數位語音處理概論

期末報告

電機四
謝宜展
B03901166

一、 Github 頁面

https://github.com/yzhsieh/Find_Article_By_Article

二、 動機

通常在各個網路論壇中，如果想要找尋某個議題的討論文章，只能使用論壇給予的搜尋功能來搜尋標題，但如果關鍵字下錯，甚至有些文章標題沒有打出關鍵字，則就沒有辦法找到所有想要的結果。

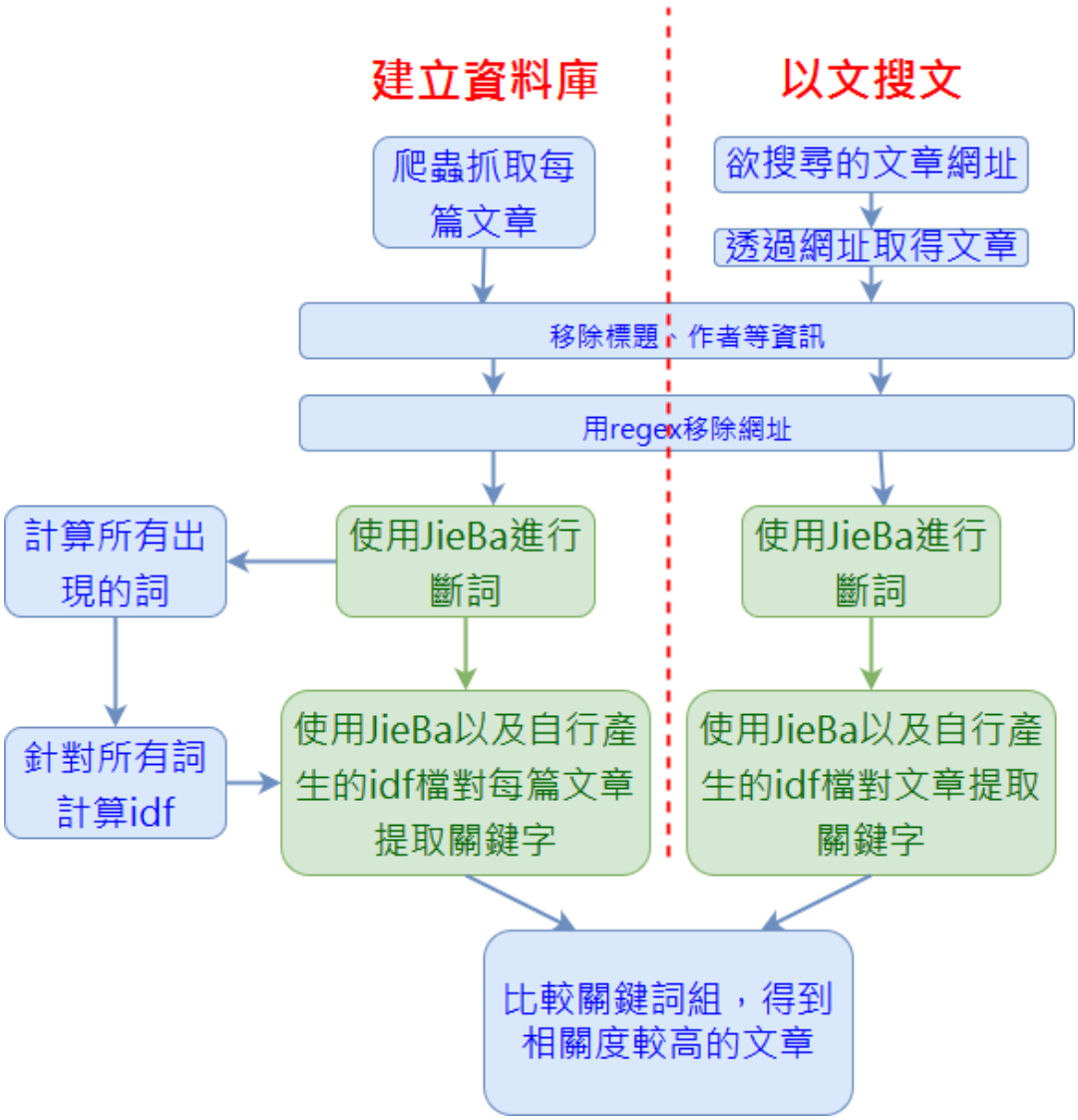
而因為課堂上有教到 article summary 的相關知識，讓我覺得可以往這個方向開發"以文搜文"的系統。然而就算是相同討論主題的文章也不太可能產生相同的 summary，因此我退而使用 tf-idf，想要找出每篇文章的關鍵詞，再透過比對關鍵詞的方式來找出相關主題的文章

三、 設計

首先利用爬蟲程式爬取論壇的文章並進行整理(此處使用 PTT NTU 版之文章)，接著使用 JieBa 套件對每個文章都進行斷詞，然後計算這些文章的 tf-idf。最後透過得到的 tf-idf 取得每個文章的關鍵詞。

進行檢索的時候也是把文章抓取下來後並斷詞，接著使用前面的 td-idf 檔案來取得關鍵詞，最後拿這些關鍵詞比較文章庫，取出比較相近的文章資訊。

流程圖如下：



四、 紀錄

爬蟲

<此部分因為跟 DSP 比較沒關係所以在報告中省略，有興趣可以參考 github 頁面>

文章整理

雖然每篇文章都會出現的字(如"標題"、"日期"字樣)因 tf-idf 可以忽略，但仍有一些每篇文章都不一樣的內容(如網址、作者 id)可能被誤判為關鍵詞，因此需要拿掉。

標題、作者 id 等因為出現地方固定，可以在爬蟲時就忽略掉，但因網址出現的地方不固定，而且作者也有可能在文章中留下網址(如貼圖連結，轉貼連結)，因此我利用以下 regex 來過濾掉文中所有網址

```
final = re.sub(r'https?:\/\/[\w.\/]*\B', '', final, flags=re.MULTILINE)
```

tf-idf

如果直接使用 JieBa 來計算 tf-idf 會出現很奇怪的結果，因此我們必須要自己來產生所有詞的 idf。

不過自己寫的 idf 會遇到一些問題，會有一些透明字元被辨認出來，但在 JieBa 載入 idf 時就會出錯，因此必須手動挑掉他，以下是我實驗時曾出現過的錯誤狀況。

(後來有進行預處理，但還是無法挑掉所有錯誤，我是去 jieba package 中的 tdidf.py 改寫一些 code 來幫助我把錯誤忽略 or 挑出來

)

1. 空白
2. \n (在文件中是完全空白的一行)
3. 有蠻多奇怪的空白 (在文件中只看的到機率)

關鍵詞比對

如果找出的關鍵詞組不錯的話，這部分就蠻容易的。

我的作法是將每篇文章的關鍵詞組與要檢索的關鍵詞組比對，看有幾個相同的關鍵詞，再輸出相同數量大於一個 threshold 的所有文章。

五、 成果

(本例子之文章庫為 PTT NTU 版從 1/15 往前爬約 19000 篇文章，下方有詳細解釋)

```
Rank : 50
Title : Re: [心情] 雨天騎腳車
Href : /bbs/NTU/M.1515490276.A.8E2.html
Date : 1/09

--------------------------------
Rank : 7
Title : [心情] 雨天騎腳車
Href : /bbs/NTU/M.1515407624.A.755.html
Date : 1/08

--------------------------------
Rank : 6
Title : [心情] 下雨天走路真的很危險
Href : /bbs/NTU/M.1493287738.A.8FA.html
Date : 4/27

--------------------------------
Rank : 6
Title : [校園] 請大家玩 Pokemon go 注意安全
Href : /bbs/NTU/M.1470487674.A.A6C.html
Date : 8/06

--------------------------------
Rank : 5
Title : [心情] 騎車不要滑手機很難嗎
Href : /bbs/NTU/M.1464762659.A.6F0.html
Date : 6/01
```

```
Rank : 50
Title : [校園] 說謊的體育室主任請下台負責
Href : /bbs/NTU/M.1514881852.A.AFE.html
Date : 1/02

--------------------------------
Rank : 9
Title : [校園] 田徑場租借事件學生會後續追蹤與說明
Href : /bbs/NTU/M.1508589938.A.C2D.html
Date : 10/21

--------------------------------
Rank : 8
Title : [新聞] 台大田徑場出借又出包？學生會批又因外
Href : /bbs/NTU/M.1514881550.A.4F8.html
Date : 1/02

--------------------------------
Rank : 7
Title : [新聞] 意識報快訊：學生會號召「堵」康正男
Href : /bbs/NTU/M.1514996552.A.CB6.html
Date : 1/04

--------------------------------
Rank : 6
Title : Re: [校園] 操場外借 校務建言回覆
Href : /bbs/NTU/M.1515138799.A.A97.html
Date : 1/05

--------------------------------
Rank : 5
Title : [校園] 放任場地被破壞體育室康正男出來面
Href : /bbs/NTU/M.1514951550.A.A95.html
Date : 1/03

--------------------------------
Rank : 5
Title : [校園] 田徑場事件跑道重獲田協認證
Href : /bbs/NTU/M.1512744202.A.85C.html
Date : 12/08
```

這個範例中要檢索的文章就是結果的第一篇，因為在大部分情況下我們用來搜尋的文章都會是文章庫的其中一篇。而繼續往下就可以看到這個系統強大的地方：找出的相關文章不會受到標題限制。左邊的範例是拿抱怨雨天騎腳踏車很危險的文章搜尋，它有找到另外一篇抱怨相同事情，但標題完全沒有提到腳踏車的文章(第三篇)。值得一提的是它還找到了抱怨騎車玩 pokemon go 的文章，也是討論騎腳踏車做出危險行為的事情。

右邊的範例則是拿最近的時事作為搜尋，除了相關的報導文章之外，它也找到了前一個與體育組有關事件的相關文章(第二篇以及最後一篇，可以用發表日期區別)。整體來說，我覺得表現比我預想的還高。

六、 討論與改進

在程式我直接使用 JieBa 的內建詞庫來斷詞，一開始在看斷出來的詞時覺得有一些些會斷在怪怪的地方，但如果要改進的話必須要使用自訂義詞庫，建立詞庫必須手動(我還沒想到自動的方式)而且曠日費時。而後來因為結果表現還不錯我就沒有打算要弄自訂義詞庫。

另外，JieBa 的斷詞有提供三種模式：全模式、精確模式以及搜尋模式。我使用的是預設的精確模式，全模式則是會將斷詞的所有可能都列出來(ex. 我走進台灣大學 -> 我 / 我走 / 走進 / 台灣 / 台灣大學 / 大學)，因為我認為斷出來的結果會使取關鍵字時取太多同義但長相不同的字，且這樣也會導致資料處理量變大，所以我沒有使用。至於搜尋模式是提取出類似給搜尋引擎用的(實際上就有點像提取關鍵字)，但因為這樣會破壞掉文章結構，也不知道它會把哪些東西刪去，因此我也沒有使用這個模式。

最後，目前關鍵字的比對方式是單純比較有多少個重複，並輸出重複個數大於一定量的文章，我其實覺得有一點點不嚴謹，如果想要得到更好的結果可能可以考慮其他方式，或是增加參考的東西(如標題)。

七、 使用方式

(本報告略過此部分，若教授或助教有興趣可以參考 github 頁面)

八、 Reference

JieBa ： https://github.com/fxsjy/jieba

一篇 tf-idf 的教學： https://stevenloria.com/tf-idf/

Tf-idf 的 wiki 頁面： https://www.wikiwand.com/zh-tw/Tf-idf

一篇正規表示式的教學：
https://atedev.wordpress.com/2007/11/23/%E6%AD%A3%E8%A6%8F%E8%A1%A8%E7%A4%BA%E5%BC%8F-regular-expression/