



大數據與商業分析

HOMEWORK1

Group 12 徐熾鎔 廖文豪 鄭皓 陳宜君 陳牧忠 藤田教譽

PRESENTATION DETAILS

TOPICS TO DISCUSS

前處理

取詞範圍

篩選方法

結果

Demo



資料前處理

PREPROCESSING

讀檔

將hw1_text.xlsx讀進

移除標點符號、STOPWORDS

將標題及內文內所有的標點符號、英文字母、一些我們判定為Stopwords的字刪除補空格

N-GRAM

將標題及內文的字進行N-gram切割，如遇空格則把此字刪掉，最後將切完的字一則一則儲存（包含標題及內文）

計算所有字的DF及TF

利用counter計算所有字的DF和TF

找出在此種類的文章

找出標題或內文含種類關鍵字的文章

○○○○○●

資料前處理

PREPROCESSING

計算在此種類的文章的文字的TF DF
利用counter計算在此種類的字在此種類的DF
和TF

找出這些字對所有文章的TF DF
利用剛剛計算所有字的DF及TF去找出我們這些字在
全部文章的TF和DF

計算TF-IDF及全部TF-IDF

計算期望值、TF卡方、DF卡方、
MI值、LIFT值

輸出CSV檔案

會先篩選TF和DF（格式如老師給的
hw1_table.xlsx）



取詞範圍

PREPROCESSING

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



篩選方法

FILTER

銀行、台灣、日本

此三類的 Keywords，我們以 $(1+\text{LOG}(\text{TF})) * \text{DF}$ 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。

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。



RESULT & DEMO