熱門話題分析系統

專題編號：108-CSIE-S010
執行期限：107年第1學期至108年第1學期
指導教授：王正豪
專題參與人員： 105590021 林敬勛
105590045 楊永健

一、摘要

本專題將開發一個能夠分析著名台灣電子布告欄（BBS）— PTT，了解台灣網友所關心的熱門話題，並提供熱門關鍵字、搜尋熱門文章、資料視覺化分析結果…等功能。

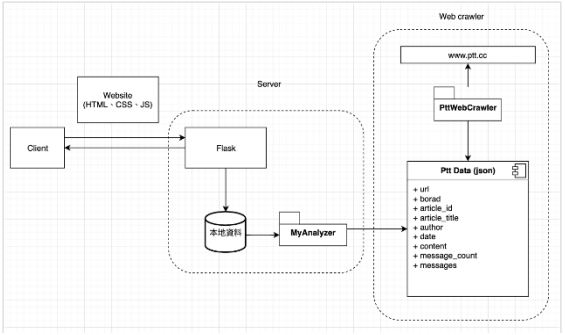
我們將會使用到兩個 Package，分別是 PttWebCrawler 和 MyAnalyzer。PttWebCrawler 用來爬取 PTT 網頁版上的資料，並將資料暫存在本機上，然後透過 MyAnayzer 分析文章內容，找出討論度高的文章，並透過切詞及推噓文的分析，找出文章內含有哪些詞容易影響到推噓文比例，最後透過 Flask 作為網站的後端，讓使用者能夠拜訪我們的網站，看到我們的分析結果。

網站主要以 HTML、CSS 作為前端架構，並利用 JavaScript 來會製圖表。

二、緣由與目的

在 ptt 中會根據不同的話題分成不同的看板，有討論運動的、討論八卦的、討論政治的、討論 3C 產品的…等等，而每個看板都有屬於他們的熱門話題，像是一到選舉八卦版就會開始熱烈討論，有新的手機資訊釋出，手機板上就會引發討論，而為了要跟上流行的腳步，確實知道當前的熱門話題，進而開發了這個系統。

三、系統架構



四、使用技術

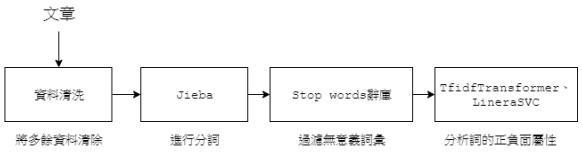
(一) PttWebCrawler

我們在這個 Package 中，我們使用 Request 來爬取網站上的資料，並使用 BeautifulSoup 把我們不需要的網頁語法去除掉，找出我們所需要資料，像是文章內容、日期、下面的留言…等等，並將這些結果輸出至本機並儲存成 json 檔方便 MyAnalyzer 分析。

而在這個 Package 中，有提供幾個在實作上會用到的 api，像是輸入 ptt 的起始頁和結束頁去爬取資料、利用日期去爬取資料…等等。

(二) MyAnalyzer

此 Package 主要會針對標題和內文，利用 Jieba 和 Stop Words 來做分詞和過濾無意義的詞彙。分析推文和留言的正面負面詞彙是依據文章的推文和噓文來做資料來源，利用 TfidfTransformer 做出特徵向量，配合 LinearSVC 進行預測訓練。熱門話題則是根據留言數來判斷。



(三) Flask

使用 Jinja 模板來渲染 Template，讓在後端處理好的數據能夠在前端讓使用者看到。

(四) 資料視覺化

使用 chartkick 來實現資料視覺化，透過給予一個型態為 dict 或 list 的資料型態，就可畫出美觀的 javascript 圖表，像是柱狀圖、圓餅圖、折線圖...等。

五、實作的功能

- 可依據起始頁數和結束頁數來爬取 PTT 上的資料。
- 可依據起始日期和結束日期來爬取 PTT 上的資料。
- 透過排程每天晚上去爬取看板上的資料。
- 利用 chartkick 畫出圖表。
- 可選擇特定區間來查看那段時間的熱門話題。
- 根據貼文來找出負面詞彙和正面詞彙。
- 根據回文來找出負面詞彙和正面詞彙。

六、未來展望

- 可新增管理者的權限，讓管理者可以直接透過網頁做爬取資料、分析資料的動作。
- 可爬取不同看板的資料，放到相對應的資料夾內，就可查看到不同看板的分析結果。

七、參考文獻

[1]"jieba-tw"
https://github.com/APCLab/jieba-tw
[2] "PTT 資料抓取與分析"
https://city.shaform.com/zh/2016/02/28/scrapy/
[3]"flask document"
http://flask.pocoo.org/docs/1.0/
[4]"chartkick"
https://github.com/mher/chartkick.py