Data Mining HW2
Topic Detection by Frequent Pattern Based Keyword Clustering

Team 5

陳世運 R06725024、陸艾眉 R06725030、劉庭維 R06725031、陳信豪 R06725048

一、作業概述

在 Text Mining 研究中，有許多將詞彙轉換成向量空間的方式，如 word2vec 與 GloVe，而透過比對向量之間的相似度，可以衡量詞與詞之間的關係。以此為基礎，在本次 Data Mining HW2 中，運用了關聯分析，找出詞彙在文章中的 Frequent Pattern，接著以 support 值建立詞彙之間的 adjacency matrix，而將其標準化後，抽出每列的值即形成我們所定義的詞向量。最後，藉由 K-Means 分群演算法，將詞彙進行分群來探討和挖掘主題，並以 t-SNE 將詞向量降成二維後於平面畫出散佈圖，以此作為資料視覺化的成果展現。

二、資料集

NTU Course - Big Data Analysis 2018 （楊立偉教授）– HW1 Dataset
2016/01/01 ~ 2016/11/30 之新聞
Link:
https://drive.google.com/open?id=1mMDG3wEWwMuNjBbTAy-RZFiqmyUr4CF2

三、執行環境和套件

- 處理環境

- 程式語言和套件

Languages/Toolkits	Version
Python	3.6.4|Ananconda
openpyxl	2.4.10
jieba	0.39
pyfpgrowth	1.0
scikit-learn	0.20.dev0
matplotlib	2.1.2
adjustText	0.7

四、資料預處理

- 步驟一：

  原始資料中有 90508 筆新聞資料，但裡面有許多是天氣或股票的新聞，為了避免影響結果，先將其濾除

- 步驟二：

  先給予特定的幾個主題字眼，文章（標題 + 內文）若含有任一主題字眼便加入訓練文本，在這邊假設我們選的主題字眼為 ['鴻海', 'Apple', '選舉']。

- 步驟三：

  將所有訓練文本進行斷詞（以下為斷詞細節流程）

  (i) 先以 n grams 的方式對每篇文章做切詞，n 取 2~7，所以一篇文章會被切七次，依序切兩、三、四、五、六、七個字。

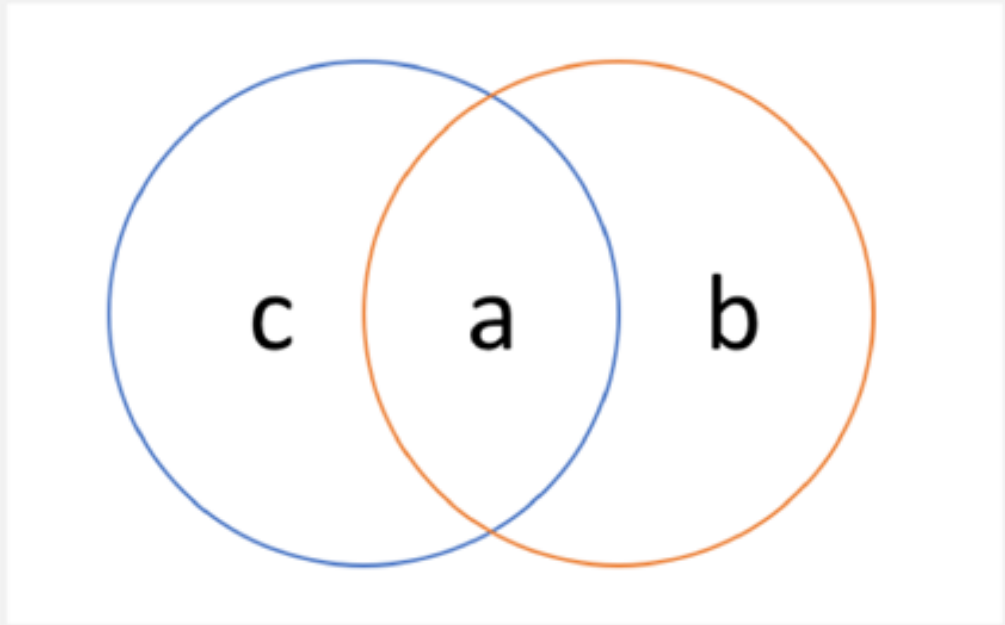
  Example:「今天天氣真好」
  2-gram：「今天」「天天」「天氣」「氣真」「真好」
  3-gram：「今天天」「天天氣」「天氣真」「氣真好」
  4-gram：「今天天氣」「天天氣真」「天氣真好」
  ……以此類推

  (ii) 統計所有詞彙 （term）於訓練文本出現出現的次數（TF）

  (iii) 濾掉 TF 次數不多的 term

  (iv) 再透過 MI/SEF (Significance Estimation Function) 濾詞，此方法為計算 term 和其 max length sub-term 的重疊次數，來決定其 max length sub-term 是否應該被濾除。

Example: Term =「行動支付」- 其 TF 為 a
Sub-Term1 = 「行動支」 - 其 TF 為 b
Sub-Term2 = 「動支付」 - 其 TF 為 c
而這三者其 TF 以圖可展示為



接著將以以下的值濾掉：

- a/(c+b-a) 過高的 Sub-Term1 與 Sub-Term2 濾除
- a/b 過高的 Sub-Term1 （行動支）濾除
- a/c 過高的 Sub-Term2 （動支付）濾除

(v) 長度為 7 的 term 通通濾掉，原因是預設切到 7-gram，而 SEF 沒有 8-gram 就無法濾除不完整的 7-gram。

(vi) 將剩下的 term 和其對應的 TF 值加入 jieba 自定義辭典，並將訓練文透過 jieba 斷詞，並濾除一些不重要的字眼（filter_words.txt），讓 jieba 切詞更為精準。

五、關聯規則

關聯規則常被用於購物籃分析，而在 Text Mining 中，可以將每篇訓練文本視為一個購物籃，而其中的每一個詞彙可以視為一個商品。以此我們便可以運用關聯規則的演算法（Apriori 或 FP-Growth）找出 term 的 Frequent Pattern。

我們運用了 Python 的 pyfpgrowth 套件來進行關聯分析。並使用 support_min = 50，conf_min = 60% 的部分結果。

(Frequent Pattern 共有 11079 條，Rule 共有 2956 條)

- **Frequent Pattern 與其支持度高低(取部分結果)**

Frequent Pattern	Support
('夏普', '日本', '董事長', '郭台銘', '鴻海', '鴻海與夏普')	53
('INCJ', '夏普', '產業革新機構', '郭台銘', '鴻海')	51
('Donald', 'Trump', '共和黨總統', '川普')	58
('候選人', '川普', '希拉蕊', '柯林頓')	90
('三星', '面板', '鴻海')	50
('Apple', 'Mac', 'iPhone')	67
('Apple', 'Pay', '信用卡')	65
…	…

根據我們從主題字詞「鴻海」、「選舉」、「Apple」，歸納以下結果：

- 產業動態

一般聯想到和鴻海相關的字詞主要受到他的產業影響，像是富士康、電子產品、代工、精密儀器等。而我們在 2016 年的財經新聞文章中，找到和鴻海相關且出現最多相關字詞是和日本夏普之間的關聯。所以，我們可以追溯回當時發生的重大事件，知道鴻海在 2016 年 3 月 30 日正式收購日本夏普集團，而在收購的過程中，媒體對於這段鴻夏戀也有各種不同面向的討論，像是鴻海想要殺價收購、鴻夏戀是否會帶來營利、協商過程、債務的移轉等。也因此，我們的分析可以看出，大部分相關的新聞都朝這一走勢，所以 frequent pattern 的結果可以用這個事件的推展來驗證，甚至，我們也可以透過關聯指出。

另外，我們也發現”鴻海”與”郭台銘”字詞在鴻海新聞當中出現的比例也相當高，因為郭台銘為鴻海董事長其經常對鴻海企業提出相關看法或是一些企業形象的塑造、相關社會公益等等，都是鴻海企業的重要特徵及重要決策，所以”郭台銘”字詞在財經新聞中出現的頻率高(frequent pattern 的結果)也表示著其與鴻海

企業的高度關聯性，鴻海董事長的一舉一動都與鴻海集團密不可分，因此也成為鴻海新聞的重要關鍵字。

- 年度大事件

2016 年正值美國總統大選，呼聲最高的候選人分別是共和黨的川普和民主黨的希拉蕊。雖然媒體為川普做的民調每次都是處於劣勢，但是川普的名人效應持續發酵甚至超過對手的政壇經驗帶來的關注，而導致國際媒體即使不看好川普，卻仍然高度關注他的一舉一動，並隨時把川普說出口的狂言誑語和政策做成新聞發布，新聞中也會將川普和他的競爭對手希拉蕊·柯林頓做各種比較。

因此我們可以發現在財經新聞當中'候選人', '川普', '希拉蕊', '柯林頓'同時出現的頻率次數有 90 次，可見在競選時期，支持率高的候選人會被媒體拿來做比較跟文章，候選人的一舉一動在競選期間都會受到相當大的關注，他們的一舉一動都會影響到後面的選情，也是競選期間選民相當關注的新聞，因此我們可以透過關聯規則，找出財經新聞中相關字詞出現的頻率高低並加以對事件做預測及分析。

- 企業產品

在智慧裝置方面，Apple 一直是手機產業的龍頭老大，每一個發表會的舉行都會有吸睛的產品問世，這些產品的業績和銷售好壞也一直是媒體關注的重點，因此，同樣是 Apple 的品牌、手機、筆電等關鍵字的相關度都會被我們篩選出來。而這類型的關聯讓我們覺得比較屬於日常的資料，沒那麼新奇。

我們可以發現搜尋出的結果 Apple 主要就為其所推出的主要產品 Mac、iPhone 都是財經新聞中與 Apple 企業相關的報導重點。另外還有'ApplePay'與'信用卡'之間的關聯，因為 Apple pay 目前主要還是以綁定信用卡為主，所以也可以找出兩者之間的關聯。

透過 Frequent Pattern 可以抓出哪些字詞相關程度的高低，並了解該主題下那些字詞與該主題具有較高的相關性並且一起在財經新聞中出現的機率高低，進而對企業可以推廣出該主題新聞的關鍵字為何並且引起媒體興趣的可能主題有哪些，像是從上述就可以發現在美國總統大選期間，我們發現'候選人', '川普', '希拉蕊', '柯林頓'同時出現的頻率次數有 90 次，可見這是媒體在選舉期間喜愛取材的新聞，這樣的新聞也許可以提高民眾的注意力與關注度所以引起媒體較高頻率的報導候選人之間的比較及競爭情況，對於新聞資料分析有很大的幫助。

- **Association Rule 與信心度(取部分結果)**

Rule -X	Rule-Y	Confidence
('投資', '虧損', '鴻海')	'夏普'	0.89
('集團副總裁', '鴻海')	'戴正吳'	0.76
('民主進步黨', '民進黨', '總統')	'蔡英文'	0.76
('副院長', '國民黨', '立法院')	'洪秀柱'	0.83
('Plus', 'iPhone')	'蘋果'	0.85
('庫克', '蘋果')	'執行長'	0.93
('候選人', '川普', '柯林頓')	'希拉蕊'	0.90
…	…	…

在 rule 的部分，將關聯分析應用在文字分析中，我們可以看到在 2016 年的各項 mining 出的關鍵字是由什麼原因對應到的，或是企業(組織)負責人是誰的關聯。像是我們從「投資, 虧損, 鴻海」對應到夏普，可以代表在鴻海收購夏普的事件中，普遍媒體都是認為鴻海這項投資是不划算的，可能會因為夏普自己的經營不善連帶造成鴻海的公司成長受到影響。另外像是「民主進步黨,民進黨, 總統」對應到的是蔡英文，也可以說明在當年民進黨推出的總統候選人是蔡英文;「副院長,國民黨,立法院」對應到洪秀柱，則是代表當前國民黨立法院副院長是洪秀柱。以及「候選人,川普,柯林頓」對應到希拉蕊，則是代表希拉蕊對候選人：川普及柯林頓的重要性，可看出希拉蕊其對兩者候選人的影響力。

因此，我們認為透過關聯分析可以找出哪些字詞之間有高度關聯性並找到關聯式法則，甚麼樣的字詞會產生不同的規則以找出彼此的關聯性，確實可以發掘當前事件的原因和即時的負責人對應，這樣的分析甚至可以讓我們找到更加特別的對應，並且助於資料分析及整理，了解資料的特性關聯進而對目標客群做不同的推薦通知，增加該新聞的點閱率等等，都是可以透過關聯式規則來完成。

六、Term Adjacency Matrix and Term Vector

- 步驟一：

  將 Frequent Pattern 依長度做排序，並保留長度大於 4（或自訂）的 FP 來建立接下來的 Adjacency Matrix。

- 步驟二：

  Adjacency Matrix 在此是根據 term 來建立的，一個 term 對應到一列和一行，而 Adjacency Matrix 中的每個值及代表某 term1 與某 term2 的關聯。

	Term1	Term2	Term3	…
Term1	A11	A12	A13	…
Term2	A21	A22	A23	…
Term3	A31	A32	A33	…
…	…	…	…	…

- 步驟三：

  Adjacency Matrix 中的值計算方式如下

  (i) 先將所有 Aij 設為 0

  (ii) 針對一個 Frequent Pattern，和其 support 值 s，若 Aij 對應其中的任兩個 term 的組合，便將其值加上 s。

  Example:

  有一 FP 為 <term1,term2, term3, term4>，其 support 值為 60 則。

  A(term1, term1) += 60

  A(term1, term2) += 60

  A(term1, term3) += 60

  A(term1, term4) += 60

  A(term2, term1) += 60

  A(term2, term2) += 60

  …… 以此類推

  (有 4 * 4 = 16 個 A 中的值要加上 60)

  (iii) 掃過所有長度大於 4 （或自訂） 的 FP，並對每個 FP 進行 (ii)

  (iv) 最後進行標準化，將 A 中的每一列除以 row sum

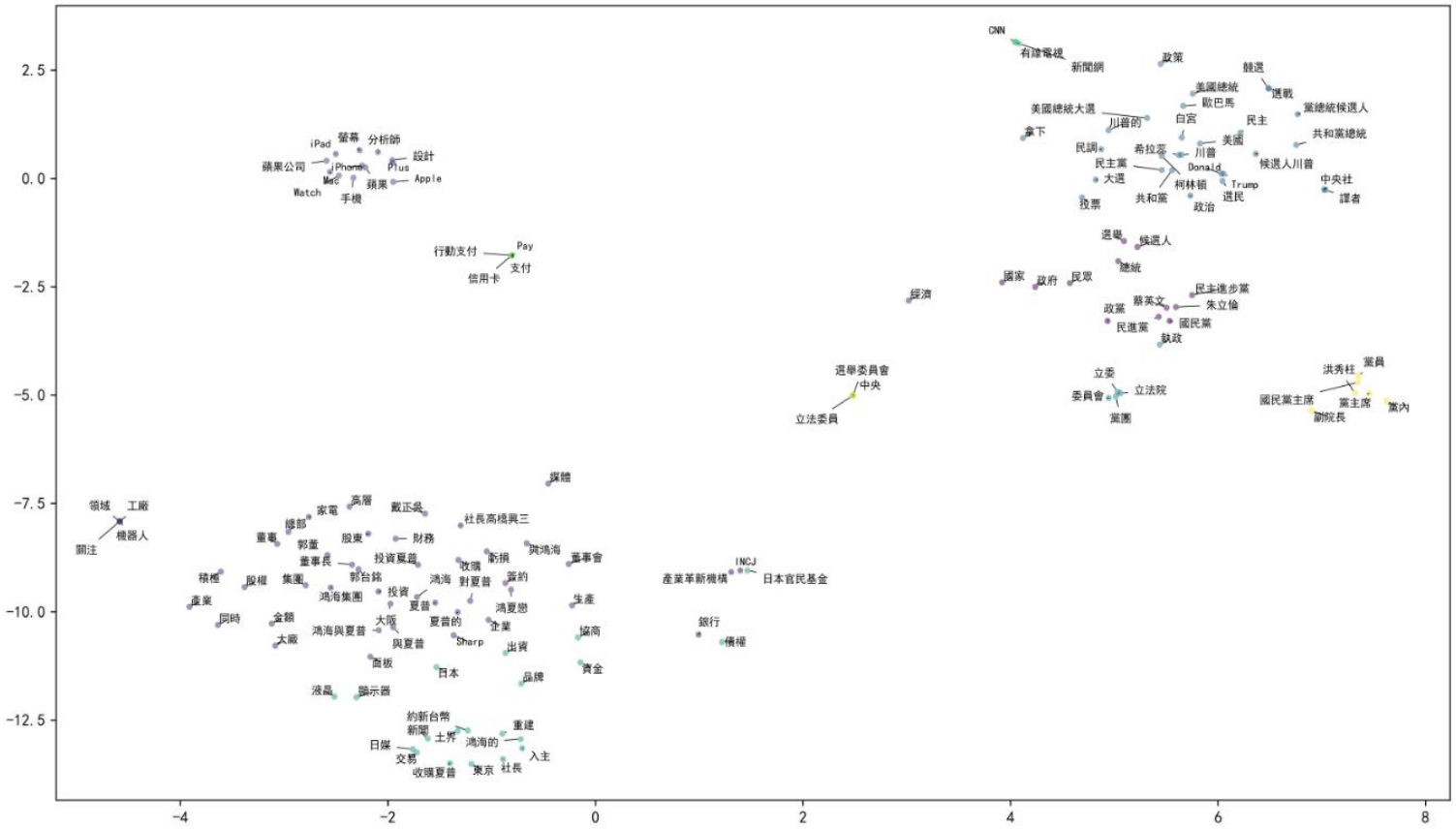
  A[i, :] = A[i, :]/sum(A[i, :])

  (v) 每一列即為其對應到的 term 的 term vector

七、Keyword Clustering and Topic Detection

形成 Adjacency Matrix 的 term 即為我們所找出的 keyword，接著我們將 term vector 透過 K-Means 進行分群，並以 t-SNE 將詞向量降至二維後進行 scatter plot，其最後結果如下，分為十群。

（圖中點的顏色代表 K-Means 分群之結果，在此 K 取 10）



從上圖中，可以明顯觀察出幾個 Topic，如「鴻夏戀」、「美國大選」、「Apple」、「行動支付」、「台灣政治」 …… 等等，並也因此證明了我們運用 Frequent Pattern 所建立的詞向量的有效性。

八、結論

關聯規則是一個常見的資料分析工具，經常用於購物籃或推薦系統，然而其於 Text Mining 中的應用卻不常被提及。以此為契機，我們嘗試將關聯規則與 Text Mining 做結合，以文本與詞彙當做購物籃問題中的籃子和商品，來進行關聯的分析，進而可以得到當年度熱門新聞或是事件發生的事件原因，更可以快速建立關鍵字之間的關聯性。此外，我們運用找出來的 Frequent Pattern，建立了詞彙的 Adjacency Matrix，並從中得到詞向量。而在最後經由分群和和維降作圖的結果，可以顯示以我們的方法建立的詞向量是有意義的，並且可以依此找出詞的主題性，從而達到 Topic Detection 的效用。