1. 先判別是否有可獨立取得網站資料JSON的API

如果沒有，下載 Python BS4

如果有，直接連接此網頁的JSON API

1. 請求 「http的網路請求」，測試是否取得必要資料

爬蟲流程

如果無法請求，透過Python呼叫瀏覽器，以取得資料

1. 資料完成抓取後，安裝python連結資料庫的套件

在資料夾cmd中下以下指令

* pip install pymysql

(conda install pymysql)

1. 安裝 python時間套件

* 在資料夾cmd中下以下指令

pip install apscheduler

(conda install apscheduler)

其他方法(最省電腦效能)：

透過網頁的res請求，取得HTML或Json資訊

Import requests

schools = requests.get(

'https://parttime.1111.com.tw/includesU/category/llllmenuCollege.js', headers={

'User-Agent': 'Mozilla/5.0 (Linux; Android 5.0; SM-G900P Build/LRX21T) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.149 Mobile Safari/537.36'

}, verify=False)

if not schools.status\_code == 200:

self.getSchool()

areas = re.search(r'var jsonCollegeCat=\'(.\*?)\'',

schools.text.replace('\n', '')).group(1)

self.areas = json.loads(areas)['n']

res = requests.get("https://www.1111.com.tw/job/{}/".format(id) ,headers={

'User-Agent': 'Mozilla/5.0 (Linux; Android 5.0; SM-G900P Build/LRX21T) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.149 Mobile Safari/537.36'

}, verify=False)

workBs = bs(res.text, "html.parser")

1. Python BS4 (網頁分析)、LXML(網頁分析)

爬蟲工具用法

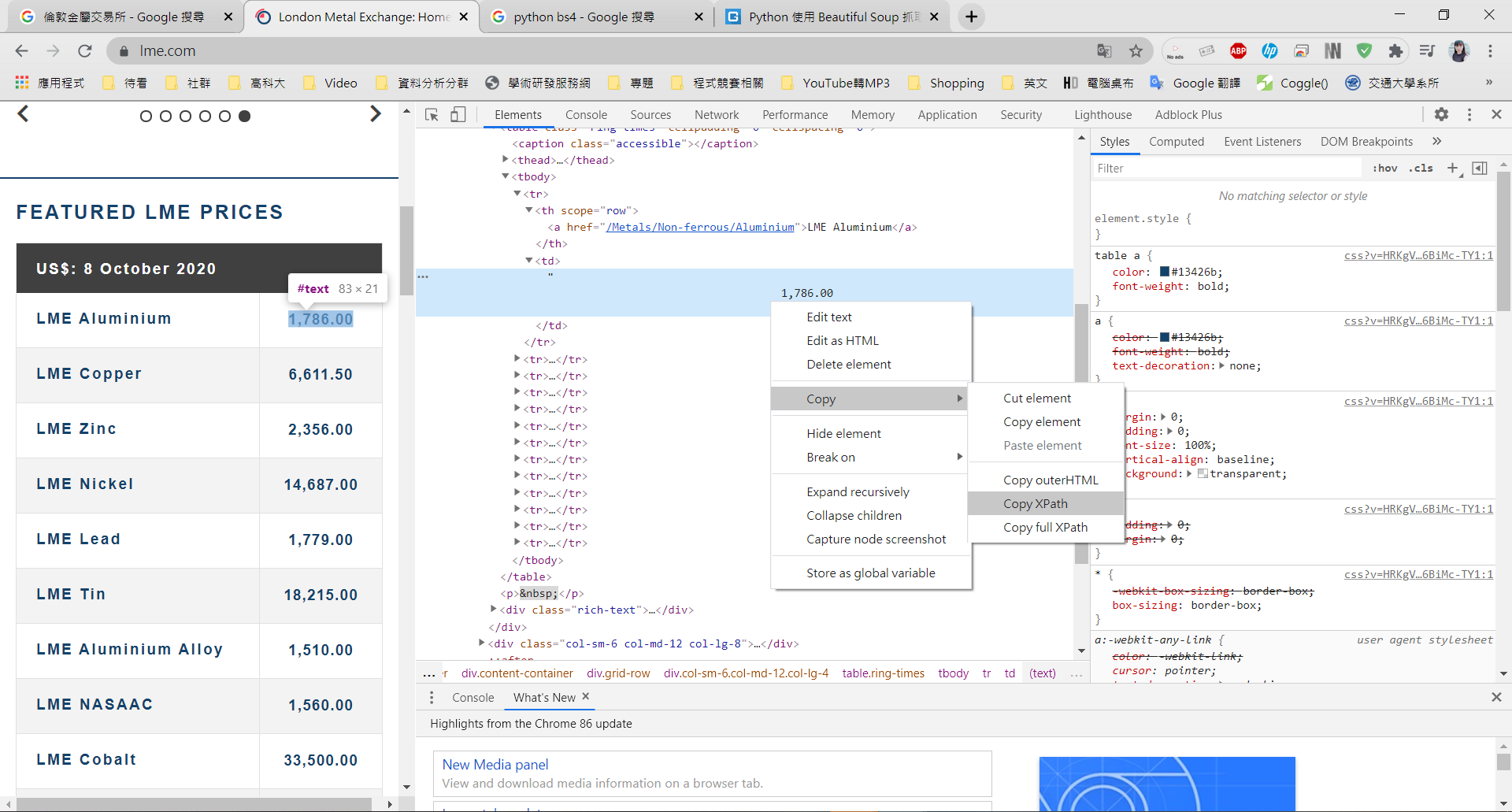
* 安裝 Python BS4

<https://blog.gtwang.org/programming/python-beautiful-soup-module-scrape-web-pages-tutorial/>

* 安裝 LXML

<https://medium.com/%E9%B3%A5-crl/python%E9%9A%A8%E7%AD%86-requests-lxml%E5%9F%BA%E6%9C%AC%E7%88%AC%E8%9F%B2-%E4%BB%A5%E5%8D%9A%E5%AE%A2%E4%BE%86%E5%8D%B3%E6%99%82%E6%8E%92%E8%A1%8C%E6%A6%9C%E7%82%BA%E4%BE%8B-f9c67de0644e>

* 做網頁路徑的分析(到要分析的網站，到要的資料區域打開F12)



取得資料位置(從紅框中的選一個使用Copy)

* 確定資料是否會因網址變化而變動

如果不會變動，就能夠正常撈資料

如果會變動，每一次撈資料都要到這裡取得新位置

Copy結果(此選用 XPath)：/html/body/div/div[2]/div/div[1]/table/tbody/tr[1]/td/text()

* 在Python 程式碼中使用

selector = etree.HTML(取得網頁原始碼)

// etree.HTML(取得網頁原始碼) 為 Python爬蟲套件

TargetString = selector.xpath('/html/body/div/div[2]/div/div[1]/table/tbody/tr[1]/td/text()')

// selector.xpath('貼上剛剛的Copy XPath')

1. Python呼叫瀏覽器

* 安裝

<https://kknews.cc/zh-tw/code/8z6enqn.html>

在專案跟目錄建立 空的python資料夾

在終端機下 pip3 install selenium 指令

* 安裝python呼叫瀏覽器的驅動程式

<https://sites.google.com/a/chromium.org/chromedriver/downloads>

下載完後，將執行檔放置python資料夾中即可