R-詞彙聚類、共現網路、中介度分析

作者：林士豪(LIN-CI-YUAN)

發布時間：國民黨初選公布前夕

APP 連結：https://reurl.cc/VaAGdb

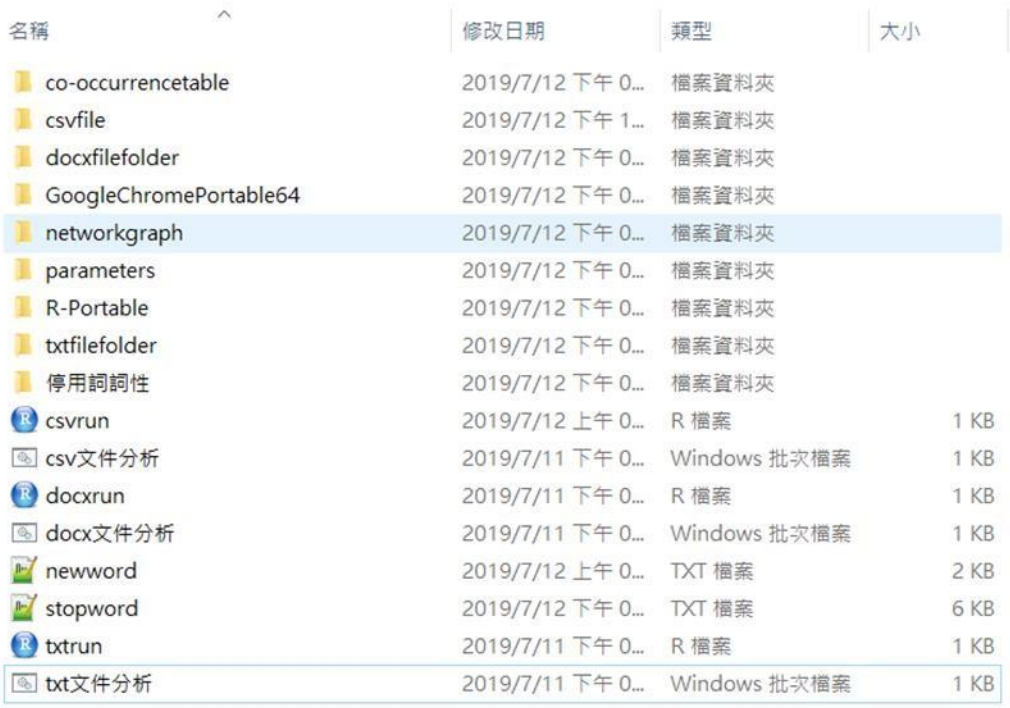
1.預先提醒：本程式預設使用 windows 作業系統，文字編碼係 UTF-8，僅有 docx 的分析文件能將其他編碼轉成 UTF-8；使用 mac 或 linux 的朋友，請參考我寫在 R 腳本中的 gist 連結。

2.分析文本：

(1)docxfilefolder：請將要分析的 docx 文件放入此資料夾中，本程式可批量處理文件，如果想先測試效果，可以參考我資料夾內預先放入的 30 則人民日報報導；pdf 也想跑共現分析的話，請先轉成 docx。

(2)csvfile：請將要分析的 csv 文件放入此資料夾中，一次只能處理一張表，如果你要處理大量的表格文字，請先事先將這些表格合併成一張表，並且欄位要對得上。

(3)txtfile：純文字檔案小，程式運算速度會更快，也占用更少的記憶體。



名稱	修改日期	類型	大小
co-occurrencetable	2019/7/12 下午 0...	檔案資料夾	
csvfile	2019/7/12 下午 1...	檔案資料夾	
docxfilefolder	2019/7/12 下午 0...	檔案資料夾	
GoogleChromePortable64	2019/7/12 下午 0...	檔案資料夾	
networkgraph	2019/7/12 下午 0...	檔案資料夾	
parameters	2019/7/12 下午 0...	檔案資料夾	
R-Portable	2019/7/12 下午 0...	檔案資料夾	
txtfilefolder	2019/7/12 下午 0...	檔案資料夾	
停用詞詞性	2019/7/12 下午 0...	檔案資料夾	
csvrun	2019/7/12 上午 0...	R 檔案	1 KB
csv文件分析	2019/7/11 下午 0...	Windows 批次檔案	1 KB
docxrun	2019/7/11 下午 0...	R 檔案	1 KB
docx文件分析	2019/7/11 下午 0...	Windows 批次檔案	1 KB
newword	2019/7/12 上午 0...	TXT 檔案	2 KB
stopword	2019/7/12 下午 0...	TXT 檔案	6 KB
txtrun	2019/7/11 下午 0...	R 檔案	1 KB
txt文件分析	2019/7/11 下午 0...	Windows 批次檔案	1 KB

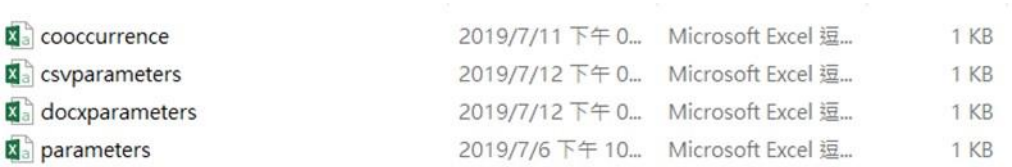
3.重要資料夾與參數：

(1)co-occurrencetable：如果順利執行程式，會自動產生一張 csv 表(能用 excel 打開)，告訴你文件中不同詞彙共同出現的次數狀況。

(2)networkgraph：docx、txt 的文件分析完成後，會產生一張共現網路圖，但是程式也會在網頁上產生另外兩種版本的共現網路圖，內容大同小異，不過網頁上的圖參數會比較詳細。

(3)parameters：最重要的參數資料夾，內含以下 csv 檔(建議用 excel 開啟)

1.以下資料夾分別對應 csv、docx、txt 三種程式。每一參數表均含有 supp、conf 欄。



cooccurrence	2019/7/11 下午 0...	Microsoft Excel 逗...	1 KB
csvparameters	2019/7/12 下午 0...	Microsoft Excel 逗...	1 KB
docxparameters	2019/7/12 下午 0...	Microsoft Excel 逗...	1 KB
parameters	2019/7/6 下午 10...	Microsoft Excel 逗...	1 KB

supp：中文翻譯叫共現率，是每個詞在分析的文件中，和另一個詞共同出現的機率，

數值是 0-1，目前文檔內有我自己測出的默認數值，如果你要換文件分析，請次幾多跑幾次，測測看需要呈現多少共現率的詞組。

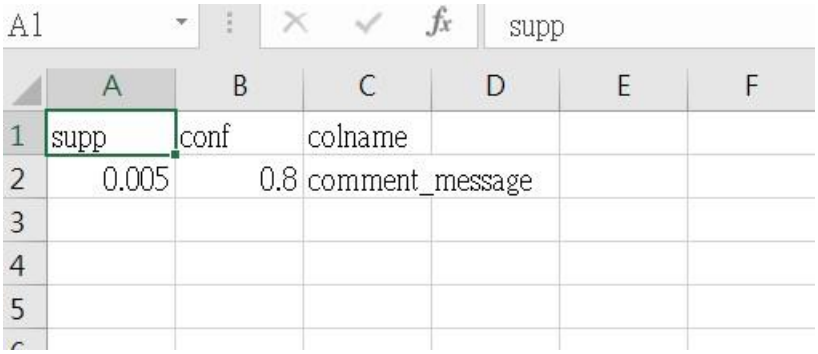
※共現網絡只會呈現共現率前 100 的節點，太多的自動跳過(會算太久)。

conf：信賴區間，supp+n 個標準差的 conf=共現率的大致上界。

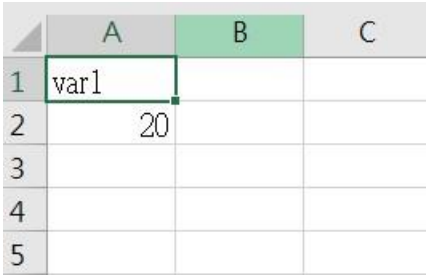
※supp 是最小共現率，最大共現率會在 supp+n*conf 中出現，但最小共現率就是 supp 填入的值，低於 supp 共現率的詞彙不會在共現網絡中。

colname：行名稱，僅在 csvparameters 出現，指定匯入的 csv 表格分析行名，該行就會被當作要分析的表格文字。

※只能分析一行，多行文字要分析請自己併成一行。



(4)coocurrence：共現次數表，如果共現率控制詞彙節點的效果不佳，可以打開此表，設置共現詞的頻率下限，以默認設置為例，共現次數低於 20 次的詞彙會被自動剔除。



5.分析平台介紹：

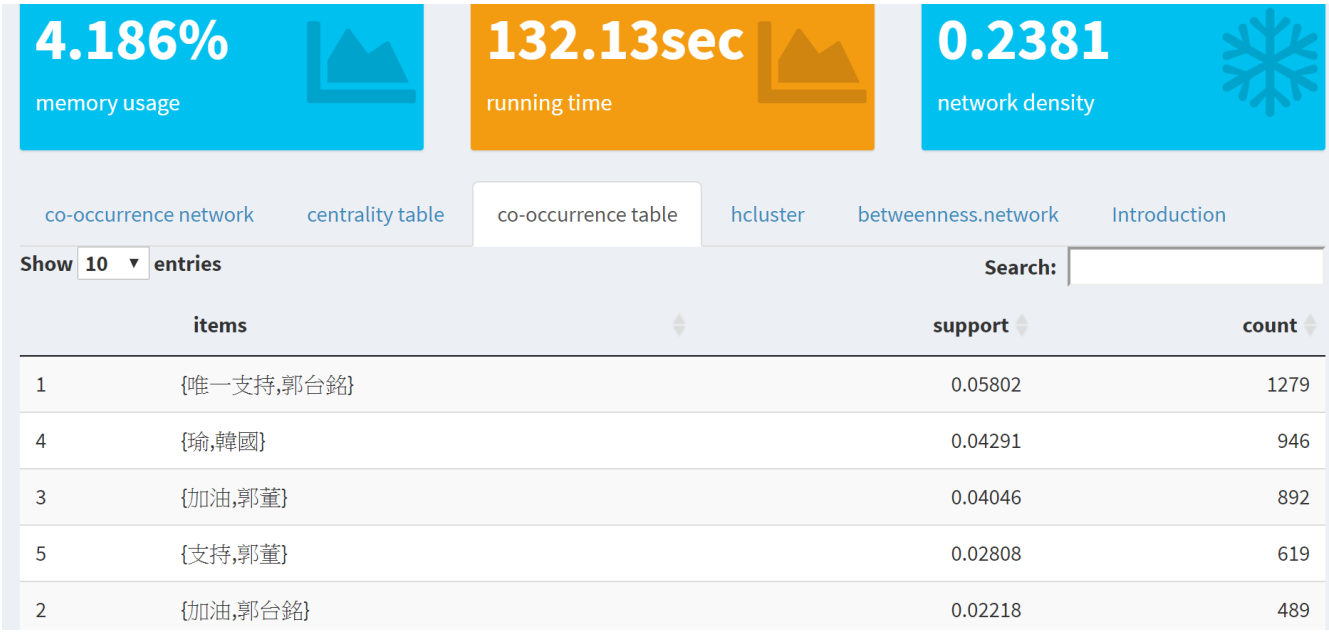
(1)Co-occurrence network：使用 networkd3 製成的共現網絡，上色方式依特徵中心度(eigen-centrality)會標記三類群組。第一類是特徵中心度最大值的核心節點、第二類是大於特徵中心度 75 百分位數的節點、第三類是低於特徵中心度平均值的節點。

※當最大值不只一個詞彙節點時，只會標記出兩種顏色。

(2)Centrality table：特徵中心度、中介度(betweenness)、內向中心性(indegree centrality)、外向中心性(outdegree centrality)。

※非閉環的網絡內向中心性=外向中心性。

(3)co-occurrence table：



	items	support	count
1	{唯一支持,郭台銘}	0.05802	1279
4	{瑜,韓國}	0.04291	946
3	{加油,郭董}	0.04046	892
5	{支持,郭董}	0.02808	619
2	{加油,郭台銘}	0.02218	489

共現表

Items：共現的詞組

Support：共現率 Count：共現次數

(4)階層聚類**(**分群**)-**平均距離法

Cluster-termrank：以本圖為例，選擇詞彙頻率前三十高的詞，進行分群。

Reanalysis cluster：選好詞彙頻率後，請按下此紐，圖就會跑出來了。

6.數字盒：

1.記憶體使用比例：就是記憶體使用比例，超過 50%藍色會轉黃色，意味著你最好換台電腦跑這個程式

2.執行秒數：超過 120 秒會轉黃色，可以選擇把文件拆分、或 docx 轉成 txt 之類的，加快執行速度。

※此秒速未計入程式所需套件(package)的載入時間。

3.共現網絡密度(network density)：這整張網到底有多密？！詳細公式請找本社會網絡的教科書來讀吧，密度大於 0.5 會轉黃，在共現網絡中 density 愈大，代表文件中的概念愈緊密(也愈洗腦)。