工 業 工 程 專 題

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

一、研究動機與目的

研究動機

隨著**網路世界**的發達與普及，人們得以透過電腦或手機等3C用品觀看第一手的新聞，不僅增進了讀者觀看新聞的即時性與便利性，同時也提供讀者一個**自由的言論平台**能夠**在各則新聞底下發表自身的觀點**並**與其他讀者交流**。但也因為這些言論平台的設立，導致下列幾個問題的產生而造成讀者在觀看新聞評論的困難以及不便：

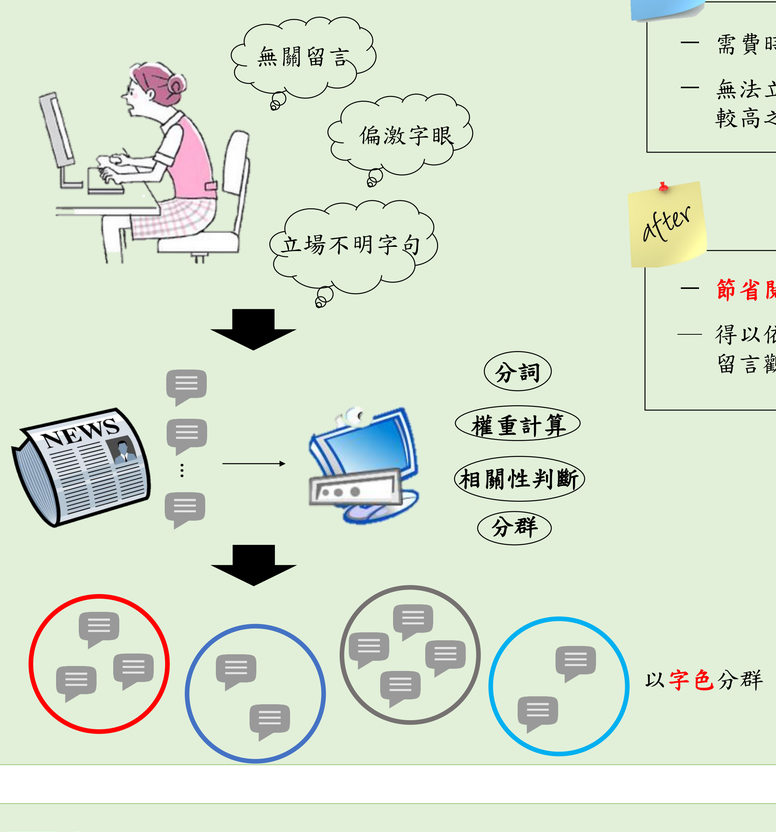
一越來越多的言論氾濫
一無關新聞的留言
一過度偏激的情緒性言論

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

本研究利用**文字探勘**的技術，找出新聞的內容特徵值，並以相同之方法找出留言的內容特徵值，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。

二、問題的發現與解決

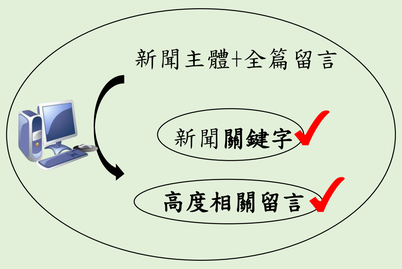


before

- 需費時逐則查看留言
- 無法立即獲得與新聞主體相關程度較高之資訊

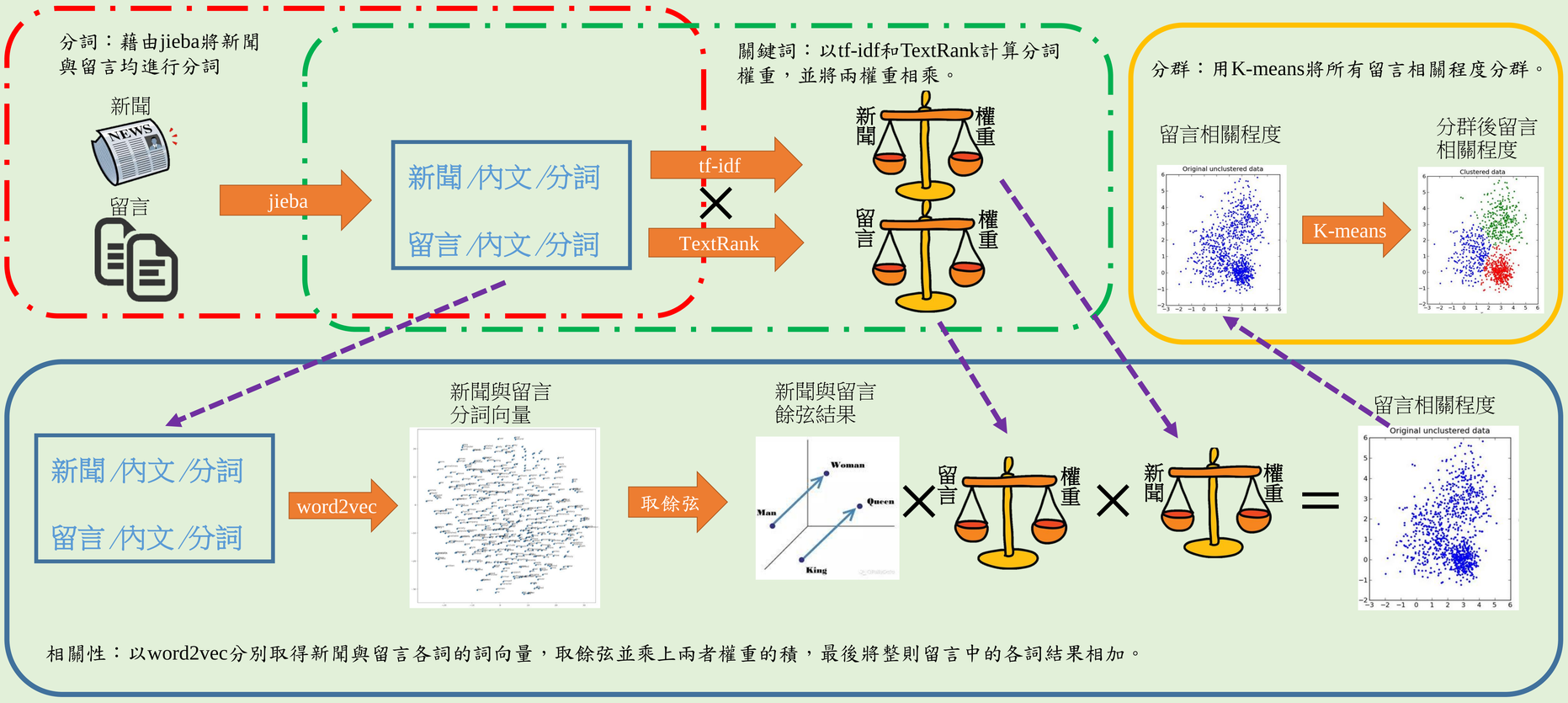
after

- **節省閱讀時間**
- 得以依照系統判別之相關性自由選擇留言觀看



三、研究方法

- 分詞 • jieba
- 關鍵詞 • tf-idf • TextRank
- 相關性 • word2vec
- 分群 • K-means



四、系統開發

工具：python JS HTML CSS

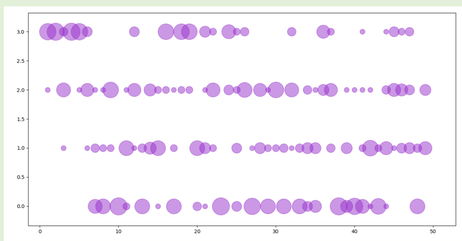
匯入新聞網址

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）

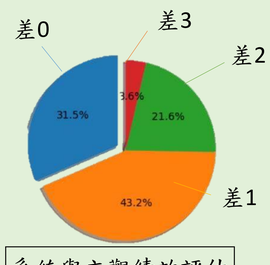
系統呈現留言與新聞之**相關分數**

五、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**

六、結論與未來展望

Future Goals

結論

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

藉由留言與新聞相關性的篩選系統，使得我們可以更加快速的篩選出對於瞭解新聞較為相關抑或是較為重要的留言，同時也能判讀出新聞的關鍵字。我們希望未來能將反諷性留言納入考量，並將系統的正確率再提升；而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。

106學年度 工業工程專題
清華大學工業工程與工程管理學系