Data Mining HW2
Topic Detection by Frequent Pattern Based Keyword Clustering

(Ryan Report)

[Abstract]

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。我們以此為基礎，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，抽出每一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。

[Dataset]

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

[Languages and Toolkits]

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

[Environment]

[Preprocessing]

1. 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

2. 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']。

3. 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。

   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」

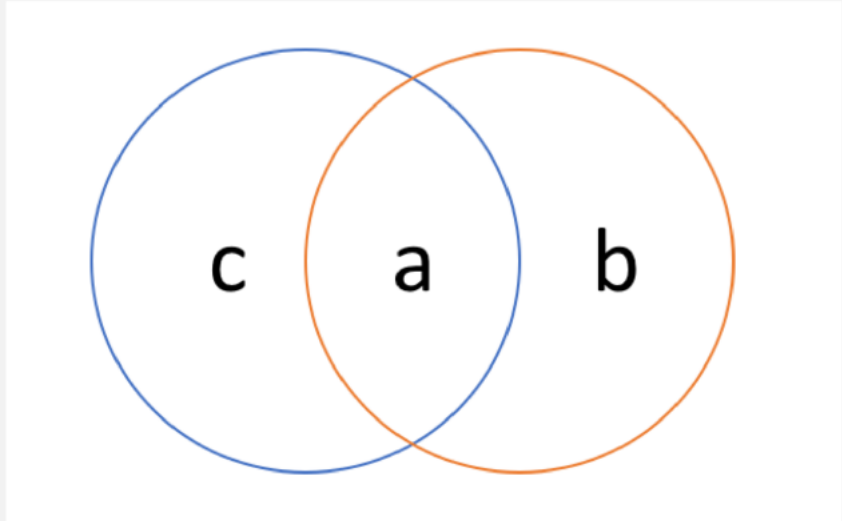
   ……

   (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
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
（因為預設是切到 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，並將訓練文本透過 jieba 斷詞，並濾除一些不重要的字眼（filter_words.txt）
（前面做的都是為了讓 jieba 切詞更為精準）

[Association]

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。

在此，我們運用了 Python 的 pyfpgrowth 套件來進行關聯分析。
以下為 support_min = 50，conf_min = 60% 的部分結果
(Frequent Pattern 共有 11079 條，Rule 共有 2956 條)

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
…	…

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
…	…	…

[Term Adjacency Matrix and Term Vector]

1. 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

2. 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	…
…	…	…	…	…

3. 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

   ……

   (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)

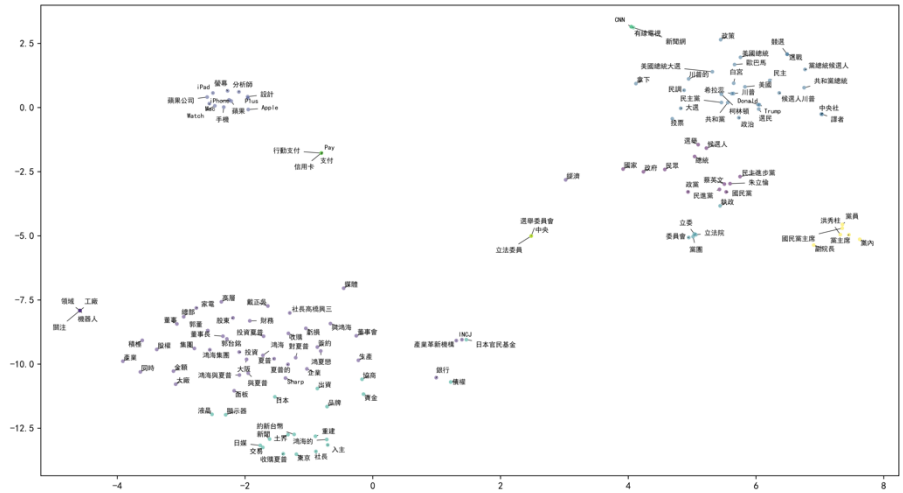
   (iv) 最後進行標準化，將 A 中的每一列除以 row sum

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

   (v) 每一列即為其對應到的 term 的 term vector

[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。



（圖中點的顏色代表 K-Means 分群之結果，在此 K 取 10）
（建議直接打開 draw.png 看大圖）

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。

[Conclusion]

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。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。