成為初級資料分析師 | Python 與資料科學應用

網頁資料擷取的開發環境

郭耀仁

Agenda

- Python 開發環境
- 虛擬環境
- 爬蟲所需模組與 Chrome 外掛

Python 開發環境

分瀏覽器與本機兩類

- 瀏覽器
    - Google Colaboratory
- 本機
    - Miniconda

直接在瀏覽器使用

Google Colaboratory (https://research.google.com/colaboratory/faq.html)

操作步驟

- 登入 Google 帳號，開啟雲端硬碟
- 點選「新增」
- 點選「連結更多應用程式」
- 搜尋「Colaboratory」
- 點選「連結」
- 新增 Google Colaboratory
- 完成瀏覽器的開發環境建置

操作步驟截圖

建立瀏覽器的開發環境 (https://www.datainpoint.com/python-essentials/01-quick-start.html#%E5%BB%BA%E7%AB%8B%E7%80%8F%E8%A6%BD%E5%99%A8%E7%9A%8

安装 Miniconda

Miniconda (https://docs.conda.io/en/latest/miniconda.html)

安装 jupyter

```
pip install jupyter
```

我為什麼推薦使用 Miniconda?

- 與 Anaconda 相比輕量許多
- 有 Python 直譯器
- 有套件與環境管理工具 conda

安裝 Miniconda 注意將 Python 加入路徑變數

虛擬環境

建立適當的虛擬環境

為了避免套件版本衝突，可以建立資源隔絕的虛擬環境

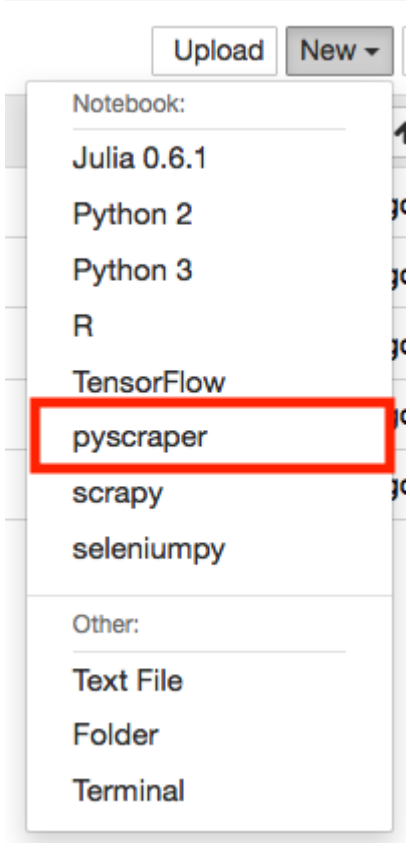
```
conda create -n pyscraper python=3.6
conda activate pyscraper
pip install ipykernel requests beautifulsoup4 pyquery selenium
```

建立對應虛擬環境的 Kernel

連結 Jupyter Notebook 與虛擬環境

```
python -m ipykernel install --user --name pyscraper --display-name "pyscraper"
conda deactivate
jupyter notebook
```

新增一個 pyscraper 的 Notebook



移除 Kernel

```
jupyter kernelspec list
jupyter kernelspec remove pyscraper
```

移除虛擬環境

```
conda deactivate
conda env list
conda remove --name pyscraper --all
```

爬蟲所需模組與 Chrome 外掛

爬蟲 Python 模組

模組名	功用
`requests`	對網站發送請求獲取資料
`pyquery`	對網站發送請求獲取資料與解析 HTML 資料
`Beautifulsoup4`	解析 HTML 與 XML 資料
`selenium`	自動化瀏覽器

爬蟲 Python 模組文件連結

- requests (http://docs.python-requests.org/en/master/)
- pyquery (https://pythonhosted.org/pyquery/)
- beautifulsoup4 (https://www.crummy.com/software/BeautifulSoup/bs4/doc/#)
- selenium (https://selenium-python.readthedocs.io/)

爬蟲 Chrome 外掛

外掛名	功用
Quick Javascript Switcher	關閉 JavaScript 功能
JSONView	讓 JSON 資料格式在瀏覽器上呈現得比較漂亮
SelectorGadget	協助 CSS Selector 定位
XPath Helper	協助 XPath 定位
EditThieCookie	觀察 Cookies

爬蟲 Chrome 外掛連結

- Quick Javascript Switcher (https://chrome.google.com/webstore/detail/quick-javascri switcher/geddoclleiomckbhadiaipdggiiccfje)
- JSONView (https://chrome.google.com/webstore/detail/jsonview/chklaanhfefbnpoihckbnefhakg
- SelectorGadget (https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjo
- XPath Helper (https://chrome.google.com/webstore/detail/xpath-helper/hgimnogjllphhhkhlmebbmlgjoejdpjl)
- EditThisCookie (https://chrome.google.com/webstore/detail/editthiscookie/fngmhnnpilhplaeedifhccc