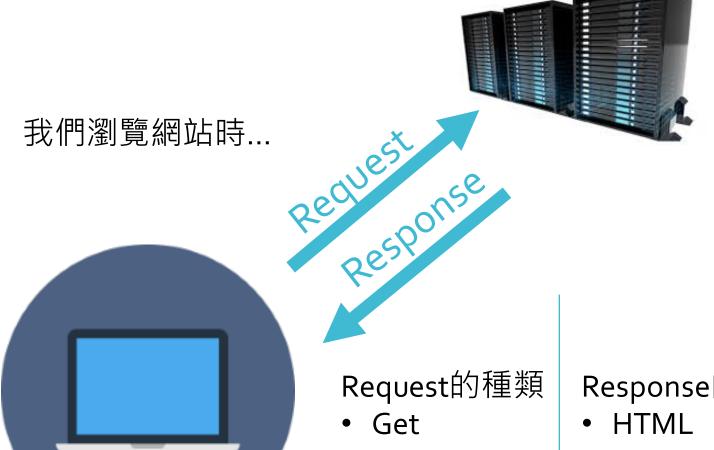




爬蟲入門

網路 運作原理



- Post

Response的種類:

- Json \ CSV

用python假裝是網頁瀏覽器 把資料抓下來

爬蟲 三大步驟

觀察下載

處理

觀察網頁 與伺服器對話 抓下來的 資料慢慢整理

爬蟲 三大步驟



初學盡量不要邊下載邊處理

觀察
下載
處理
下載
處理

全部下載好,全部處理!

觀察 下載 下載 處理 處理

爬蟲 三大步驟

• 優點:

- 簡單分隔程式碼
- 下載的資料都存好好的
- •一次下載完,縱覽所有資料格式不一致
- •程式當掉了,資料不用重爬

觀察

- 先手動看瀏覽器怎麼爬資料:
 - · Chrome → 開發者工具 → 檢查 → Network (網路)
 - 執行想要爬蟲執行的動作
 - · 觀察是Get/Post
 - · 觀察參數(Query parameters / Form data)

下載

- 下載資料
 - ·試試看requests (能於80%以上的網站)
 - 下載的網頁內容不完整
 - · → 可能是網頁用了 Javascript
 - · 改用selenium (安裝比較複雜)

資料整理

- HTML裡面的Table → Pandas.read_html
- CSV → Pandas.read_csv
- JSON → Json.dumps
- ・HTML中的某個元素 → beautifulsoup

下載 課程範例

- Desktop/finlab_course/
 - Course6.ipynb

謝謝大家的收看我們下個單元見!