1. # 2020 Fall DSP Final Projcet Latent Semantic Analysis (LSA)

形式：程式實作 + 結果報告

https://github.com/vincent10400094/DSP_Final_Project

組員

b07902016 林義閔 b07902114 陳柏衡

1. 動機

如果要做文章分類，或是統計語言模型特性時，之前我們學過的都是一些以機率或是神經網路為基底的模型，而且通常是用「labeled data」來訓練的。因此在第十四章第一次聽到 lsa 的觀念時，我們覺得使用線性代數的嘗試很有意思，決定自己動手操作看看，觀察其背後的運作的原理以及將其應用在一套中文新聞資料集中的結果。

2. 實驗方法

我們參考第十四章的內容 [1] 以及原始論文 [2] 提供的方法實作 LSA，並將其使用於一套中文新聞的語料上，觀察將「字詞」、「新聞」各自投影在一個較低維空間的結果，並嘗試使用投影出來的向量找出「字詞間」、「新聞間」的關係。

2-1. 方法、名詞解釋

- $V = \{w_1, w_2, \dots w_M\}$ 為總共有 M 個字詞的集合（在我們的實驗中，$M = 20578$）
- $T = \{d_1, d_2, \dots d_N\}$ 為總共有 N 則新聞的集合（在我們的實驗中，$N = 3266$）
- W 為 $M \times N$ 矩陣，第 i 個 row 代表 w_i 而第 j 個 column 代表 d_j。
- t_i 第 i 個詞在所有語料 T 中出現的次數

2-2. 資料處理

我們使用一套在網路上找到的「台華新聞語料庫」[3] 作為我們的資料，其中包含約 3266 則中文新聞，且資料是已經事先分詞好的。

```
Obama 大勝 美國 首位 黑人 總統 駐美 特派員 曹郁芬 華府 報導
```

拿到資料後我們的處理分成以下兩個部分：

1. 統計
    1. 語料 T 中所有詞的數量（M）
    2. 每個詞在 T 中總共出現的次數（t_i）
2. 為所有的詞建立 index，也就是指定每個詞對應到 W 的哪一個 row（我們只取總出現次數 ≥ 3 的詞，否則詞的總數會太多）。

處理後我們得到一個表格，之後用於建立 W 矩陣。

Index	詞	總出現次數
0	總統	63
1	美國	217
2	首位	8
3	黑人	4
4	歐巴瑪	13
5	的	18000
...		

2-3. 建立、分解 W 矩陣

1. 計算代表所有語料的矩陣 W，每個 entry 的計算方法如下：

$$w_{i,j} = (1 - \epsilon_i)\frac{c_{i,j}}{n_j}$$

其中 * $c_{i,j}$ 為第 i 個詞在第 j 則新聞中的出現次數 * n_j 為第 j 則新聞的長度 * ϵ_i 為第 i 個詞於所有新聞中的 entropy，越高代表其越有指標性，越低代表越多的文章都包含這個詞（在 N 個項目的分佈中，entropy 最高為 $\log N$，最小為 0，因此最前方乘上一個 normalize 項使 $0 \leq \epsilon_i \leq 1$）

$$\epsilon_i = -\frac{1}{\log N}\Sigma_{j=1}^{N}\frac{c_{i,j}}{t_i}\log\frac{c_{i,j}}{t_i}$$

2. 接者我們使用 SVD (singular value decomposition) [4] 的方法將 W 拆成三個矩陣（已經照特徵值大小排序好）：

$$W_{M\times N} = \underline{U}_{M\times M} \times \underline{S}_{M\times N} \times \underline{V}^T_{N\times N}$$

並取出前 R 高的特徵值及特徵矩陣（我們取 $R = 150$），等於是取 $\underline{U}$、$\underline{V}$ 的前 R 個 column（特徵向量）分別作為 U、V，而取 $\underline{S}_{R\times R}$ 的前 R 個特徵值作為 $S_{R\times R}$。

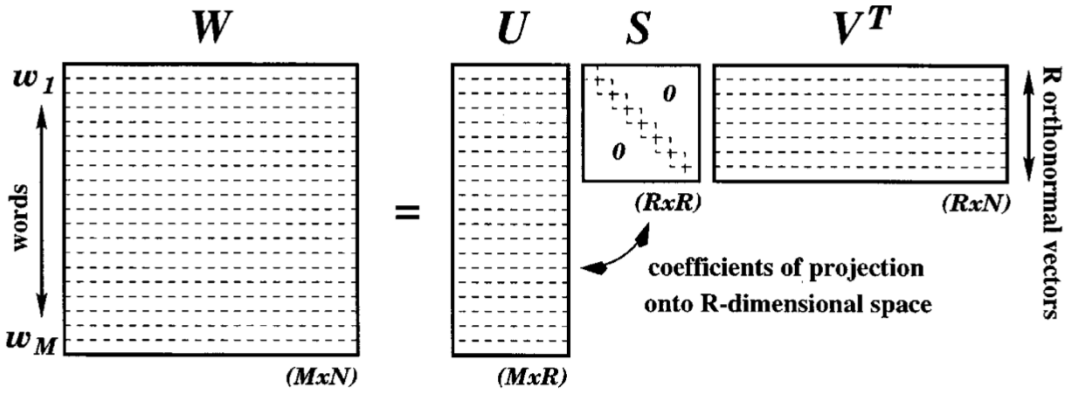
$$W_{M\times N} \approx \hat{W} = U_{M\times R} \times S_{R\times R} \times V^T_{N\times R}$$

U、V 就分別是對應 R 個特徵向量的 R 個 basic vector，構成一個 R 維的子空間。

2-3. 字詞分群、新聞分群

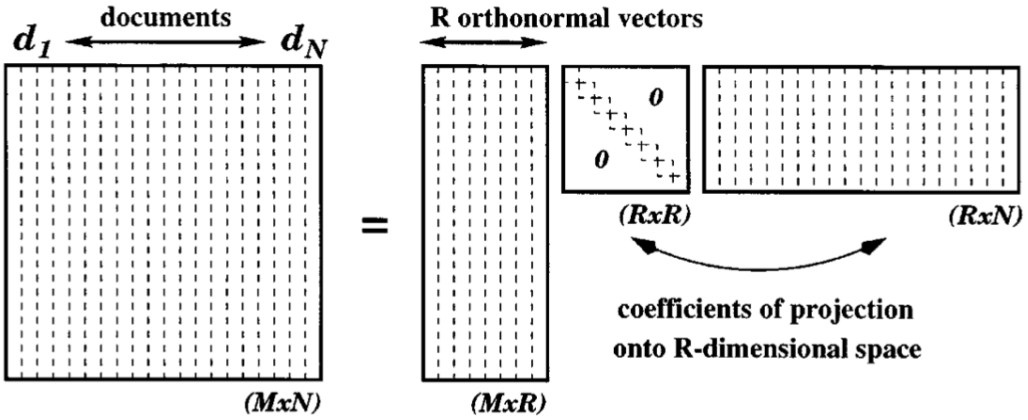
接著就是最重要的部分：將詞、新聞投影至一個維度較小的空間（被稱為語意空間），在語意空間中字詞、新聞將會以取出來最重要的 R 個特徵所表示。

2-3-1. 字詞分群



由這張論文中的圖片說明，$U \times S$ 即為將 M 個詞投影在 R 維子空間的結果。

2-3-2. 新聞分群



而 $S \times V^T$ 則是將 N 則新聞投影在 R 維子空間的結果。

2-3-3. K-means cluster

透過 K-means [5] 去對字詞（或新聞）在 R 子維空間投影的向量去做分類，將字詞分為 100 類，新聞分為 50 類。

3. 實驗結果

3-1. Word entropy

在建立矩陣 W 的同時還會計算每個詞 w_i 在整個語料 T 中分佈的 entropy。根據 entropy 的定義：

- Entropy 越低、(1-entropy) 越高，代表詞的分佈越平均、攜帶越少的資訊量（例如「的」、「在」等等，幾乎在所有的新聞都會出現，在考慮時就以較低的權重去考慮）。
- Entropy 越高、(1-entropy) 越低，代表詞的分佈較為稀疏、攜帶越多的資訊量（例如「黑人」、「歐巴瑪」等等，較能代表該新聞的主題）

[詞]	[1-entropy]
總統	0.401
美國	0.526
首位	0.185
黑人	0.069
歐巴瑪	0.067
的	0.938
在	0.927

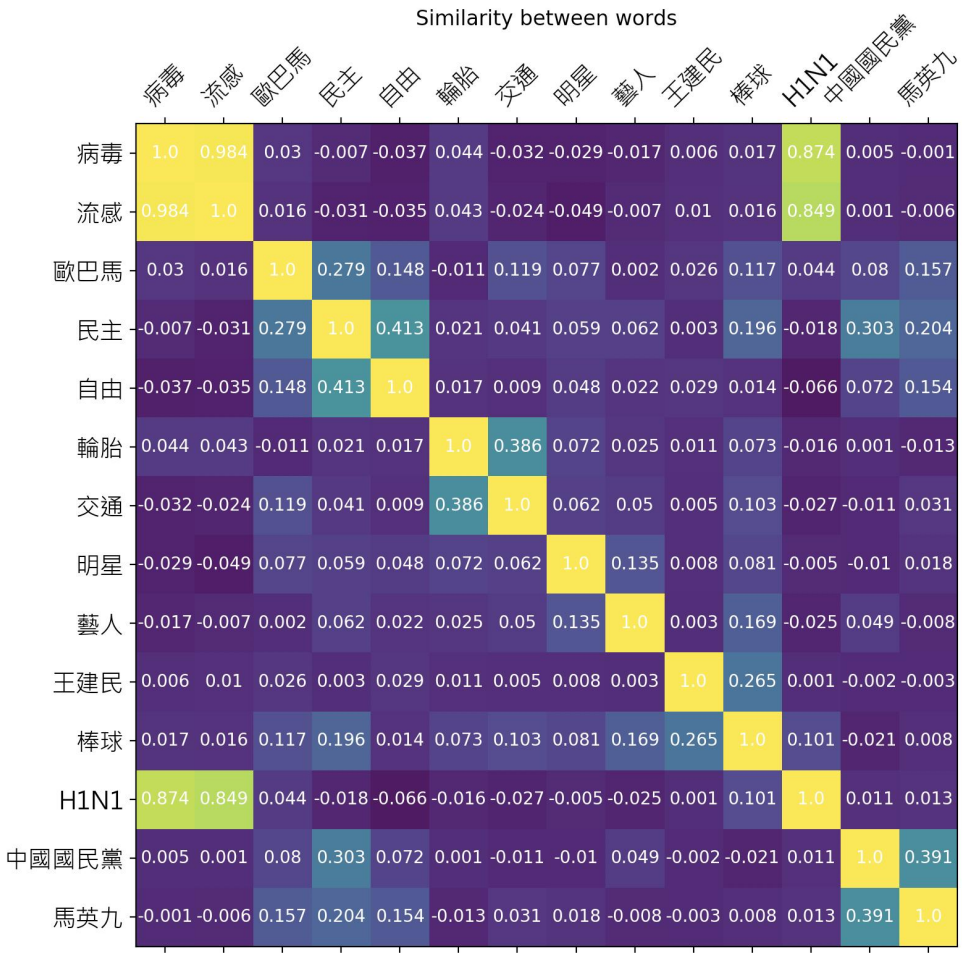
3-1. 字詞分群結果

首先我們觀察字詞們投影到 R 維空間後相似度的關係。相似度的定義如下（即 cosine similarity）：

$$\text{sim}(\underline{u}_i, \underline{u}_j) = \frac{\underline{u}_i \cdot \underline{u}_j}{|\underline{u}_i| \cdot |\underline{u}_j|}$$

其中底線的向量 $\underline{u}_i$、$\underline{u}_j$ 是詞投影至語意空間的結果，而 $-1 \leq \text{sim}(\underline{u}_i, \underline{u}_j) \leq 1$，越高代表在語意空間的相似度越高。

我們取了 14 個詞並交叉計算相似度，作圖如下：



如果將相似度較高的詞分類，大約可以獲得以下分類：

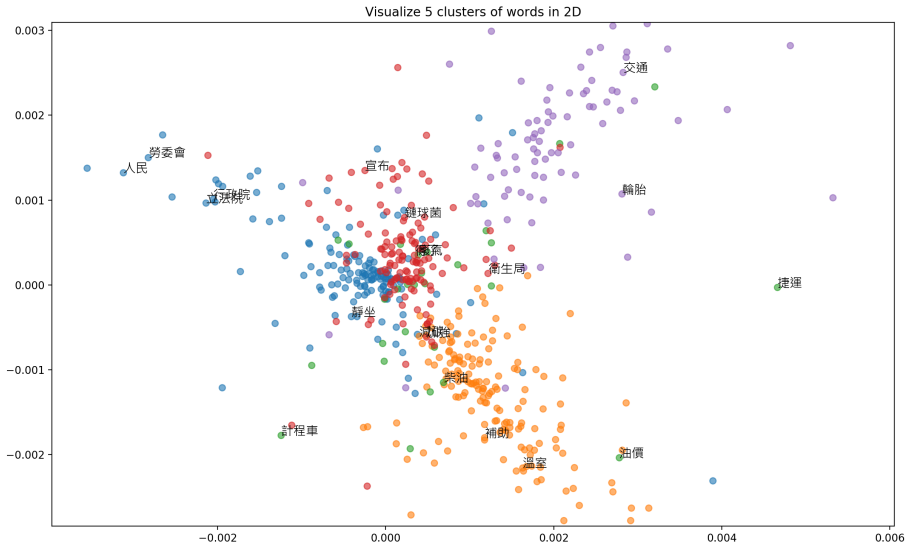
類別編號	詞
Class 0	病毒、流感、H1N1
Class 1	民主、自由
Class 2	輪胎、交通
Class 3	明星、藝人
Class 4	王建民、棒球
Class 5	中國國民黨、馬英九

其中可以觀察到以下幾點：

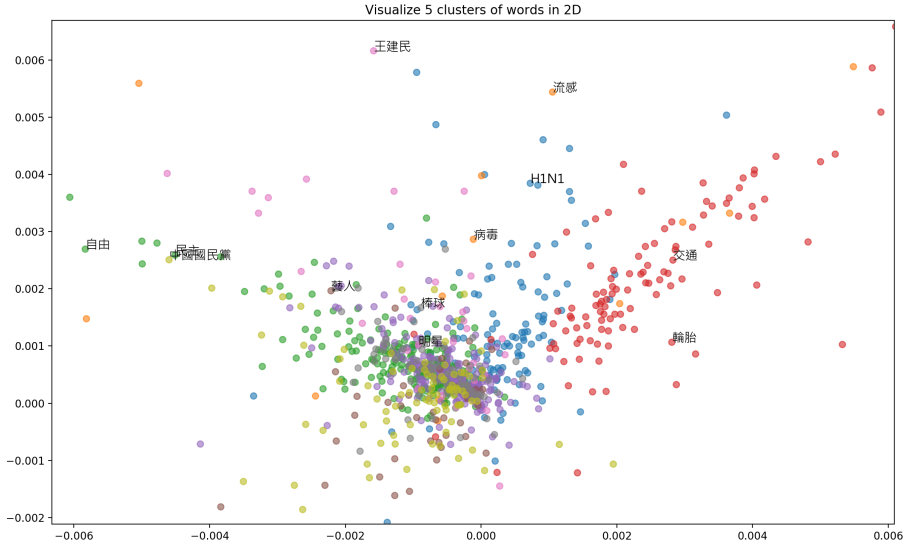
1. 對角線的值都為 1.0，因為向量與自己的夾角為 0，而 $\cos 0 = 1$。
2. 在 class 0 中的相似度是很明顯的，互相的相似度都有超過 80%，我們想其背後的原因是有「病毒」的文章中幾乎都有出現「流感」、「H1N1」，其中一個例子如下（H1N1 新型 流感 包含 四種 病毒 民視新聞 黃文玲 綜合報導 H1N1 新型 流感 病毒 來勢洶洶 ， 這個 新型的 病毒 ， 外界 對 H1N1 它的 認識 並不多 ， 目前 全球的 專家 ， 正努力 的 分析 ， 這個 新型 病毒 ， 究竟 有什麼 特點 ？ 請看 以下 這個 報導 。）
3. 除了相關的詞會有較高的相似度外，語意上很不相干的詞也反應出較低的相似度。例如「流感」與「自由」的相似度為 0.035，甚至是小於 0

的。

接者我們試著視覺化用 K-means 演算法將詞向量分群的結果。首先我們從 R 個維度中找到兩個對於分類最有代表性的維度，並將詞向量投影於此二維平面上，作圖如下（不同顏色代表被 K-means 演算法分到不同的類別）：



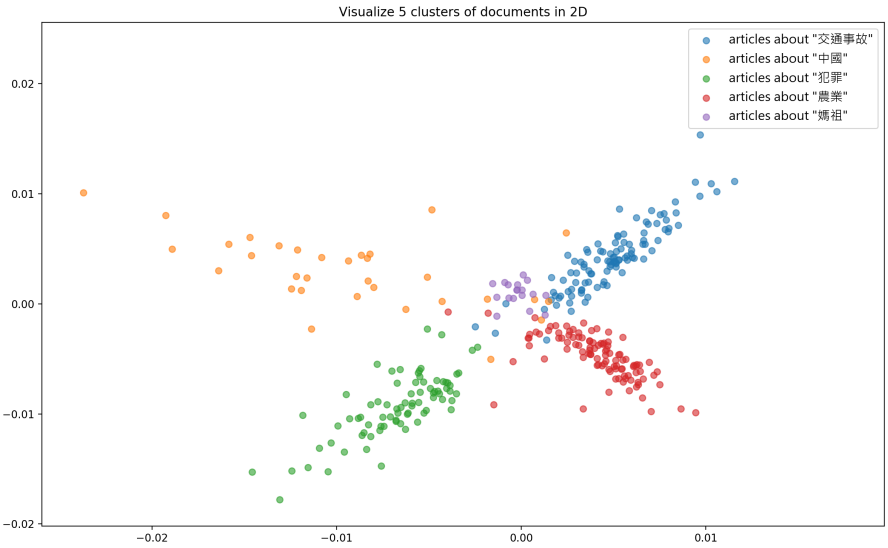
在圖中我們可以發現，像是橙色（溫室、油價、柴油等等，可能是與能源相關的主題）就分佈在左下角的位置，而藍色（人民、勞委會、行政院等等，可能是與政治相關的主題）就分佈在左上角，右上角則是與交通比較相關的。



而將先前分類的例子視覺化後發現類別的趨勢還是有顯示出來，像是雖然「民主」與「中國國民黨」是不同類，但他們的位置卻很相近（他們也有不低的相似度）。而因為我們只是取其中兩維，因此不是完全反應他們在語意空間中的關係，像是「棒球」跟「明星」在圖中很相近，但其實他們的相似度只有 0.081。

3-2. 新聞分群結果

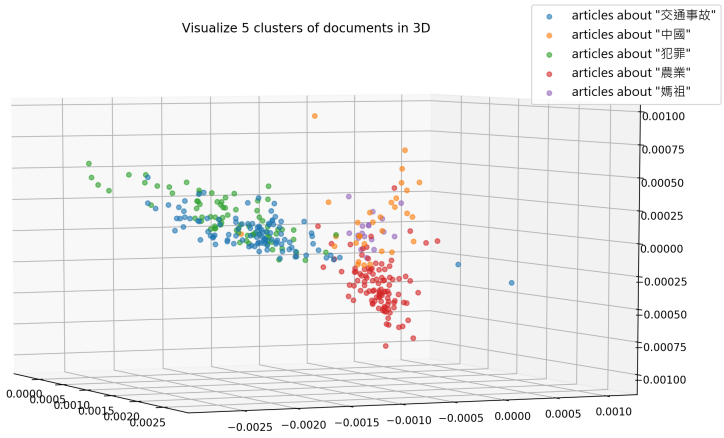
使用同樣的方法，將用新聞向量投影在二維平面後可以得到以下這張圖（不同顏色代表被 K-means 演算法分到不同的類別，而類別的label 則是我們根據該類別中的文章內容人工加上的標籤）：



與字詞分析有類似的結果，可以發現被 K-means 演算法分類到同類別的新聞，在此兩維度中皆有較為聚集的現象，而不同類別的則會分佈在空間中不同區塊。

前面我們已經知道不同類別會在某些維度中產生明顯的分群，但因為整個向量仍是高維度，我們難以觀測向量整題在 R 維空間的分佈狀況，因此進一步透過 PCA [6] 將新聞在 R 維子空間的向量降維到三維後，將其作圖在三維空間。而在三維空間的作圖中我們可以發現，不同類別的向量確實也會分群，且在空間的位置也跟其類別有所關係，像跟"交通事故"有關的新聞與跟"犯罪"有關的新聞分佈較為接近，這也是相當合理的，因為兩者之間互相的關係明顯跟其他類別高，兩者用到的詞彙較為接近(可能都有傷亡之類的詞彙等)，

因 此 其 向 量 也 較 為 接 近 。



心得

在第一次聽完 LSA 的概念後雖然覺得很有意思但也是有點一知半解，在這次的經驗中，我們透過參考原始論文、自己動手建立矩陣、分解矩陣更加了解其背後的概念，像是空間之間投影的關係、在「語意空間」中的分佈等等。雖然我們還沒辦法去猜測「語意空間」中每個維度代表的意義，但是透過視覺化的方式來呈現資料也讓我們覺得很有趣，不同類別的詞、新聞確實能在語意空間中表達出他代表的意義。

分工

	負責項目
林義閔	程式部分、報告撰寫
陳柏衡	程式部分、資料視覺化

參考資料

1. http://speech.ee.ntu.edu.tw/DSP2020Autumn/Slides/14.0.pptx
2. “Exploiting Latent Semantic Information in Statistical Language Modeling”, Proceedings of the IEEE, Aug 2000
3. https://github.com/sih4sing5hong5/icorpus
4. https://numpy.org/doc/stable/reference/generated/numpy.linalg.svd.html
5. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
6. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html