

東吳大學評價

組長：06170501 厲彥伯

組員：06170145 周欣德

組員：06170139 許靖玟

目錄

第一章：緒論

第一節：動機與目的

第二章：模型與分析步驟

第一節：模型介紹與操作

第二節：流程圖

第三節：分析步驟

第三章：實驗與分析結果

第一節：分析中的修改

第二節：結果呈現

第四章：結論

第五章：工作分配說明

第六章：參考資料

第一章：緒論

第一節：動機與目的

近年來大多數人對於未知的事物，多透過搜尋引擎取得相關資訊，當大家越關注網路評價，我們越應該注重網路上的聲量，因此我們想去探討東吳大學在網路的評價為何？同時將好、壞、中立之評價收納起來，優勢之處是值得推廣的地方，而不足之處進而讓學校得知並加以改善。

第二章：模型與分析步驟

第一節：模型介紹與操作

1. TextRank 模型介紹

TextRank 主要原理來是 Google 所發展的 PageRank，PageRank 主要是用來衡量網站之間的重要性，而 TextRank，類似於 TF-IDF 主要是用來找出文字權重與關鍵字，透過文章中高頻的字，計算相似度，將高頻的字高於門檻的單詞挑出來作為重點摘要。TextRank 有兩種方式計算，一種是無方向性(類似原本的 PageRank)，一種是有方向性。

2. TextRank 模型操作

- (1) 將爬蟲後的資料，整合成文本數據
- (2) 將每則留言（句子）切割成單詞
- (3) 將每個單詞以向量表示，併入矩陣中
- (4) 將相似矩陣轉換為以詞句為節點，用於詞句 TextRank 計算
- (5) 最後，排名在一定數量的構成最後的關鍵詞句

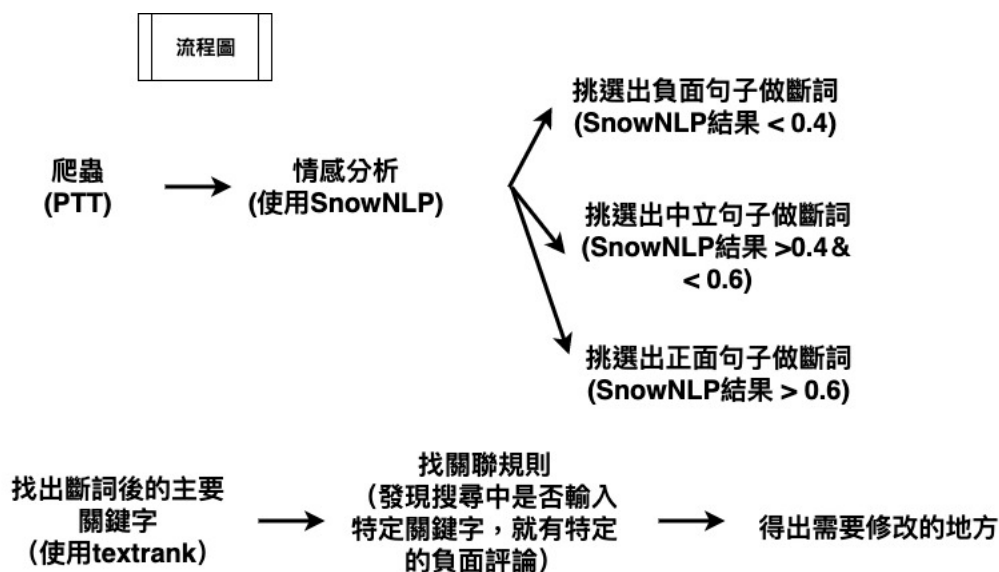
3. SnowNLP 介紹

SnowNLP 是基於 TextBlob 所開發的，為針對中文，專門處理自然語言的程式套件，內容包含中文斷詞、詞性標注、句子切割、情感分析、字體轉換等功能。

4. SnowNLP 實際運用

這次主要運用到句子切割及情感分析，以逗號或句號當作切割的依據，將每一句話進行情感分析，並將分析結果分為三類，分數 ≤ 0.4 、分數 ≥ 0.6 、分數介於 $0.4 \sim 0.6$ ，以利後續的操作。

第二節：流程圖



圖一：流程圖

第三節：分析步驟

(1)取得資料：首先我們透過爬蟲得到 ptt 上有關「東吳」的留言與標題，合計資料約 2500 筆資料。但是後來發現這些資料對於探討東吳大學負評的資料有限，所以改爬取「東吳負評」的留言與標題，資料合計約 3100 筆。

(2)資料清理：我們整理出我們需要的資料格式，同時也剔除不必要的內容。如東吳推廣部那些並不在本次探討的內容。

(3)斷詞：先將爬蟲後的結果進行斷詞，但有些較口語化的詞 jieba 內建並不認識，我們透過建立字典，再將內容重新進行斷詞。

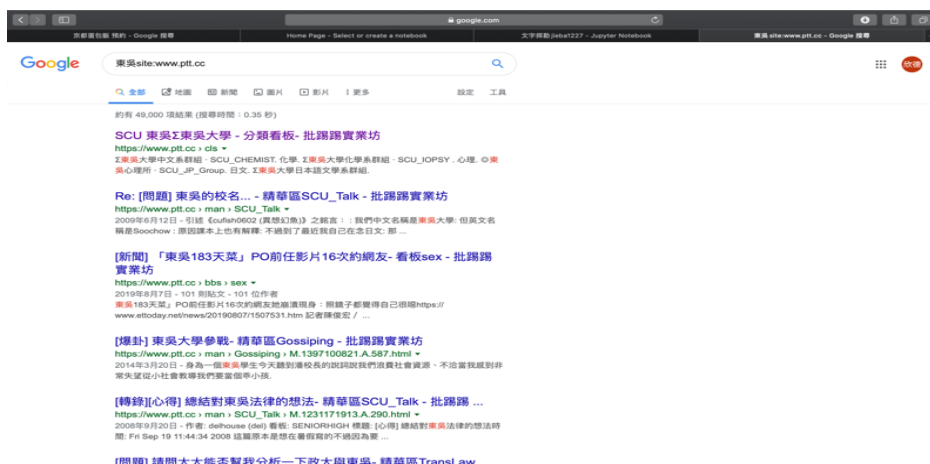
(4)分析情緒：透過 SnowNLP 將每則留言做情感分析，並萃取出評價結果 ≤ 0.4 的留言，當作是分析成分，並認定此區間的結果為負評。 $0.4 < \text{and} < 0.6$ 為中立， ≥ 0.6 為正面。

(5)應用 TextRank 算法生成關鍵詞句及摘要。

第三章：實驗與分析結果

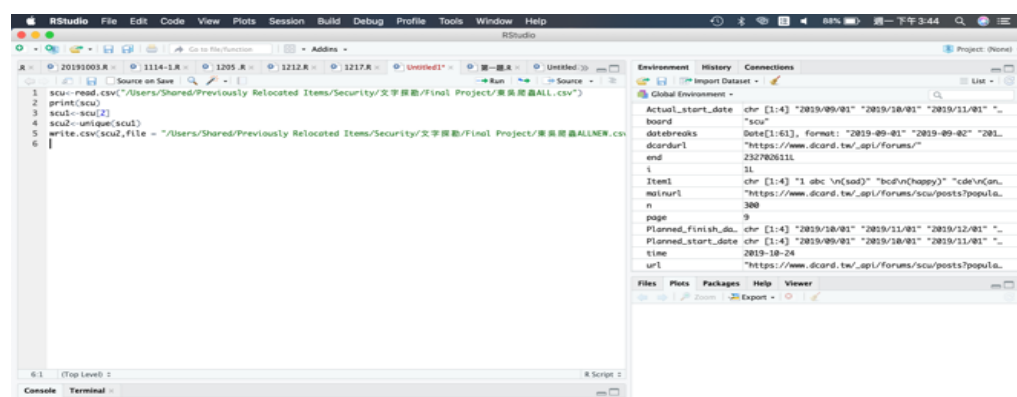
第一節：實驗過程

1.資料取得：爬蟲「東吳」、「東吳負評」的網頁 1~10 頁。

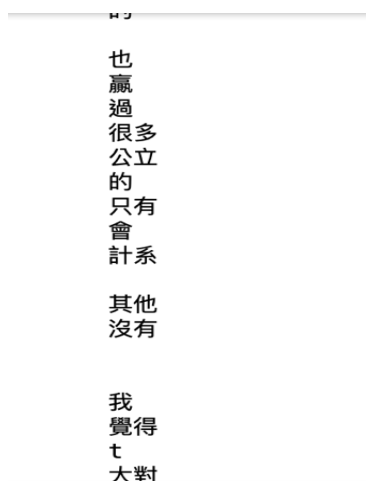


圖二：為爬取的網頁

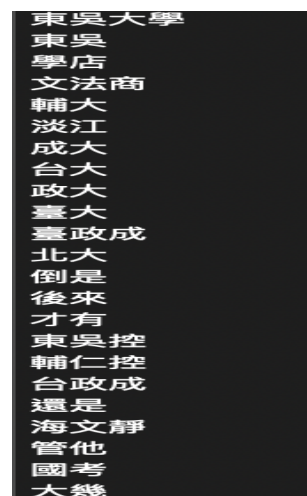
2.資料整理：這個步驟耗費較長的時間，我們整理有使用到 Excel、R、Python。首先爬取時就將同一使用者的留言視為同一則，才能避免同一筆資料被分成多筆資料。再者利用 R 去重複化。最後再透過 Excel 人工方式去除不要的資料



3.斷詞：我們第一次使用 **jieba** 分詞後，發現有許多詞語是內建沒有的，定義辭典後並重新斷詞，如輔大、政大等等。



圖四：第一次測試斷詞後，跑出來的結果。



圖五：建立的字典，約有 252 筆。

第二節：結果呈現

1.利用 **SnowNLP** 跑出的分數，我們將執行結果分別分成正面、中立、負面三個結果。結果如下。

| | 正面 | 中立 | 負面 | 合計 |
|-----|---------------|---------------|---------------|-------------|
| 數目 | 1489 | 595 | 1322 | 3406 |
| 百分比 | 43.71% | 17.47% | 38.81% | 100% |

2.利用 **textRank** 找出分別下列的結果：

(1)使用整份資料找出主要的關鍵詞，發現主要關鍵詞以東吳大學競爭學校如北大、淡江、政大居多。同時也出現東吳較知名的科系如會計、法律等等。

| 關鍵詞 | |
|-----|-----------------------|
| 法律 | 0.006235195212281076 |
| 北大 | 0.004608523269429693 |
| 學校 | 0.004607746829047792 |
| 老師 | 0.00416545031926811 |
| 淡江 | 0.004071564350777886 |
| 應該 | 0.0036129694005615654 |
| 學生 | 0.003541309182698474 |
| xd | 0.003532628052141913 |
| 大學 | 0.003257944002225973 |
| 法律系 | 0.0032497160058931537 |
| 覺得 | 0.0032120823675875596 |
| 沒有 | 0.003209967507739951 |
| 中正 | 0.0031997335157977593 |
| 不過 | 0.003137897288651214 |
| 會計 | 0.0029882884530784324 |
| 考上 | 0.002895690529859707 |
| 政大 | 0.002787860287978107 |
| 可能 | 0.0027524323528693986 |
| 畢業 | 0.0026868713941647416 |
| 好像 | 0.0026646589946336388 |

發現主要關鍵詞以東吳大學競爭

圖六：TextRank 對於整份資料所找出的關鍵字

| | |
|-----|-----------------------|
| 北大 | 0.005461466401958841 |
| xd | 0.005199054080944319 |
| 中正 | 0.004840997211544183 |
| 法律系 | 0.004772934589345601 |
| 大學 | 0.004446395044415421 |
| 老師 | 0.00439893427243637 |
| 淡江 | 0.004315649667047309 |
| 學校 | 0.0042701760704646225 |
| 現在 | 0.004152273181220675 |
| 考試 | 0.003953174635965803 |
| 畢業 | 0.0038663827451400905 |
| 學生 | 0.003667854038052862 |
| 考慮 | 0.003141000624815945 |
| 會計 | 0.0030641649421606306 |
| 知道 | 0.002957882560799138 |
| 政大 | 0.002893235470633703 |
| 資源 | 0.002874591030868017 |
| 好像 | 0.0027767994065333573 |

關鍵詞句：

北大法
來說
不會
法律不
後來
沒考上
念法律
中正法律
來看
看不
去念
不去
去看

摘要：

1271 0.0027872556606589943 看你也有選法律系 如果最後來念東吳 建議拼書卷轉法律,0.99867362

圖七：TextRank 對於正面留言所找出的關鍵字

(3) 使用資料中，中立的留言，找出主要關鍵詞，發現結果裡面的評論多了日文系。

| 關鍵詞 | |
|-----|-----------------------|
| 老師 | 0.006291377856896284 |
| 北大 | 0.0059725144518148945 |
| 會計 | 0.004158408019619078 |
| 法律系 | 0.004124948100844061 |
| 中正 | 0.0040832190573803434 |
| 東華 | 0.003975095002892952 |
| 出來 | 0.0038593324333030254 |
| 大學 | 0.0038082015943545384 |
| 成績 | 0.003538938949159709 |
| 好像 | 0.0033537241065899545 |
| 台大 | 0.00334725060198556 |
| 日文系 | 0.003277106720947076 |
| po | 0.0032239106065366143 |
| 現在 | 0.003175956943427779 |
| 台北 | 0.0030303308852883772 |
| 今年 | 0.003024245558572363 |
| 看到 | 0.0029141501610210384 |
| 覺得 | 0.0028697866782140063 |
| 補習 | 0.002868375812881566 |
| 法律 | 0.0028259424253719075 |

關鍵詞句：

摘要：

410 0.005561553259377565 北大推,0.440641503

圖八：TextRank 對於中立留言所找出的關鍵字

(4) 使用資料中，負面留言，找出主要關鍵詞，結果如下。

| | | |
|-------|-----------------------|--|
| 關鍵詞 | | |
| 淡江 | 0.00552606226760374 | |
| 學校 | 0.005387092814497222 | |
| 應該 | 0.005245977024336474 | |
| 不過 | 0.004553604085414182 | |
| 老師 | 0.00413569546966127 | |
| 沒有 | 0.004134966497277372 | |
| 覺得 | 0.004030087325180337 | |
| 北大 | 0.003814319153138295 | |
| 政大 | 0.0036376538441524916 | |
| 學生 | 0.003545039797421854 | |
| 可能 | 0.003488137061819308 | |
| 企管 | 0.0032526454940936343 | |
| 選擇 | 0.003197885234419353 | |
| 大學 | 0.0029285902217415947 | |
| 會計 | 0.0028468319477019167 | |
| 好像 | 0.002709405323368019 | |
| 經濟 | 0.0026286786509079 | |
| 台大 | 0.002626498485480473 | |
| 興趣 | 0.0025137200195579027 | |
| 法律 | 0.002406680427639925 | |
| 關鍵詞句： | | |
| 會後 | | |
| 會想 | | |
| 想說 | | |
| 來說 | | |
| 會推 | | |
| 話說 | | |
| 想去 | | |
| 不過大 | | |
| 後來 | | |
| 人推 | | |
| 會說 | | |
| 沒看 | | |
| 會去 | | |
| 應該會 | | |
| 會選 | | |
| 摘要： | | |
| 474 | 0.0022310622071603646 | 可能以後還會有學弟妹爬到這篇 我想說一下面試日文系的,0.039035289 |

這幾份主要關聯詞都是與我們競爭較大的學校如，臺北大學、政治大學、中正大學、淡江大學等等。除了學校以外，再來就是科系方面，外界總是認為東吳大學知名度集中於特定的科系，如法律系、會計系、日文系。

圖九：TextRank 對於負面留言所找出的關鍵字

第四章：結論

我們發現在 ptt 討論板上，東吳大學主要的聲量來自於幾個面向。首先，外界對於東吳大學的印象，過度集中於法律系、會計系、日文系這三個學系上，對於其他學系評論甚少。同時也出現很多認為東吳大學其他學系程度落差過大的負面聲音，會造成如此現象代表學校是否過度集中資源於特定科系。近一步觀察主要關聯詞也都是與我們競爭較大的學校如臺北大學、政治大學、中正大學、淡江大學等等，所以學生選擇學校時，會有一定程度參考網路上的內容。我們發現東吳的學生在業界普遍有一定程度的評價，讓大家對於學校的印象為一所前端私立大學。但仍有些地方值得學校去改進與探討，撇除先天的地理位置造成交通不便之外，其他方面有宿舍不足、學校的教育方式偏向補習班教學、選課不友善等等。這些批評的言論，依然在網路上出現，因而影響未來學生選擇東吳的意願。

第五章：工作分配

厲彥伯：爬蟲、ppt 製作

許靖玟：word-模型介紹、流程圖、SnowNLP

周欣德：資料整理、word-結論、實驗步驟、TextRank

第六章：參考資料

<https://medium.com/@danjtchen/textrank-文字探勘-找出關鍵字-以-八卦版標題為例-b16620370872>

<https://www.jiqizhixin.com/articles/2018-12-28-18>

<https://cloud.tencent.com/developer/article/1065715>

<https://codertw.com/程式語言/365456/>