Group 12

[銀行、台灣、日本]

此三類的 Keywords，我們以(1+LOG(TF))*DF 卡方值之結果作為指標依據。將公式中原本的 IDF 以 DF 卡方值當作權重代替，因為 DF 卡方值包含了相關類別的資訊在裡面，比 IDF 更周全。

ex.

8799	docs	
編號	詞	(1+LOG(TF))*DF卡方值
1	日本	24087.38081
2	夏普	7793.95261
3	安倍	4632.589985
4	熊本	4567.005655
5	東京	4492.176216
6	日圓	4240.689403
7	都市天氣預報	4131.488521
8	都市	3563.574393
9	日股	3456.313862
10	台日	2371.388205
11	夏變	2354.892551
12	食品	1958.654979
13	核災	1881.202312
14	日媒	1839.177532
15	美日	1517.242578
16	日銀	1255.891176
17	熊本強震	1255.891176
18	赴日	1178.395155
19	日韓	1174.276204
20	核災食品	1093.082852
21	國際	1007.121272
22	亞股	952.8113283
23	福島	950.6821125
24	負利率	946.4781568
25	蓮舫	927.9378078
26	知事	871.6991528
27	北韓	865.800508
28	京都知事	830.6718284
29	東京都知事	830.6718284
30	護漁	828.5803926

[信用卡、匯率、台積電]

此三類的 Keywords，我們將 TF 卡方值、DF 卡方值各自排名（Rank)後，以兩者排名各 50% 的權重，最後找出新的排名做為最終指標依據。其中我們發現單獨看一種卡方值較有偏頗的顧忌，因此以併看的方式挑出與類別最相關的字詞。

ex.

1746	docs						
編號	詞	TF卡方值(原	DF卡方值(原	TF卡方RAN	DF卡方RAN	TF+DF卡方I	FINAL RANI
1	台積電	231489	88092	2	2	4	2
2	台股	68476	12078	4	10	14	5
3	權值股	22971	14545	8	9	17	7
4	大立光	18188	18119	16	4	20	8
5	張忠謀	20068	10368	10	11	21	9
6	股台積電	10952	10082	23	15	38	15
7	廠台積電	10000	10000	27	17	44	17
8	外資	23211	6222	7	51	58	21
9	董事長張忠	8712	8064	39	25	64	24
10	收盤報價	8100	8100	51	20	71	26
11	包括台積電	8281	7569	45	27	72	27
12	半導體	18944	5751	13	65	78	32
13	晶圓	11683	5663	21	67	88	38
14	龍頭台積電	8100	7225	56	34	90	39
15	股今	9553	5764	28	63	91	40
16	聯電	8064	6936	57	35	92	41
17	蘋概股	8762	5954	35	58	93	42
18	金融股	8201	6248	49	48	97	44
19	台積電股	8464	5776	41	61	102	46
20	台積電董事	7564	6498	65	43	108	49
21	積電董事	7564	6498	66	44	110	50
22	加權	7169	6767	77	38	115	53
23	聯發科	6800	6820	81	36	117	54
24	加權指數	7344	6241	72	49	121	55
25	概股	9060	5157	32	89	121	56
26	美股	8601	5396	40	82	122	57
27	權指數	7344	6241	73	50	123	58
28	發科	6638	6820	87	37	124	59
29	奈米	23701	4324	6	118	124	60
30	天收盤報價	6724	6724	85	42	127	62

(補充：整理資料方法）
STEP 1 我們移除所有的標點符號/英文字母/注音，並且也移除連接詞（ex. 的、在、了、和、與、於等等），用 replace 來補刪去的詞為空格。

STEP 2 切一篇文章中的所有的字串(標題與內文），在切字串時，如果字串中有空格則會直接被刪除，並且以一篇一篇文章的形式獨自儲存，以利之後在數文章篇數。

STEP 3 分標題跟內文，如果標題跟內文有出現六大中主題字詞，就會把它加到分類裡面。
用 counter 數這個詞總共出現幾次在這篇分類裡(TF)，用 counter 數有幾篇文章有這個詞(DF)，並計算找的詞在全部標題與文章的次數(全部 TF、全部 DF)。

STEP 4 做後續 tf-idf、MI、LIFT 等等的計算

種類	主題字詞	取詞範圍	字數	篇數
銀行	銀行	tf>=50 and df>=6	7136	6674
信用卡	信用卡	tf>=10	4554	653
匯率	匯率	tf>=20 and df>=2	4947	1951
台積電	台積電,"張忠謀"	tf>=20 and df>=2	5337	1746
台灣	台灣,"臺灣","台北","新北","臺北","高雄","台中","桃園","台南"	tf>=50 and df>=40	38026	39931
日本	日本,"東京","名古屋","大阪","神戶","澀谷","京都","安倍","晉三"	tf>=50 and df>=8	8740	8799

種類	主題字詞	取詞範圍	字數	篇數
銀行	銀行	tf>=50 and df>=6	7136	6674
信用卡	信用卡	tf>=10	4554	653
匯率	匯率	tf>=20 and df>=2	4947	1951
台積電	台積電,"張忠謀"	tf>=20 and df>=2	5337	1746
台灣	台灣,"臺灣","台北","新北","臺北","高雄","台中","桃園","台南"	tf>=50 and df>=40	38026	39931
日本	日本,"東京","名古屋","大阪","神戶","澀谷","京都","安倍","晉三"	tf>=50 and df>=8	8740	8799