

# D16：實作 Day：Wiki的爬蟲實作練習



簡報閱讀



範例與作業



問題討論

Wikipedia爬蟲練習

重要知識點

重要知識點複習

解題時間

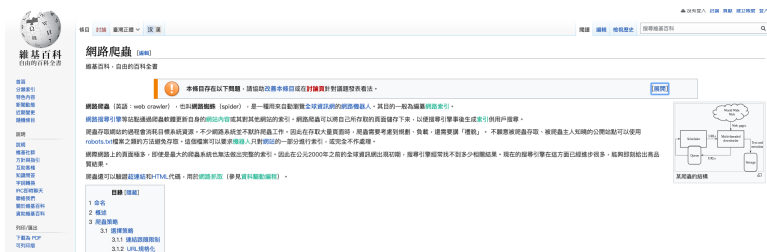
## Wikipedia爬蟲練習



出題教練：張齊文

## 重要知識點

### Wikipedia(維基百科)爬蟲



範例1：選定一個關鍵字，爬取該關鍵字的文章內容。

爬蟲存取網站的過程會向伺服器系統提問，不少網頁系統並不設計回應工作，因此存取取大量資料時，爬蟲需要考慮到時間、負載、搜尋策略「權限」，不願意被伺服器存取、被伺服器主人知曉的公開連結可以使用 robots.txt 檔案之操作方式避免存取，這個檔案可以讓系統知道人只對網站的一部分資料有權利，爬蟲完全不能存取。

網頁網路上的資源極多，即使是最大的網路系統也無法輸出完整的索引，因此在公元2000年之前的全球網頁輸出初期，搜尋引擎經常找不到多少相關結果，現在的搜尋引擎在這方面已經進步很多，能夠輸出結果品質較好。

爬蟲還可以透過利用伺服器 HTML 代碼，爬出網站內容（參見資料結構與編碼）。

目錄 (目錄)

- 1 命名
- 2 權限
- 3 爬蟲策略
- 3.1 搜尋策略
- 3.1.1 網路資源限制
- 3.1.2 LPL 爬蟲

## 範例2：擷取文章中，延伸出的外部連結關鍵字。

維基百科

網路爬蟲

網路爬蟲 (英語：web crawler)，也叫做網路蜘蛛 (spider)，是一種用來自動抓取網頁內容的程式。爬蟲可以將自己存取到的資料儲存下來，以便搜尋引擎事後生成索引供用戶搜尋。

爬蟲存取網站的過程會向伺服器系統提問，不少網頁系統並不設計回應工作，因此在存取大量資料時，爬蟲需要考慮到時間、負載、搜尋策略「權限」，不願意被伺服器存取、被伺服器主人知曉的公開連結可以使用 robots.txt 檔案之操作方式避免存取，這個檔案可以讓系統知道人只對網站的一部分資料有權利，爬蟲完全不能存取。

網頁網路上的資源極多，即使是最大的網路系統也無法輸出完整的索引，因此在公元2000年之前的全球網頁輸出初期，搜尋引擎經常找不到多少相關結果，現在的搜尋引擎在這方面已經進步很多，能夠輸出結果品質較好。

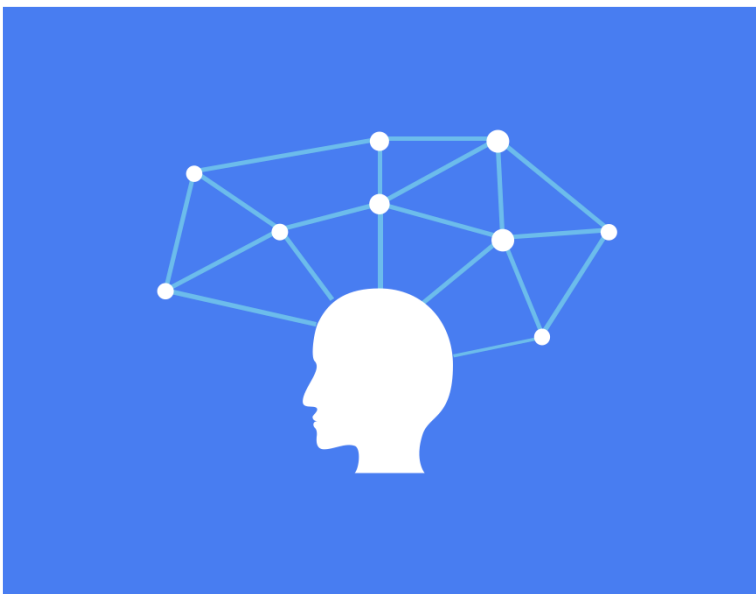
爬蟲還可以透過利用伺服器 HTML 代碼，爬出網站內容（參見資料結構與編碼）。

目錄 (目錄)

- 1 命名
- 2 權限
- 3 爬蟲策略
- 3.1 搜尋策略
- 3.1.1 網路資源限制
- 3.1.2 LPL 爬蟲

本文

## 重要知識點複習



練習：定義一個爬蟲函數，重複前面兩個步驟的流程，可遞迴爬取更多關鍵字的解釋文章。其流程如下：

- 爬取當前關鍵字的解釋，並存入檔案(因為文章內容太多會佔滿整個頁面，所以存程檔案，方便後續檢視)。
- 萃取出當前關鍵字所引用的外部連結，當作新的查詢關鍵字。

## 解題時間



Sample Code & 作業  
開始解題



[下一步：閱讀範例與完成作業](#)