

00

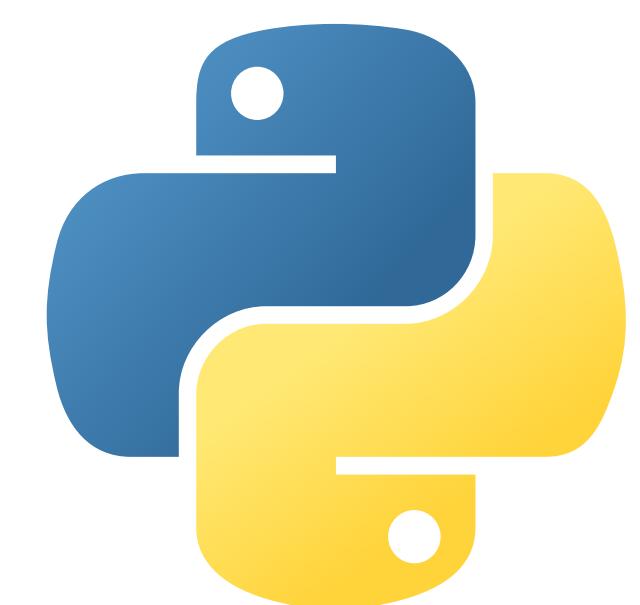
Day
38-40

期末專題

cupay



出題教練：楊鎮銘



python

專題時程：2020/4/13 (一) 至 2020/4/27 (一)

Python網路爬蟲 期末專題時間

結果發布：2020/5/18 (一)

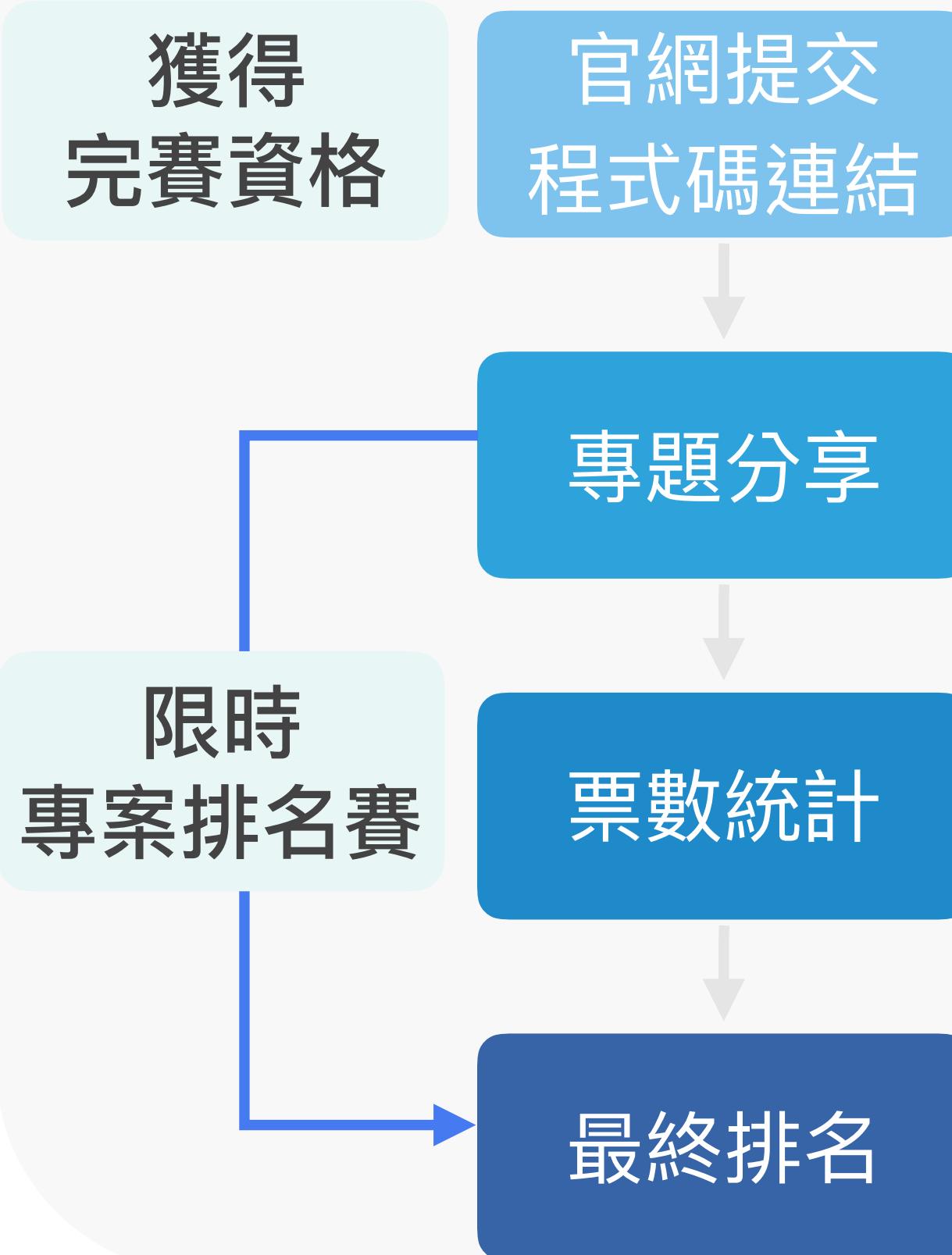
04

APRIL

S	M	T	W	T	F	S
				1	2 D33-D35	3
5	6 D36-D37	7	8	9	10	11
12	13 期末專題 開始 D38-D40	14	15	16	17	18
19	20	21	22	23	24	25
26	27 期末專題 結束 D38-D40	28	29	30		

期末專題進行流程

完賽流程



完成專題後請您將您的程式碼放置 github，並回到官網提供連結完成作業提交

請於共學社團建立期末專案文章分享

統計共學社團成員和駐站專家按讚票數

獲得票數最高的前 10% 名學員將獲得期末排名紀念品

如何建立期末專題：方法一 發布文章

step1：前往共學社團建立文章

The screenshot shows the CUPOY platform interface for publishing an article in a study group. The steps are highlighted with red boxes and numbered 1, 2, and 3.

- 點擊貼文** (Step 1): The "Post" tab is highlighted with a red box.
- 點擊文章** (Step 2): The "Article" button is highlighted with a red box.
- 點擊完整版** (Step 3): The "Full Version" button in the rich text editor toolbar is highlighted with a red box.

The interface includes a sidebar with group statistics (5 posts, 216 answers, 22 announcements), member lists, and a news feed at the bottom.

如何建立期末專題：方法一 發布文章

step2：用文章形式分享期末專題內容

Python期末專題 Cupoy (格式參考)

取消 預覽 完成

編輯 插入 格式

① 標題撰寫格式

一般段落 Python期末專題|百日馬拉松顯示名稱

一、專題摘要 (解釋實作與說明需要解決的問題，限300~500字。)

1. 期末專題主題
2. 期末專題基本目標

二、實作方法介紹 (介紹使用的程式碼、模組，並附上實作過程與結果的截圖，需圖文並茂。)

1. 使用的程式碼介紹
2. 使用的模組介紹

三、成果展示 (介紹成果的特點為何，並撰寫心得。) ② 期末專題文章基本格式內容

四、結論 (總結本次專題的問題與結果)

五、期末專題作者資訊 (請附上作者資訊)

1. 個人Github連結
2. 個人在百日馬拉松顯示名稱

如何建立期末專題：方法二 分享簡報

在 Cupoy 建立文章，內容可使用公開簡報分享。

學員可以使用第三方簡報分享平台，譬如 slideshare 完成本次期末分享內容。



The screenshot shows a presentation slide from SlideShare. The title of the slide is "BeautifulSoup 用法". Below the title, there is a code block in Python syntax:

```
# 用 html.parser 分析 response.text · 並存入 soup 中
soup = BeautifulSoup(response.text, 'html.parser')

# 搜尋所有 html 中標籤為 div 且 class 為 r-ent 的目標
soup.findAll('標籤', 'class')
```

At the bottom of the slide, there is a logo for "北科程式設計研究社" and navigation controls for the presentation. The status bar at the bottom right indicates "37 of 46" and "414 views".

如何建立期末專題：方法二 分享簡報

公開分享至 slideshare 後，再回到 Cupoy 共學社團建立文章，簡述流程並分享簡報連結。



Python期末專題 Cupoy(格式參考)

取消 預覽 完成

編輯 插入 格式

① 標題撰寫格式

一般段落 Python期末專題 百日馬拉松顯示名稱

一、專題摘要 (解釋實作與說明需要解決的問題，限300~500字。)

1. 期末專題主題
2. 期末專題基本目標

二、實作方法介紹 (介紹使用的程式碼、模組，並附上實作過程與結果的截圖，需圖文並茂。)

1. 使用的程式碼介紹
2. 使用的模組介紹

三、成果展示 (介紹成果的特點為何，並撰寫心得。) ② 期末專題文章基本格式內容

四、結論 (總結本次專題的問題與結果)

五、期末專題作者資訊 (請附上作者資訊)

1. 個人Github連結
2. 個人在百日馬拉松顯示名稱

如何獲得分數

建立完成文章公佈在社團後，可獲取其他學員與專家like，即為您的期末排名分數

The screenshot shows a community page on the Cupay platform. The top navigation bar includes links for 首頁 (Home), 熱門新聞 (Hot News), 主題社群 (Topic Communities), 我的 (My), 搜索 (Search), and a red-highlighted 投稿 (Post) button. A notification icon with a '3' is visible.

The main content area displays a community profile for "深度學習與電腦視覺學習馬拉松專屬社團". The profile includes a banner for an activity from 2019/11/25 to 2020/02/17, a summary of 3 posts, 199 answers, and 21 announcements, and member statistics (458 members).

Below the profile, there are three posts by user "楊哲寧":

- 影像前處理、標籤前處理與Loss Function(Day15)**: Includes a detailed description about image normalization and loss functions, and a link to a Colab notebook.
- 運用Colab練習程式與訓練模型**: Describes using Colab for training models, mentioning Cifar10 and Breaking Captcha datasets.
- Python (jupyter notebook、keras) 環境架設**: Provides instructions for setting up Python environments using Jupyter notebooks and Keras.

Each post has engagement metrics: likes, comments, and shares. The first post's like count (2) is highlighted with a red box. A callout at the bottom of the post list encourages users to like posts to earn points.

On the right side, there is a sidebar for creating new communities, featuring a "測試" community with 16 members and a "建立社區" button.

期末專題文章架構

請學員分享你的專題實作結果，格式不拘但請包括下列內容：

01

專題摘要

解釋實作與說明需要解決的問題，
限300~500字。

02

實作方法介紹

介紹使用的程式碼、模組，並附上實作過程與結果的截圖，需圖文並茂。

03

成果展示

介紹成果的特點為何，並撰寫心得。

04

結論

總結本次專題的問題與結果。

期末專題主題

本次期末專題您可從以下兩個主題中**擇一**進行：



Cupoy

官網新聞



Ptt

討論版

期末主題 1 - Cupoy

專案目標：

請任選 Cupoy 新聞服務之某一種分類 (如熱門新聞、科技、商業....)，使用你學習過的爬蟲程式，爬取前 500 篇的文章：

<https://www.cupoy.com/newsfeed/topstory>

The screenshot shows the Cupoy news feed interface. At the top, there's a navigation bar with links for '首頁', '熱門新聞', '主題社群', and '我的'. Below the navigation is a search bar and a red '投稿' button. On the left, there's a sidebar with '我的訂閱' (My Subscriptions) and '分類' (Categories). The '熱門新聞' (Top News) section displays two articles: one about a plane crash and another about Xiaomi's Poco子品牌. To the right, there's a '新聞來源' (News Sources) section listing various media outlets like 鏡週刊, Yahoo!奇摩, 上報UpMedia, and ETtoday運動, along with their update counts and follower counts. A '熱門標籤' (Hot Tags) section at the bottom lists tags such as 龍頭, 龍華廠, 龍洞, 鼠年, 點擊, 點加強, 黑色, 黑田, 黑歷史, and 黑人.



期末專題實作提示 (基本目標)

TARGET 1

透過開發者工具觀察網站在列出 News Feed 這邊是屬於動態網站還是靜態網站，或是有 API 可以直接送 requests

TARGET 2

根據網站特性選擇 requests / BeautifulSoup / selenium 等工具進行爬蟲整理

TARGET 3

整理成 pandas.DataFrame 後做簡單的統計可以用 matplotlib.pyplot 或是 pandas 內建的 function 畫圖 (histogram / pie chart ...)

期末專題實作提示 (進階目標)

TARGET 1

爬下文章，透過 jieba 等斷詞將文章拆解

TARGET 2

可以簡單的計算同樣文字出現的頻率或是透過 TFIDF 的統計方式計算

TARGET 3

將經常出現的 stop words 過濾掉之後對頻率進行排名

TARGET 4

將結果透過 wordcloud 文字雲的方式呈現

期末主題 2 - Ptt

專案目標：

根據版的熱門程度跟屬性，可選定以下任一種：

1. 八卦版：<https://www.ptt.cc/bbs/Gossiping/index.html>
2. 政黑板：<https://www.ptt.cc/bbs/HatePolitics/index.html>

批踢踢實業坊 › 看板 Gossiping

看板 精華區

搜尋文章…

Re: [新聞] 民眾黨徵助理起薪30K被罵翻 柯文哲：雇少 XSR700

1 Re: [新聞] 年前全漲價！從珍奶到鍋貼 小數點也要賺 popy8789

2 Re: [新聞] 民調：4成5反對蔡英文兼任黨主席 逾7成2 Rrrxddd

1 Re: [新聞] 館長道歉林右昌沒答應展店 「他們可能是 NuclearSnake

3 [新聞] 川普德州取暖之旅 保證美中貿易協議將嘉 CavendishJr

批踢踢實業坊 › 看板 HatePolitics

看板 精華區

搜尋文章…

Re: [討論] 柯黑只會吹捧817萬？？？ devidevi

[討論] 罷免韓之後，再來要罷免小英 csfgsj

[討論] 侯漢廷共謀？ canttorati

5 [黑特] 親藍政論節目現在全力轟郭 哪招？ kapasky

Re: [新聞] 年輕人偏好投給第三勢力？近3成首投族投 IronCube

1 [討論] 助理薪水問題 paladin90974



期末專題實作提示 (基本目標)

TARGET 1

爬下文章，透過 jieba 等斷詞將文章拆解

TARGET 2

可以簡單的計算同樣文字出現的頻率或是透過 TFIDF 的統計方式計算

TARGET 3

將經常出現的 stop words 過濾掉之後對頻率進行排名

TARGET 4

將結果透過 wordcloud 文字雲的方式呈現

期末專題實作提示 (進階目標)

TARGET 1

透過不同帳號，但是相同 IP 且政治用語的詞頻分佈類似的定位成網軍

TARGET 2

進一步分析帳號是否在特定期間 (e.g. 選舉) 有明顯的活動特性

TARGET 3

如果不同帳號但是政治用語的詞頻分佈類似，進一步判斷這些高頻率的單字是 positive / negative 來歸納兩個帳號之間是否具有相同政治立場

期末專題知識點目標

- 專題結束後你可以學會
 - 了解不同網站實作的爬蟲細節
 - 對於爬蟲流程的分析與判斷有完整的 Overview
 - 可以分析針對不同網站所需的爬蟲複雜度
 - 搭配不同領域知識做出獨特的應用
 - 清楚說明爬蟲流程與作法



完賽時間

IT'S YOUR ACHIEVEMENT

請跳出 PDF 至官網 Sample Code & 作業
開始解題

