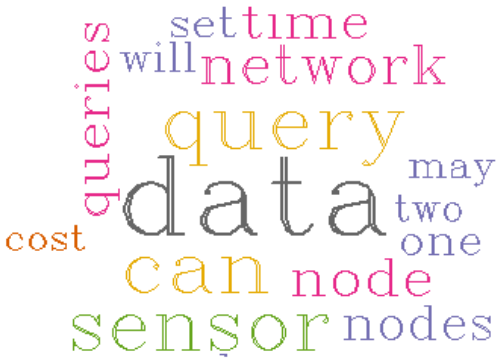
Information Retrieval : Project 2

Student : 610721204　陳克威



1. Document representation

1-1 DataSet：

此資料集為分散的論文資料集，總共有 1,167 筆文件資料，此專題目標為將文件分群，使用 R 語言實作。

1-2 Data preprocessing：

(1) 將資料讀入建立語料庫(Corpus)。
(2) 語料庫：將所有英文字母轉換為小寫。
(3) 語料庫：去除數字。
(4) 語料庫：去除標點符號。
(5) 語料庫：去除數字。
(6) 語料庫：去除多餘的空白符號。
(7) 語料庫：去除英文停詞 (使用 tm package)。
(8) 將語料庫建立成 DocumentTermMatrix 的形式。

	absence	address	addresses	approach	assume
1	1	1	1	1	1
2	0	0	0	0	0
3	0	0	1	0	2
4	0	0	0	0	0
5	0	0	0	1	0

(9) 計算此 DocumentTermMatrix 的 TF-IDF 權重值，將其轉換成 Weight-TF-IDF 版本的 DocumentTermMatrix。

	absence	address	addresses	approach	assume
1	0.048	0.033	0.042	0.016	0.024
2	0	0	0	0	0
3	0	0	0.023	0	0.026
4	0	0	0	0	0
5	0	0	0	0.036	0

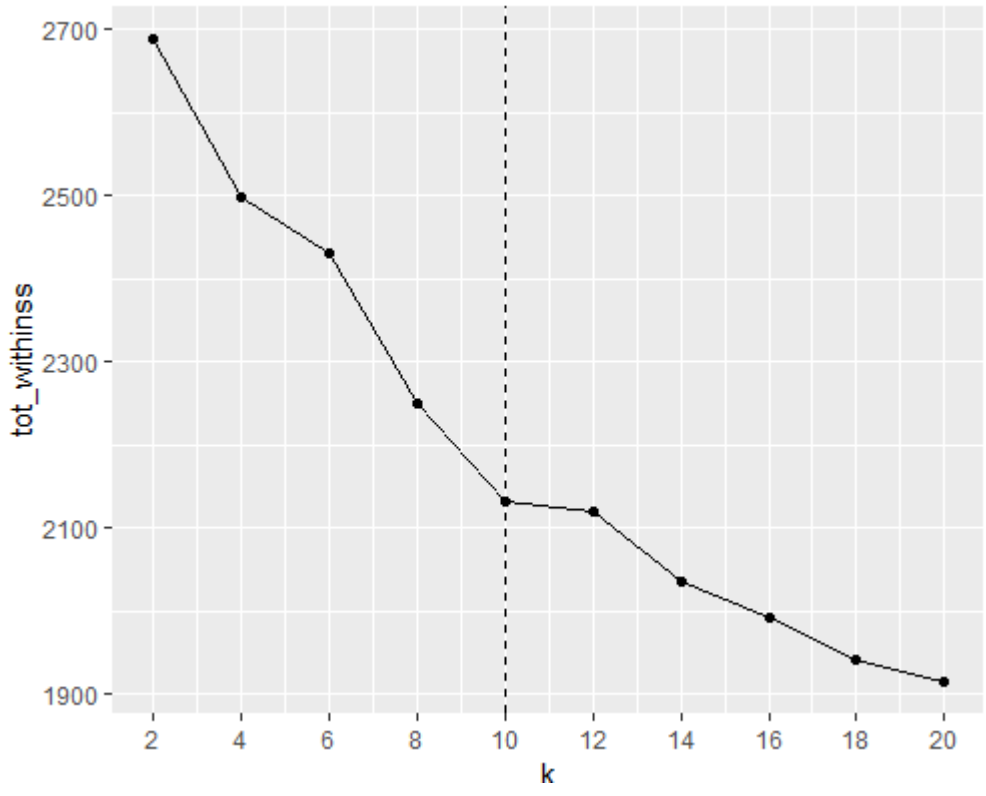
2. K-means with cosine similarity：

2-1 Cosine similarity ：

因 R 語言內建的 K-means 不提供 Cosine similarity，因此這裡參考[1]的程式碼，將 Cosine similarity 建立成一個 Function 方便之後計算距離。

2-2 Elbow method (clustering) ：

分群的目的是「使群內的總變異最小；使群間的總變異最大」[2]，為了決定分群數量，這裡選擇使用 Elbow method 來決定分群數量，基於下圖的結果，這裡選擇 K=10 為我們的分群數。



3. Result presentation

	8	6	9	7	5	10	1	3	4	2
num_doc	411	183	141	122	63	63	59	49	46	29

總共分為 10 群，每一群的文件數量如圖所示，因為 K-means 的質量很大程度取決於 Initial centroid，因此在做 Elbow method 來決定最終分群數量時可以設定 set.seed()，來避免測試結果有所不同。

根據文件本身的性質不同，可以嘗試不同的前處理方式，特別是停詞(stop words)的部分，另外若 DocumentTermMatrix 的維度過大，可能會需要做降維的處理。

4. Reference

1. Thomas W. Jones, document clustering(2019-04-17), https://cran.r-project.org/web/packages/textmineR/vignettes/b_document_clustering.html. (last access:2019/12/25)

2. skydome20, R 筆記 – (9)分群分析(Clustering) (2016/06/06), http://rpubs.com/skydome20/R-Note9-Clustering. (last access:2019/12/25)