**利用文字探勘與分類器技術**

**建置社群推文檢索工具**

指導教授:黃建宏

組員: 張國榮、呂嘉哲、簡孝齊、高承泰

**摘 要** **系統功能**

**旅遊文章分類：**讓使用者利用爬蟲軟體輸入一篇旅遊文章或者語音檔案，並由訓練完成的**NLTK**與**BERT**判斷該文章為三大分類中的哪一類。

**文章特徵詞提取：**將文章分類完成後，將根據其類別的語料庫進行 **TF-IDF** 特徵詞提取，萃取出該文章中較重要的關鍵字，除了可做為訓練特徵外，並結合網路爬蟲程式可以讓使用者明白當前旅遊的趨勢與熱門關鍵字。

**文章字詞相關性：**提取出關鍵字後，將列出文章特徵詞之間的相關性，並透過量化的數據來了解彼此的類別相關程度大小，並判斷是否為旅遊文章，讓使用者得知。

在現今流量巨大的各類文章中，如何快速

分辨出我們所需要的旅遊文章，並加以分

類。如何有效率的篩選出其中的精華成為

一項新的難題，而文字探勘近年因機器學

習的熱潮而水漲船高，我們希望透過文字

探勘來自動分析網路上的文章並加以分類

，來提高獲取資料的效率。

**研究動機**

由於現在3C產品的普及，人人皆有一機，

整天待在家裡不出門了解台灣的美實在太

可惜了，台灣的景點與美食都還沒吃遍玩

透，怎麼可以說沒地方玩了呢。為了讓台

灣使用者更方便的瀏覽旅遊相關資料，我

們希望透過抓取網路上的文章加以統整，

經過學習過後的分類器，整理出旅遊各個

標籤分類，進而從網路上提高獲取旅遊資

**介面圖**

料的效率。

