436-4cd4-b66b-a8f2eb7670e4/dataset_libros.csv

Como se ha referido en el punto anterior, el origen de los datos es principalmente de la web QueLibroLeo, con datos adicionales provenientes de Google Books.

La web no tiene ningún tipo de restricción para hacer scraping. Adjunto el contendi del fichero robos.txt: https://quelibroleo.hola.com/robots.txt

```
User-agent: *
Allow:
```

Licencia

Database released under Open Database License, individual contents under Database Contents License. Se ha elegido esta licencia con el

fin de permitir a los usuarios compartir sus datos con libertad y sin temor a los derechos de autor o cuestiones de propiedad.

Miembros del equipo

La actividad ha sido realizada de manera individual por Jose Manuel Gómez López.

Código principal

```
import scrapy
import logging
from scrapy.crawler import CrawlerProcess
import requests
import csv

libros = []

# Definimos clases para hacer scraping
class LibrosSpider(scrapy.Spider):
    name = 'libros'
    allowed_domains = ['quelibroleo.hola.com', 'www.quelibroleo.com']
    start_urls = [
        'https://quelibroleo.hola.com/noticias/libros/los-100-mejores-libros-de-
la-literatura-universal-en-espanol']

    def parse(self, response):

        # Usamos xpaths para encontrar las estructuras html que necesitamos
        libros = response.xpath(
            "//div[@class = 'entry-content']/p[position() >= 4 and not(position()
> 104)]/a[1]")

        # Recorremos el resultado con un bucle y extraemos los nombres de lso
libros y los siguientes enlaces para scrapear
        for libro in libros:
            nombre = libro.xpath(".//text()").get()
            enlace = libro.xpath(".//@href").get()
            url= enlace.replace("http://", "https://")
            yield scrapy.Request(url=url, callback=self.parse_libro,
meta={'nombre_libro': nombre})

    #Función para obtner información adicional de cada libro usando su nueva url
```

```
    def parse_libro(self, response):
        global libros
        print('Extrayendo…' + response.url)
        nombre = response.request.meta['nombre_libro']

        # Volvemos a usar xpaths para obtener los campos que necesitamos
        linea = {
            'nombre_libro': nombre,
            'genero': response.xpath("//span[text()='Género']/following-
sibling::a/text()").get(),
            'editorial': response.xpath("//span[text()='Editorial']/following-
sibling::a/text()").get(),
            'año': response.xpath("//span[text()='Año de edición']/following-
sibling::text()").get(),
            'isbn': response.xpath("//span[text()='ISBN']/following-
sibling::text()").get(),
            'idioma': response.xpath("normalize-
space(//span[text()='Idioma']/following-sibling::text())").get()
        }

        # Vamso añadiendo lineas a la lista. Cada linea será un diccionario con
todos los campos de cada libro
        libros.append(linea)

# Definimos la clase google_books para hacer uso de su REST API. La autenticación
a la la API se realiza con una llave (API Key)
class google_books():
    googleapikey="*******************************"

    def search(self, value):
        global libros
        # Definimos lo parametros de busqueda y la llave para la autenticación
        parms = {"q":value, 'key':self.googleapikey}
        # Hacemos uso de la librería requests para la llamada REST
        r = requests.get(url="https://www.googleapis.com/books/v1/volumes",
params=parms)
        print (r.url)
        #Convertimos en json el resultado de la búsqueda
        resultado = r.json()

        numero_paginas='NULL'
        average_rating='NULL'

        #Recogemso los campos que necesitamos cuando esten disponibles para
añadirlso al dataset del web scraping y completar la nfomación
```

```
        if ("items" in resultado):
            if ("volumeInfo" in resultado['items'][0]):
                if ("pageCount" in resultado['items'][0]["volumeInfo"]):

numero_paginas=resultado['items'][0]["volumeInfo"]["pageCount"]

                if ("averageRating" in resultado['items'][0]["volumeInfo"]):

average_rating=resultado['items'][0]["volumeInfo"]["averageRating"]

        res = {'numero_paginas': numero_paginas, 'average_rating':
average_rating}

        return  res



if __name__ == "__main__":


    process = CrawlerProcess({
        'USER_AGENT': 'Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)',
    })

    #Llamamos al spider para empezar el scraping
    process.crawl(LibrosSpider)
    process.start()
    print(libros)

    #Inicializamos la conexión con google books creando un objeto de su clase
    book = google_books()

    #Buscamos los libros a través de la APi de Google haciendo uso del ISBN de
cada libro
    for libro in libros:
        busqueda = "isbn" + libro['isbn']
        print(busqueda)
        info_adicional = book.search(busqueda)
        print(info_adicional)
        #Añadimos la información adicional a la lista de diccionarios
        libro.update(info_adicional)

    print(libros)
```

```
    #Por último, creamos el dataset con la información recogida de ambos procesos
    csvfile = "dataset_libros.csv"
    keys = libros[0].keys()
    with open(csvfile, 'w') as output_file:
        dict_writer = csv.DictWriter(output_file, keys)
        dict_writer.writeheader()
        dict_writer.writerows(libros)
```

Ficheros incluidos

* **/Descripción_del_Proyecto.txt**: Archivo con expliación del contexto de la recogida de datos y del dataset
* **ListaLibros/spiders/libros.py**: punto de entrada al programa. Inicia el proceso de scraping.
* **ListaLibros**: Directorio con ficheros de configuración de Scrapy
* **ListaLibros/spiders/ejemplo_output.txt**: Ejemplo de salida del scraper

Recursos

1. Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
2. Masip, D. El lenguaje Python. Editorial UOC.
2. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
4. Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.