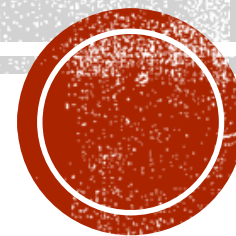


新聞留言篩選器 -以內容特徵值為基礎

指導教授：侯建良 博士

學生姓名：邱家堯 連鈞瀚 陳奕廷



研究動機

- 隨著網路世界的發達與普及，人們得以透過電腦和手機觀看第一手的新聞，同時也提供讀者一個自由的言論平台能夠發表意見與交流。但也因而產生下列幾個問題造成讀者在觀看新聞評論的困難以及不便：

- ✓ 越來越多的言論氾濫
- ✓ 無關新聞的留言
- ✓ 過度偏激的情緒性言論



因此我們希望藉由一套標準化的系統程式，將與新聞**相關程度**較為密切的留言篩選出來，以節省讀者閱讀留言的時間，並藉此提升留言與新聞間的相關性及可讀性。

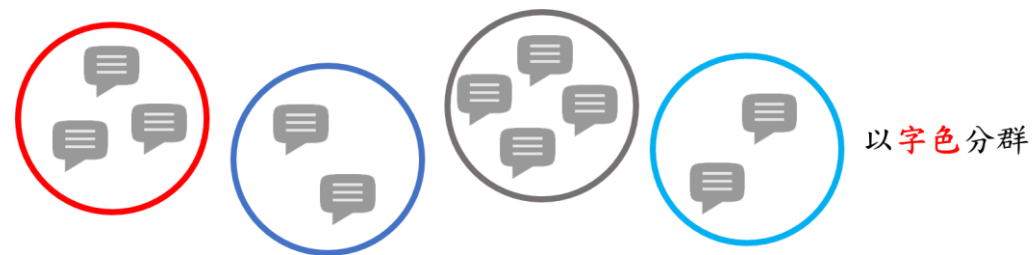


研究目的

- 利用 **文字探勘** 的技術，找出下列兩項數值：

- ✓ 新聞的內容特徵值

- ✓ 留言的內容特徵值



並透過兩者間的**相關程度分析**，篩選出與新聞主體相關程度較高的留言，並以**字體顏色**區分留言的相關程度，以期**節省讀者閱讀時間**，並提供其較為客觀而正確的言論。



研究方法



• jieba

以**jieba**將新聞與留言均進行分詞

• tf-idf

• TextRank

以**tf-idf**和**TextRank**計算分詞權重，並將兩權重相乘

• word2vec

以**word2vec**分別取得新聞與留言各詞的詞向量，取餘弦並乘上兩者權重的積，最後將整則留言中的各詞結果相加。

• K-means

用**K-means**將所有留言相關程度分群



系統開發



網路新聞留言篩選器

輸入新聞網頁網址

分析

結果

1

匯入新聞網址

台灣運毒什麼時候可以死刑? [0.0302380953207]

3

系統呈現留言與
新聞之相關分數

新聞留言

賺賺賺 [0.0]

人權有很多方面的，不要狹隘的理解它。台灣從過去到現在犧牲了多少人才有現在的環境，人權在很多方面還是沒有做得很好，還需要很多的努力，譬如說最近很熱的「賦稅人權」。珍惜我們得來不易的環境，並努力改善不足的，給下一代更好的生活環境是這一代該努力的。 [0.0199565319284]

3好+1好=系賀 [0.0]

看到印尼總統愛人民的心有多堅決。嚴刑峻罰還是具有一定程度的喝止作用。台灣執政者要好好思考，犯罪者老是抱持僥倖心態，吸毒造成的社會問題有多麼大？ [0.033647012004]

各人造業各人擔~~不要牽拖任何理由，毒品罪刑的嚴重性別告訴我你不知道~~被槍斃是剛好而已 [0.0262653964868]

台灣運毒什麼時候可以死刑？ [0.0302380953207]

應該查查通訊 到底是誰指使他們運毒的組織👉 [0.0231293738499]

唉.....下輩子別再犯錯了..... [0.00974938222468]

台灣應該要跟他們一樣 [0.026382187527]

一條命15萬虧大了 [0.0108829332059]

=..=!!!如果能回來~繼續運毒嗎~~~~~ [0.0255764445009]

這...這...高才生是塔殼ㄉㄟ賽嗎???!! [0.0]

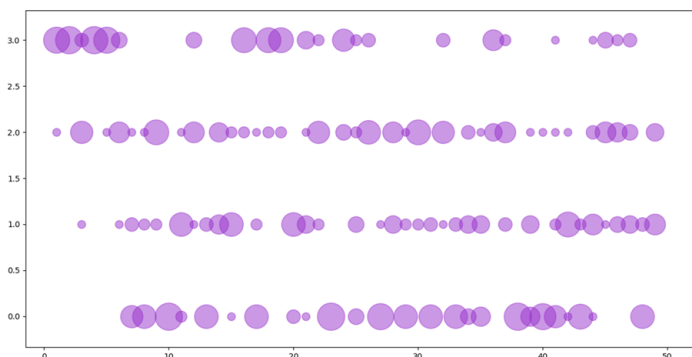
取得對應該則
新聞之所有留
言並獲得新聞
留言分群結果
(以字色呈現)

2

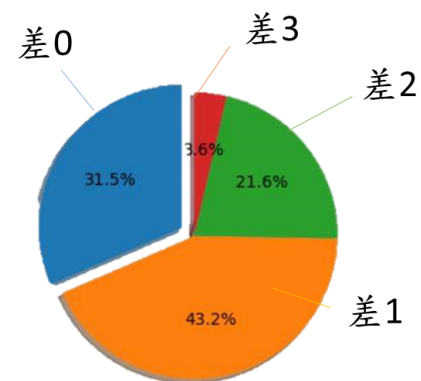


結果分析

經由樣本的蒐集及留言分則比對，可以發現各則留言的評分幾近集中，由此可證明出每則留言**評分有相似看法**



樣本的評分集中程度圖

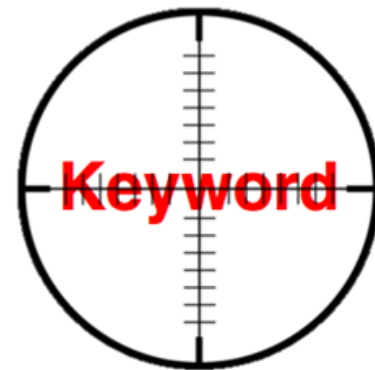


系統與主觀績效評估比較圖

根據系統程式跑出之結果與主觀績效評估比較之下，可以發現大約七成五的樣本誤差值均在1評分水準以內，代表**系統與人為的判斷出入甚小**



結論與未來展望



尋找新聞與留言的特徵值 → 篩選出與新聞相關程度較高留言

主觀績效評估+多組樣本+系統結果 → 符合一般大眾判讀結果

字色分群 → 方便判讀及節省閱讀時間

成本最小化與利益最大化

■ 未來展望

對於某些有閱讀部落客文章、網路新聞或是對於閱讀長篇文章略覺麻煩的讀者，我們希望能發展出一套系統，讓我們可以判斷出該文章之摘要，以減少讀者閱讀的時間。

