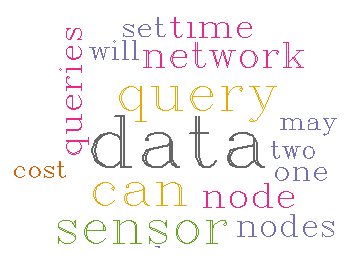
Information Retrieval : Project 2

Student : 610721204　陳克威



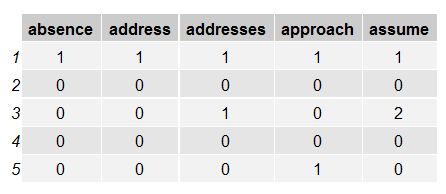
1. Document representation

1-1 DataSet：

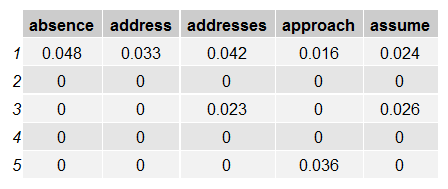
此資料集為分散的論文資料集，總共有1,167筆文件資料，此專題目標為將文件分群，使用R語言實作。

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。



1. 計算此DocumentTermMatrix的TF-IDF權重值，將其轉換成Weight-TF-IDF版本的DocumentTermMatrix。



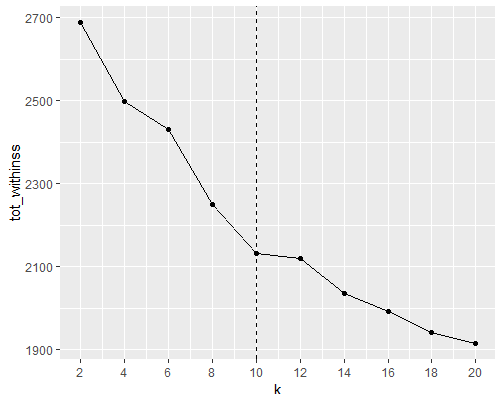
1. K-means with cosine similarity：

2-1 Cosine similarity ：

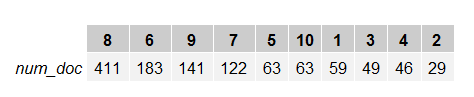
因R語言內建的K-means不提供Cosine similarity，因此這裡參考[1]的程式碼，將Cosine similarity建立成一個Function方便之後計算距離。

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。



1. Result presentation



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。

1. 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)