

## 實務專題計畫摘要報告 YOLO Travel

專題編號：107-CSIE-S018

執行期限：106年第1學期至107年第1學期

指導教授：王正豪

專題參與人員：104820006 莊秉凡

104820010 賴曉柔

104820014 鄭宜文

### 一. 摘要

本專題將開發一個蒐集各大旅行社旅遊資訊的網站：YOLO Travel。透過網路爬蟲與資料庫系統結合成之 Web Application。以自動排程的方式自動化爬蟲程式，減少人工執行的重複性，並提供使用者簡單明瞭的操作介面，使用基於內容的簡易推薦系統。

YOLO Travel 網站主要以 HTML 與 JavaScript 作為前端，並以 Python 為後端的架構進行實作，透過 Django 框架建構網站功能並整合爬蟲程式及資料庫。  
關鍵詞：Recommender System、Django、Web Crawler、Sqlite、Spark

### 二. 緣由與目的

我們發現現今旅行社雖然都有線上網站可以搜尋該社全部的行程，卻沒有一個可以讓人們一次比較各個旅行社相似行程的平臺。因此我們小組打算開發一個旅行社行程的比價網站，蒐集各大旅行社的最新行程，讓使用者可以選擇自己想要的旅行條件，並輕鬆的在不同旅行社之間做比較，在這個平台我們設計簡單、易於使用的介面，讓使用者以最高的效率，找到最理想的行程。

### 三. 研發工具

前端	後端	爬蟲
- HTML/CSS - Bootstrap	- Python - Django	- Requests 套件 - Beautiful Soup - Selenium

### 四. 研究報告內容

#### (一) 研究問題：

我們將設計簡單的操作介面，讓使用者能夠根據需求，輸入自己想要的條件，快速精準的篩選後，清楚地列出所有符合的項目。

Python 是個適合做資料處理以及分析的程式語言，也具有強大的運算處理能力。搭配 Python 多功能的框架，可以快速架設後端，而其前後端之間的關聯性簡單易懂容易操作，再加上資料庫的驅動優勢，因此我們選擇 Python 處理後端運算。

在前端的部份以 HTML/CSS 搭配 JavaScript 的形式來呈現，並透過 Bootstrap 框架來設計，其自動化調整頁面等功能，讓我們的 Web 應用更加方便與美觀，同時也能做出許多使用者介面元素的功能。

透過字彙分析工具將行程內容切割成許多關鍵字，再利用這些關鍵字做統計與月份做連結後，將結果作為推薦系統的參考依據，顯示出每個月最熱門的行程，以及每一行程的相關行程推薦。

通過這樣的設計，我們可以擁有明確的程式架構，讓使用者方便使用、開發者易於維護。

#### (二) 研究方法及步驟：

##### 1. 建立虛擬環境及 Django 設定：

為了讓專案中所使用到的 Python 套件便於管理，使用 Virtualenv 這個工具建立 Python 虛擬環境，就能讓專案套件

與其他專案及系統套件隔離，減少資源浪費與可能漏洞。透過 Django 的 MTV 模式有效率的控制資料庫和網站資訊的呈現，本專案使用 Django 預設之 Sqlite 輕量化關聯式資料庫，來存入與提取爬蟲後的網站資料，降低使用複雜度。

## 2. 爬蟲程式撰寫並存入資料庫：

本專案所有資料皆取自各旅遊網站，我們一個個瀏覽並了解每個網站之特性及網頁編排方式，依照不同格式撰寫專屬爬蟲程式，取得我們所需的資料後，存入資料庫。Django 使用 ORM 操作資料庫模型，把資料用物件導向的方式整理，易於使用與管理，加上 Django 有內建的後台網站，圖形化介面能使操作更加靈活，減少錯誤發生。

## 3. 排程任務與簡易推薦系統：

Django 擁有眾多的套件可供使用，其中一種原生套件 Django-Q 即為其好用的排程工具，將透過此工具與內建後台的結合來管理排程任務，我們預計以 24 小時為單位更新資料庫，減輕伺服器負擔。當取得資料後，以基於內容相似性的推薦方式分析資料的相似程度來實作資訊的呈現，將使用者可能有興趣的資訊排在前方供其瀏覽。

## 4. 前端網頁與後端伺服器的結合：

Django 的 MTV 模式對於動態網頁的撰寫提供良好的靈活度，特殊的 Template tags 增加程式碼的可擴充性。我們預計將完成後的網站部署到 AWS 之 EC2，並使用 nginx 及 uWSGI 來部署 Django 專案。

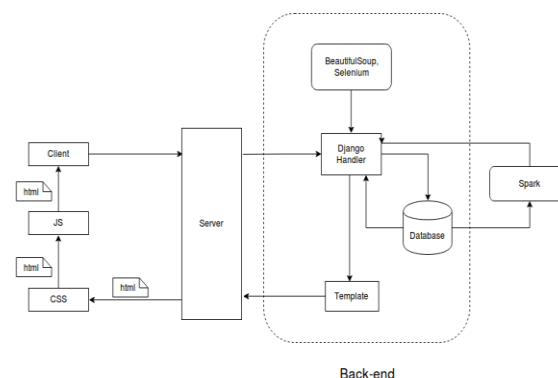
## 5. 字串切割與資料統計：

Jieba 是文本分析的工具，並支持自定義字典，所以可以更快速與準確的切割字詞。我們透過 Jieba 切割標題，將切割後的字詞紀錄並統計。

Spark 是個速度快且容易使用的運算工具，而且也支援我們後端所使用的 Python 語言，在統計上以及配合資料庫彙整都很方便。利用剛剛 jieba 切割完的字詞並透過月份做分類之後，再使用 Spark 做 MapReduce 統計所有的關鍵字，

可以觀察到每個月份較常出現的熱門字彙，從資料庫提出相關的行程，向使用者做推薦。

## 五. 系統架構圖



YOLO Travel 系統架構圖

## 六. 參考文獻

- [1] Royi Ronen, Noam Koenigstein, Elad Ziklik, Nir Nitzan "Selecting content-based features for collaborative filtering recommenders"
- [2] Keunho Choi, Yongmoo Suh "A new similarity function for selecting neighbors for each target item in collaborative filtering"
- [3] 江易軒 "A Design and Implementation of Platform to Create Multi-language Interface."
- [4] D. V. Annenkov and E. A. Cherkashin, "Generation technique for Django MVC web framework using the stratego transformation language," 2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2013, pp. 1084-1087.