# 數據處理練習

數據分析 2016 楊舜宇

# 練習簡介

在這個練習中，你們將會撰寫程式來處理與分析網站log，共有三個部分，第一項過濾出原始log中含有oid的紀錄，第二項為擷取出原始log的資料欄位，第三項為統計各個IP當日的流量，第四項為視覺化呈現流量分布。藉由這個練習你們能知道如何去擷取、處理與視覺化資料，以利後續分析。

# 網站log資料

練習所使用的log為中研院數位典藏資源網1，這是一個介紹歷史文物與動植物資訊的網站，log檔案存在 ”\2016實習\楊舜宇\log\數位典藏網\” 中，裡面有2015/6/1 ~ 2016/7/20的log，練習時請將檔案下載到本地端。

單條log資料如下所示:

157.55.39.22 - - [18/Jun/2016:00:09:20 +0800] "GET /content/repository/resource\_content.jsp?oid=1123204&queryType=qc HTTP/1.1" 200 37009

格式由7個部分組成，依序為IP、請求時間、HTTP Method、請求Url、HTTP協定版本、HTTP回應狀態碼、回應資料大小(bytes)，其中請求Url有包含oid2是我們主要分析的內容，其他紀錄則是載入網頁其他資源(如: 圖片)所留下，並不重要。

註: 中研院數位典藏資源網1 (<http://digiarch.sinica.edu.tw/>)

oid 2:文物、生物資訊網頁編號

# 練習一

這個部分要你們撰寫一個Linux Shell腳本，將包含oid的log條過濾出來，並輸出成檔案。首先為了在Window下建立shell執行的環境，請先安裝Cmder，再來將log檔案從NAS上複製到本地端，不要直接修改原始資料，完成以上步驟後，即可開啟文字編輯器撰寫腳本，檔名存成logFilter.sh。

你的程式必須能指定log來源資料夾與輸出資料夾，執行後將每個log檔案中包含oid的log過濾出來，並分別輸出到輸出資料夾的對應檔案中。

程式輸入: log來源資料夾、輸出資料夾

使用方式: ./logFilter.sh –s <來源資料夾> -o <輸出資料夾>

幫助訊息: 若程式參數輸入錯誤不完整或輸入-h參數，顯示幫助訊息  
 幫助訊息要包含以下項目:

1. 程式名稱
2. 作者
3. 撰寫日期
4. 程式簡介
5. 參數操作說明
6. 使用範例

提示: 善用linux指令grep

# 練習二

這個部分要你們使用python或java撰寫一隻資料擷取的程式，將原始log中的文字轉換成可供機器讀取的csv檔。你們必須使用regex(正規表達式)來擷取資料，並將各個欄位以適當的變數類型儲存，最後將資料分別存到對應的csv檔案中。

開發環境建設:

Java:

Oracle JDK 8 (記得設環境變數)

Intellij (IDE)

Python:

Anaconda python 3

PyCharm (IDE)

程式碼要求:

1. 請在程式碼開頭撰寫程式基本資訊

* 作者
* 撰寫時間
* 程式功能
* 如何使用

1. 不可以將所有程式碼寫在單一main function處理
2. 資料要以對應格式儲存，不可以全部都存成字串
3. 變數命名要有意義

提示:

Python: 請善用內建模組 re, csv, dict, datetime, dateutil

Java: regex, SimpleDateFormat, Map

進階功能:

這部分要你平行化處理資料擷取程式，請使用java來實作程式碼，將原本的parse的function改成一個Runable任務或繼承Thread，並利用java ExecutorService來自動平行畫這些任務，並計算程式所消耗的時間。

另外，為了瞭解現在執行的進度，你必須使用一個AtomicInteger的變數來計算已經完成的任務數量，並使用另一個thread每隔一秒輸出現在進度，輸出格式如下:

<輸出時間> 已完成<已完成任務數量>/<總任務數量>個任務 經過<已經過時間> seconds

Ex.

2017/2/7 11:00:23 已完成45313/52151個任務 經過2 seconds

說明: AtomicInteger是一個thread safe的資料型態，可在多個thread間被共同存取而不發生問題。

# 練習三

這個部分要你們統計練習二輸出的csv，統計各個IP在當日瀏覽的次數，最後一樣輸出成csv檔，格式為IP, 次數。

# 練習四

為了瞭解大部分正常人每日瀏覽網頁的數量，這個部分你們要使用Python matplotlib來呈現瀏覽次數的分布。你們必須先統計相同瀏覽次數的數量，並使用長條圖，以5次分一組顯示。

圖表標題: 瀏覽次數分布

X軸: 瀏覽次數

Y軸: 相同瀏覽次數的數量

X軸間隔: 5

Y軸間隔: 1000

參考: <http://matplotlib.org/examples/statistics/histogram_demo_histtypes.html>