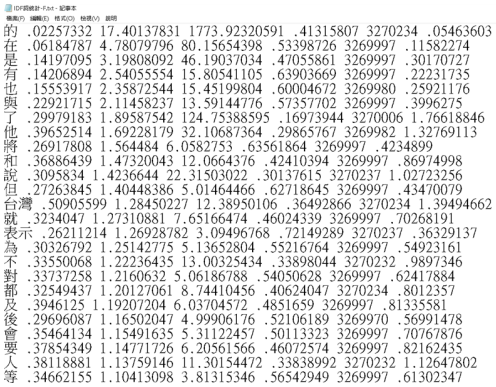
「教師跨領域研究社群」-文字探勘與資料庫建置

時間：2022-11-11
地點：傳播館 Q403

唐大崙分享自己過去耗費數年，從網路上爬取新聞文章，所建立的 300 多萬則新聞資料庫中，辛苦建立出來的、近 60 萬個中文詞的統計特徵資料庫，包括 TF、IDF、分布標準差等資料，依據此資料庫便能輕鬆判斷一則新聞的關鍵詞究竟是哪一些詞彙，也可以依據關鍵詞與詞彙所在相對位置的權重調整，計算兩篇文章相似度。這種相似度計算結果，與人的主觀判斷相當接近，表示這個演算法可以被接受用來追蹤語意相似的文章。不過在語意追蹤計算之前，需要預先建立正確斷詞的過度檔案，這是比較複雜的過程。但是，這樣的資料庫是否適用於法律契約的文章？仍是個必須真實嘗試才能確定的實徵性問題。因此，唐大崙與蔡明修老師約定下次再進一步拿真實契約文章來嘗試，也討論到先合作開設一個免費為人計算中文文章相似性的服務網站，等技術更精進，再進一步客製化為建築契約文件做語意追蹤的服務，這將可開啟另一種跨領域的合作範例。

詞頻文字雲 與 詞關鍵度文字雲 的差別

唐大崙以同一篇新聞做詞頻分析 對比於 關鍵詞分析所做的文字雲顯示，有很大差異。
唐大崙老師先分享文字探勘部分的經驗與做法。



的	.02257332	17.40137831	1773.92320591	.41315807	3270234	.05463603
在	.06184787	4.78079796	80.15654398	.53398726	3269997	.11582274
是	.14197095	3.19808092	46.19037034	.47055861	3269997	.30170727
有	.14206894	2.54055554	15.80541105	.63903669	3269997	.22231735
也	.15553917	2.35872544	15.45199804	.60004672	3269980	.25921176
與	.22921715	2.11458237	13.59144776	.57357702	3269997	.3996275
了	.29979183	1.89587542	124.75388595	.16973944	3270006	1.76618846
他	.39652514	1.69228179	32.10687364	.29865767	3269982	1.32769113
將	.26917808	1.564484	6.0582753	.63561864	3269997	.4234899
和	.36886439	1.47320043	12.0664376	.42410394	3269997	.86974998
說	.3095834	1.4236644	22.31503022	.30137615	3270237	1.02723256
但	.27263845	1.40448386	5.01464466	.62718645	3269997	.43470079
台灣	.50905599	1.28450227	12.38950106	.36492866	3270234	1.39494662
就	.3234047	1.27310881	7.65166474	.46024339	3269997	.70268191
表示	.26211214	1.26928782	3.09496768	.72149289	3270237	.36329137
為	.30326792	1.25142775	5.13652804	.55216764	3269997	.54923161
不	.33550068	1.22236435	13.00325434	.33898044	3270232	.9897346
對	.33737258	1.2160632	5.06186788	.54050628	3269997	.62417884
都	.32549437	1.20127061	8.74410456	.40624047	3270234	.8012357
及	.3946125	1.19207204	6.03704572	.4851659	3269997	.81335581
後	.29696087	1.16502047	4.99906176	.52106189	3269970	.56991478
會	.35464134	1.15491635	5.31122457	.50113323	3269997	.70767876
要	.37854349	1.14771726	6.20561566	.46072574	3269997	.82162435
人	.38118881	1.13759146	11.30154472	.33838992	3270232	1.12647802
等	.34662155	1.10413098	3.81315346	.56542949	3269997	.61302347

唐大崙說明多篇文章之間的相似度資料可用以進行像這類分群的語意追蹤網絡圖。
唐大崙談到要計算出關鍵詞需要像這樣的中文字詞分布的統計資料庫。