軟體開發-Python

爬蟲實務

黃啟倫

爬蟲

- 網路蜘蛛（Web spider）
- 網路爬蟲（Web crawler）
- 爬蟲的基本邏輯
  - 利用URL進行定位
  - 讀取html碼
  - 利用迴圈大量執行
  - URL或是網頁固定格式(EX:YahooMovie)
  - 友善爬蟲(善待對方的主機，不可造成負擔)

爬蟲前置工作

- 整頁或是特定資訊
  - 網頁的規律程度
  - 資料乾淨程度
- 觀察URL
  - 尋找規律ID
- 檢查網頁原始碼
  - 觀察可抓的資料
  - 規律性
  - 基本上黑字就是可以抓

宣告

```
# -*- coding: utf-8 -*-
import urllib.request, urllib.parse, urllib.error
# 令一變數起始IDFirst
global IDFirst
IDFirst=942599
# 令一變數myURL，為URL
global myURL
myURL="http://udn.com/news/story/6928/"
# 令一變數storage，為儲存位置
global storage
# 使用HTMLParser
import html.parser
```

建立主程式

```
print ("The program is for parser full information")
print ("Copyright  2014 Soochow University, CHI-LUN HUANG. All rights reserved."
print ("processing start... ")
print ("Goal:"+str(IDFirst))

storage= "D:\YahooInfo\Data" +str(IDFirst)+".txt"
url =(myURL+str(IDFirst))
ExampleWeb = urllib.request.urlopen(url)
webContent = ExampleWeb.read().decode('utf_8')
ExampleWeb.close()

#try:
    # 開啟檔案在尾端加入東西
fout = open (storage, "w",encoding=("utf-8"))
fout.writelines('%s ' % str(IDFirst))
fout.writelines('\n')

Parser = HTMLParserExample()

try:
    # 將網頁內容拆成一行一行餵給Parser
    for line in webContent.splitlines():
        Parser.feed(line)
except (html.parser.HTMLParseError, data):
    print(("# Parser error : " + data.msg  ))

Parser.close()
fout.close()


print ("processing End... ")
```

Storage:儲存位置

url:網址

使用套件:解碼，獲取網頁資料

Assign class

讀網頁:跟讀檔相像，一行一行的讀取並呼叫class

建立程式(class)

- 爬蟲套件說明

```
# 解析器，繼承自HTMLParser
class HTMLParserExample(html.parser.HTMLParser):

    # 用來抓標籤的資料
    def handle_data(self, data):

        fout.writelines('%s ' % data)
        fout.writelines('\n')
```

合併程

ParserExample.py - C:\Users\user\Desktop\software\ParserExample.py (3.4.2)

File Edit Format Run Options Windows Help

```
# -*- coding: utf-8 -*-
import urllib.request, urllib.parse, urllib.error
# 令一變數起始IDFirst
global IDFirst
IDFirst=942599
# 令一變數myURL, 為URL
global myURL
myURL="http://udn.com/news/story/6928/"
# 令一變數storage, 為儲存位置
global storage
# 使用HTMLParser
import html.parser


# 解析器, 繼承自HTMLParser
class HTMLParserExample(html.parser.HTMLParser):
    # 用來抓標籤的資料
    def handle_data(self, data):


        fout.writelines('%s ' % data)
        fout.writelines('\n')

print ("The program is for parser full information")
print ("Copyright  2014 Soochow University, CHI-LUN HUANG. All rights reserved.")
print ("processing start... ")
print ("Goal:"+str(IDFirst))

storage= "D:\YahooInfo\Data" +str(IDFirst)+".txt"
url =(myURL+str(IDFirst))
ExampleWeb = urllib.request.urlopen(url)
webContent = ExampleWeb.read().decode('utf_8')
ExampleWeb.close()

#try:
    # 開啟檔案在尾端加入東西
fout = open (storage, "w",encoding=("utf-8"))
fout.writelines('%s ' % str(IDFirst))
fout.writelines('\n')

Parser = HTMLParserExample()

try:
    # 將網頁內容拆成一行一行餵給Parser
    for line in webContent.splitlines():
        Parser.feed(line)
except (html.parser.HTMLParseError, data):
    print(("# Parser error : " + data.msg  ))

Parser.close()
fout.close()

print ("processing End... ")
```

作業

- 以聯合新聞網為例
  - http://udn.com/news/story/6928/942599
- 增加三個功能
  - 新增迴圈使用range來大量爬取網頁
  - 新增篩選門檻與”東吳大學”有關的新聞
  - 新增寫檔將與東吳大學有關的網址寫入txt

迴圈使用range

- 要有開始與結束值
- global IDFirst
- IDFirst=942599
- global IDLast
- IDLast=942600
- 加入for (因為多了一個for 下方程式位階要改變)
- for i in range(IDFirst,IDLast):

- url =(myURL+str(i))

篩選門檻

- global match
- match=False
- if ('東吳大學' in data):
- 　　global match
- 　　match = True

寫檔

- if match==True:
- 	match=False
- 	fout.writelines('%s ' % url)
- 	fout.writelines('\n')

執行Demo

建議參考資料

- 開發者群組
- %s%d解釋
- 官網常用的地方
- 寫檔方式(r_a_w)
- 爬蟲套件說明