第二次讀書會報告

組別：11

組員：

邱昱榕 N114320004

朱怡樺 N114320005

林威呈 N114320011

董力慈 N114320019

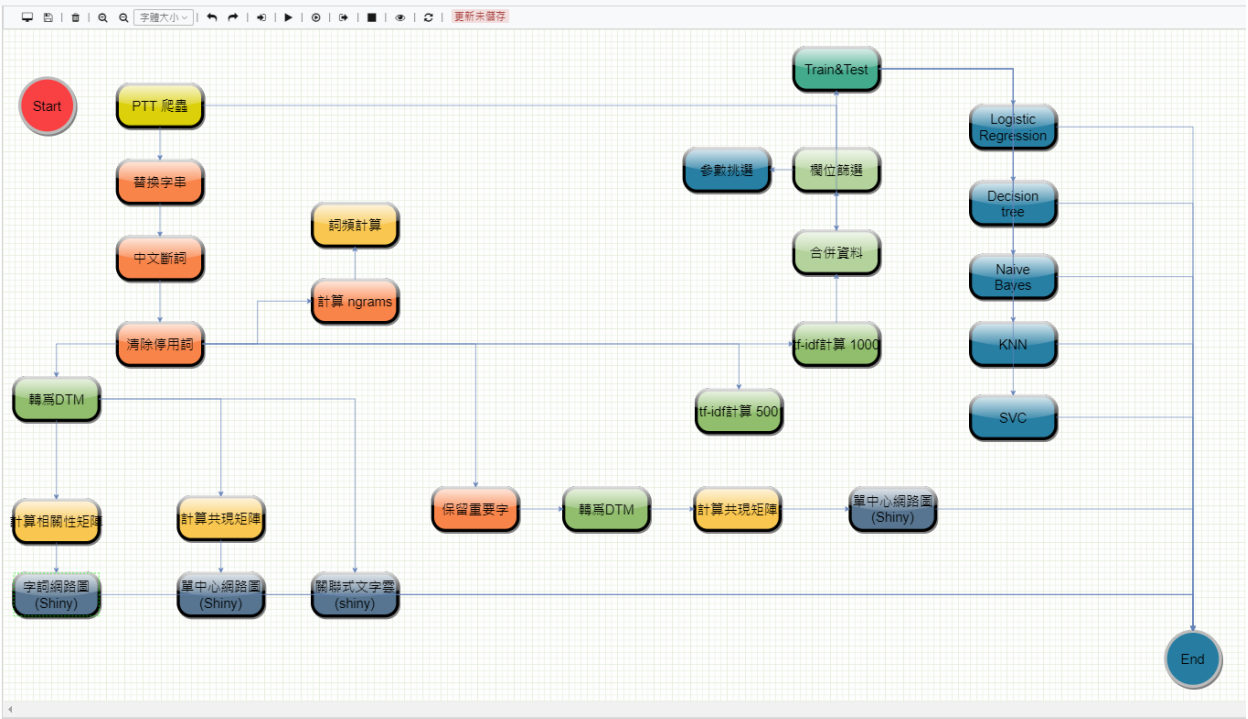
花秀旻 N114320021

趙丞德 N114320023

陳昱維 N114320026

李晉安 N114320028

下圖為我們這次整個分類器訓練的流程圖



本次報告主題標的：

取 Tarflow 中，soft_Job (軟體工作) 、Stock(股票) & Tech_Job (科技工作)

標的資料期間：

2024/01/01-2024/01/31

(36) 針對 result 欄位做<中文斷詞>

定義詞彙的部分有<現金股利、股票股利、台積電...等等>

(40) 使用功能<清除停用詞>，詳細設定可以參閱下圖，我們透過多次情緒分析結果的文字雲，來回觀察並增訂所需停用之字詞

選擇清除英文字母，因文章內容常用有網址、無效英文縮寫等

選擇清除數字，例如股票代碼等與情緒分析較無關之資料，以及技術性專有名詞等無情緒表達的字詞。

1. 常有文本的固定格式字詞，例如有依版規、內文、標題、資料來源等詞彙
2. 感謝詞包含謝謝、感謝等非表正面情緒的禮貌性用詞
3. 文章內表達時間序，例如過去、最近、現在、目前、未來、今天、今日、今年、去年、明年等詞彙
4. 常出現但與情緒分析較無相關之詞彙，例如成交量、股名、股票代碼、政府、產業等詞彙


文字探勘工作流程設計平台
昱維test6
問題回報
清除停用詞（40）
參數設定
Input - 36
任務結果
語言 *
Chinese
使用預設停止詞
是
是否清除單字元
是
是否轉為小寫英文
是
清除英文字母 *
是
清除數字 *
是
清除換行符號 *
是
清除特殊標點符號 *
是
清除html tag *
是
自定義停止詞
imgur
JPG
HTTP
資料來源
內文
儲存更改

計算 ngrams 的過程：

分詞：將文本分割成單詞或字符的序列。

建立 n-gram：根據 n 的值，將分詞後的文本序列組合成 n-gram。(n=2)

統計頻次：計算每個 n-gram 在文本中出現的頻次或概率。

"計算 ngrams" 功能是為了幫助我們快速生成並使用 ngrams 特徵進行文本分析或處理。

最後輸出結果 ngrams 數量為 : 237848

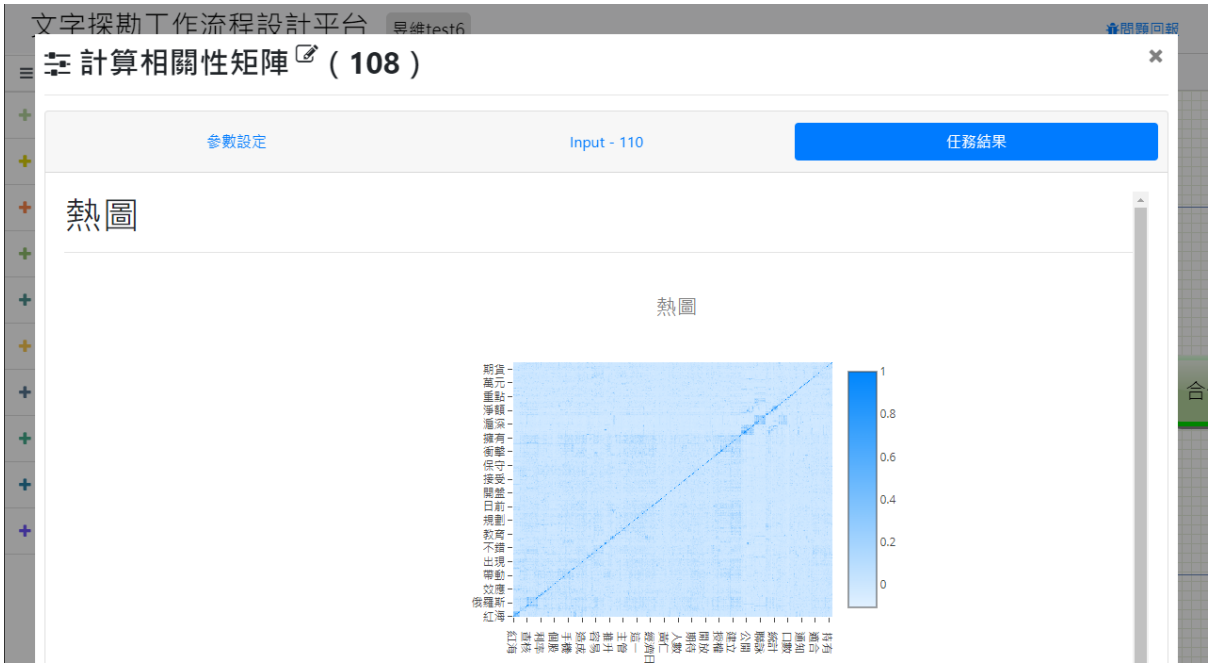
通過詞頻計算，我們了解文本中各個詞彙的重要性和出現頻率，從而進行詞彙的分析和挖掘。

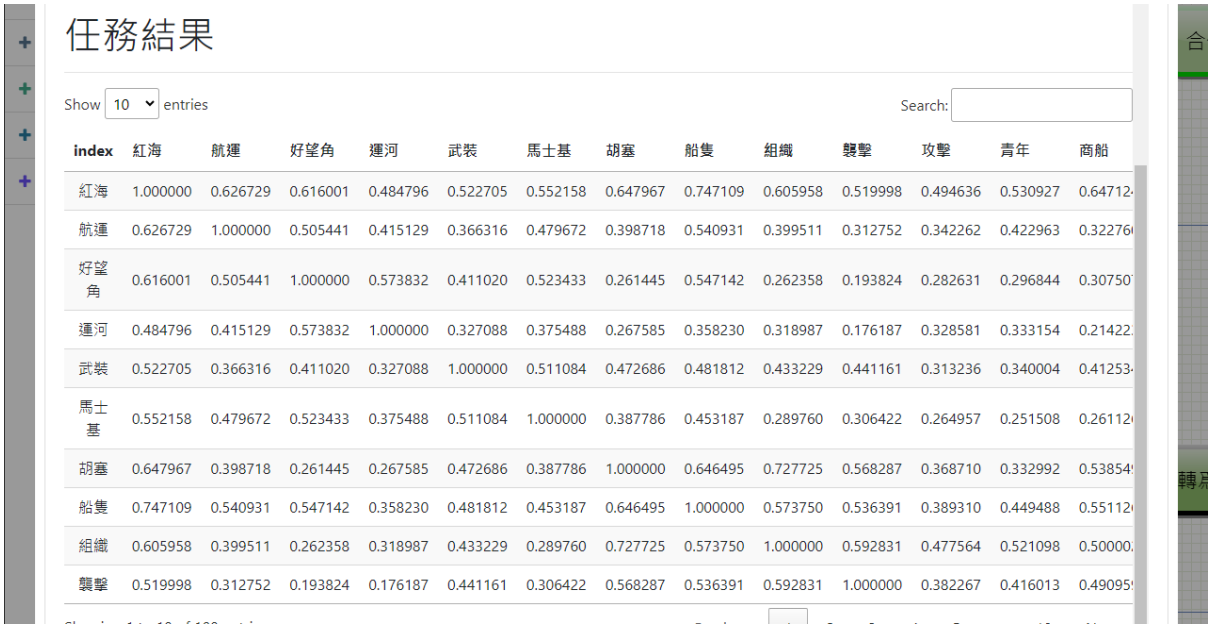
此時我們利用詞頻計算所產生出來的文字雲，可以馬上了解本文章的重點

根據詞彙表和詞頻信息，生成文檔詞頻矩陣（DTM），其中每一行代表一個文檔，每一列代表一個詞彙，單元格中的值表示對應詞彙在該文檔中的詞頻或其他統計信息，詞彙數量的值為1000

計算相關性矩陣的目的是幫助我們理解文本數據中各個文本之間的相似度或關聯程度。這對於文本分類、主題建模、情感分析等任務非常有用，可以幫助識別相似主題、找到相似文本或文本群體等。

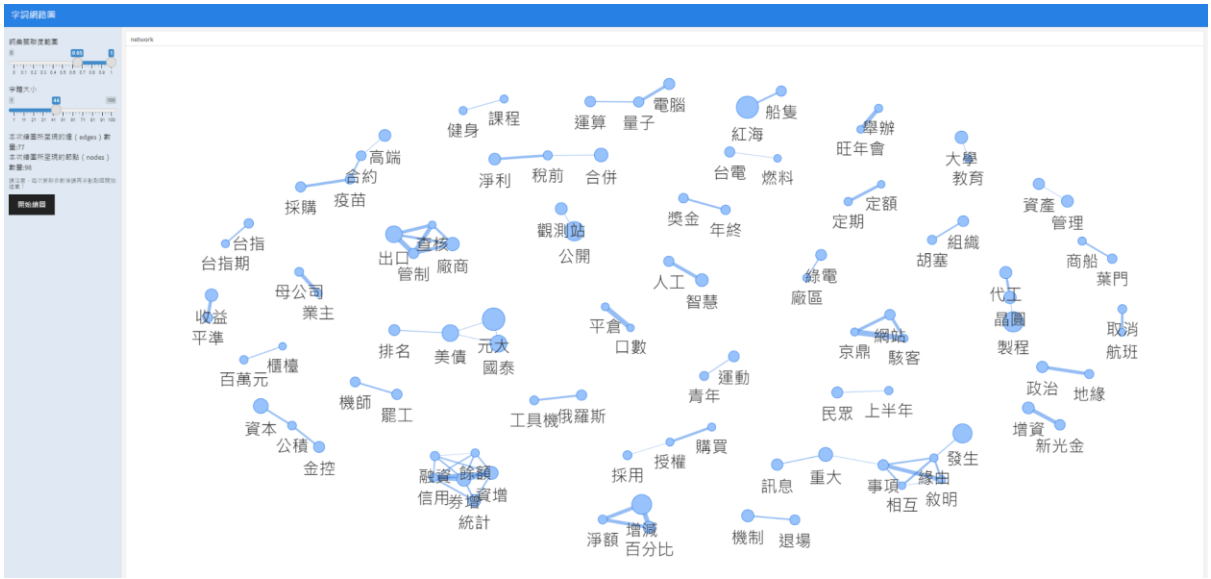
以下熱圖為展示數據之間的相似性、相關性和趨勢。

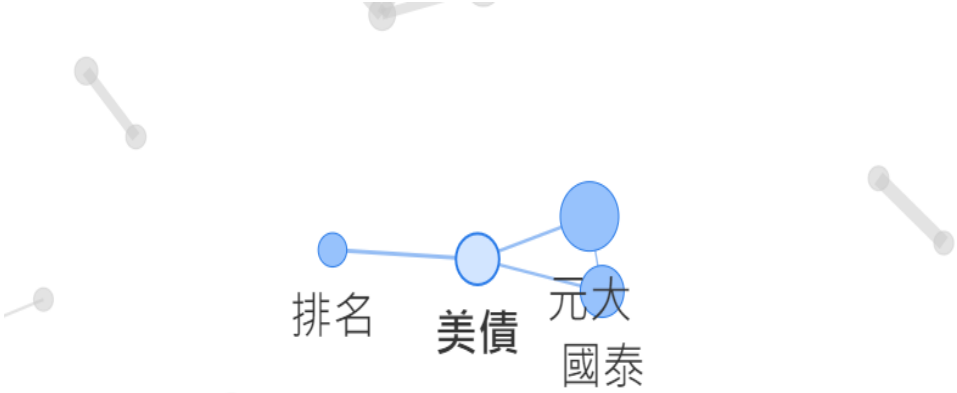


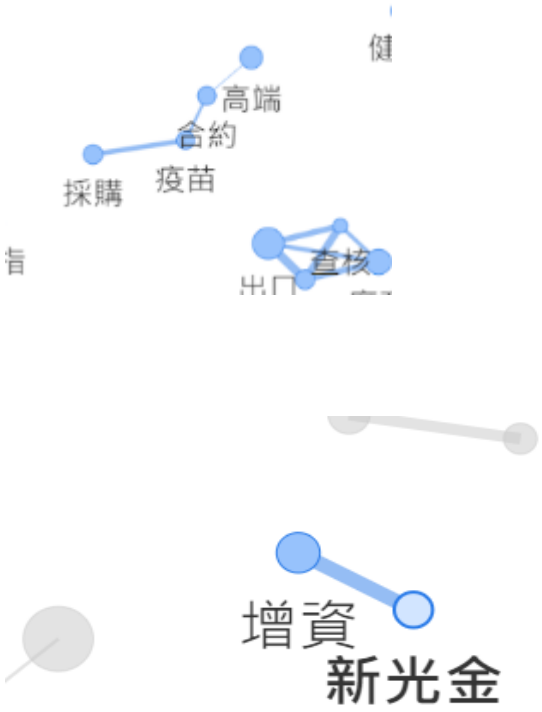


index	紅海	航運	好望角	運河	武裝	馬士基	胡塞	船隻	組織	襲擊	攻擊	青年	商船
紅海	1.000000	0.626729	0.616001	0.484796	0.522705	0.552158	0.647967	0.747109	0.605958	0.519998	0.494636	0.530927	0.64712
航運	0.626729	1.000000	0.505441	0.415129	0.366316	0.479672	0.398718	0.540931	0.399511	0.312752	0.342262	0.422963	0.32276
好望角	0.616001	0.505441	1.000000	0.573832	0.411020	0.523433	0.261445	0.547142	0.262358	0.193824	0.282631	0.296844	0.30750
運河	0.484796	0.415129	0.573832	1.000000	0.327088	0.375488	0.267585	0.358230	0.318987	0.176187	0.328581	0.333154	0.21422
武裝	0.522705	0.366316	0.411020	0.327088	1.000000	0.511084	0.472686	0.481812	0.433229	0.441161	0.313236	0.340004	0.41253
馬士基	0.552158	0.479672	0.523433	0.375488	0.511084	1.000000	0.387786	0.453187	0.289760	0.306422	0.264957	0.251508	0.26112
胡塞	0.647967	0.398718	0.261445	0.267585	0.472686	0.387786	1.000000	0.646495	0.727725	0.568287	0.368710	0.332992	0.53854
船隻	0.747109	0.540931	0.547142	0.358230	0.481812	0.453187	0.646495	1.000000	0.573750	0.536391	0.389310	0.449488	0.55112
組織	0.605958	0.399511	0.262358	0.318987	0.433229	0.289760	0.727725	0.573750	1.000000	0.592831	0.477564	0.521098	0.50000
襲擊	0.519998	0.312752	0.193824	0.176187	0.441161	0.306422	0.568287	0.536391	0.592831	1.000000	0.382267	0.416013	0.49095

通過分析字詞網路圖，可以發現詞彙之間的群體結構或關聯模式，從而識別文本中的主題或語義結構。

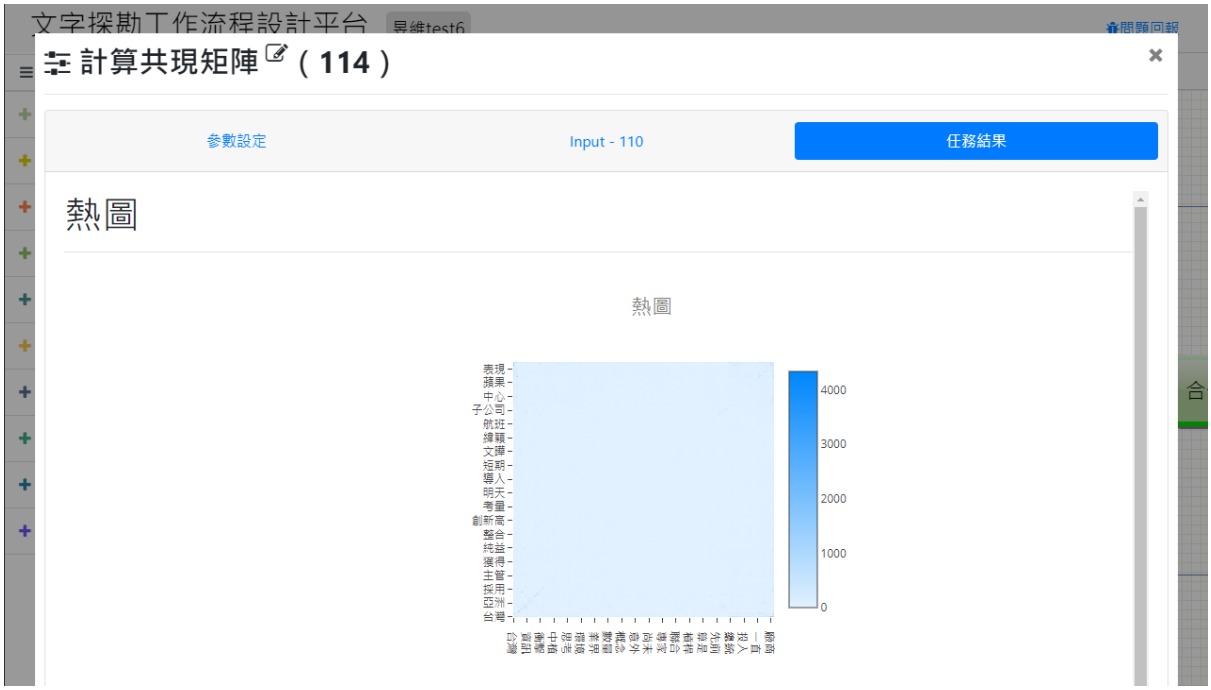





健
高端
合約
採購
疫苗
旨
查核
出口

增資
新光金

我們為了捕捉詞彙之間的語義關聯性,根據詞彙表和共現計算結果，生成共現矩陣，其中每一個元素表示對應詞彙之間的共現次數或其他統計信息。



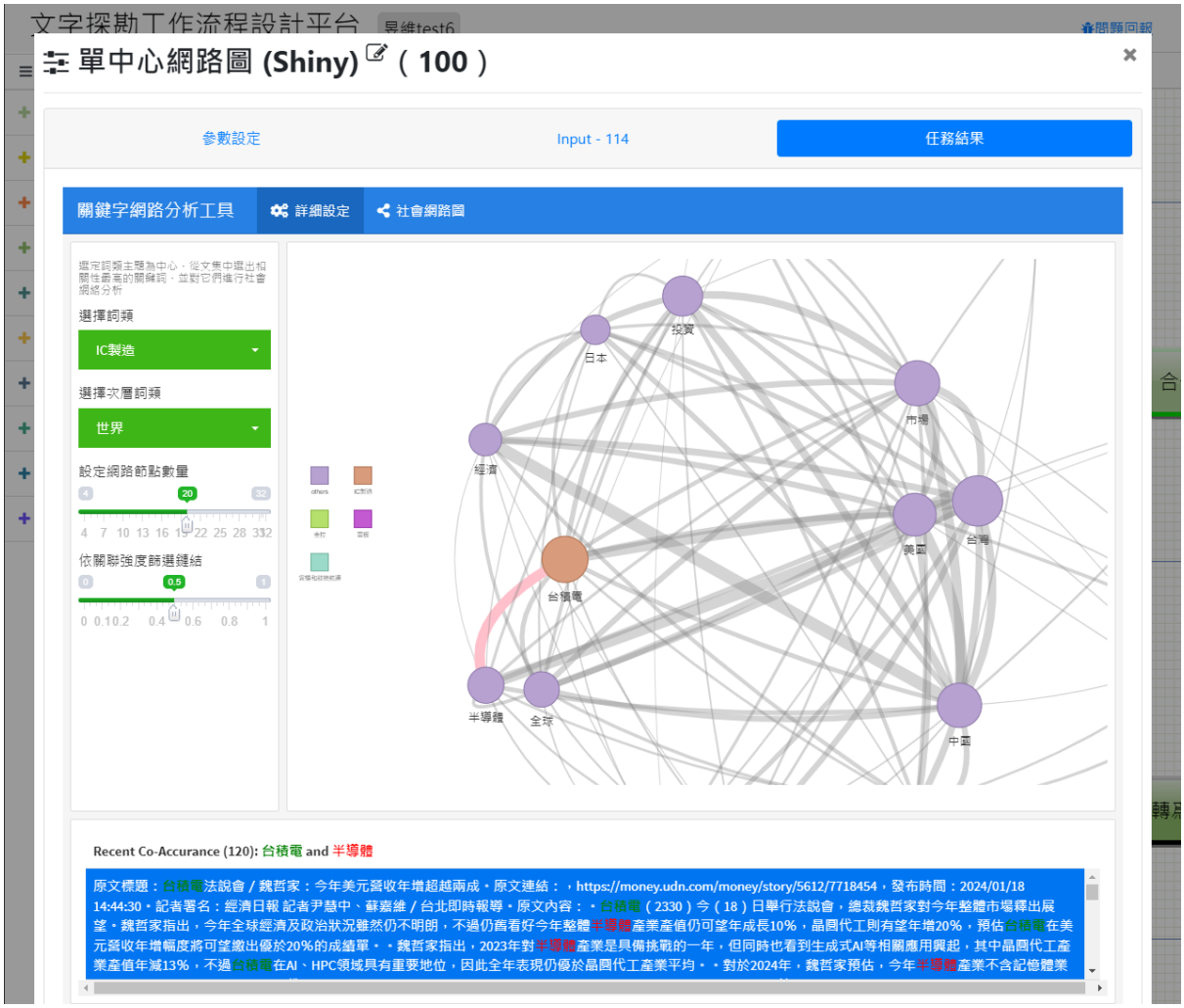
任務結果

Show 10 entries　　Search:

index	台灣	美國	營收	晶片	需求	預期	成長	持續	供應鏈	市值	科技	主要	發展	國際	日本	奈米	英特爾	製程
台灣	0.0	3211.0	417.0	755.0	497.0	364.0	886.0	914.0	1022.0	761.0	1273.0	658.0	967.0	1479.0	1044.0	436.0	89.0	464
美國	3211.0	0.0	308.0	818.0	333.0	374.0	540.0	554.0	327.0	142.0	562.0	606.0	473.0	806.0	770.0	389.0	345.0	413
營收	417.0	308.0	0.0	310.0	511.0	561.0	936.0	449.0	273.0	109.0	89.0	356.0	107.0	185.0	235.0	381.0	507.0	377
晶片	755.0	818.0	310.0	0.0	325.0	207.0	274.0	290.0	202.0	79.0	234.0	178.0	197.0	152.0	154.0	489.0	489.0	477
需求	497.0	333.0	511.0	325.0	0.0	242.0	403.0	402.0	144.0	79.0	214.0	239.0	226.0	154.0	217.0	358.0	175.0	310
預期	364.0	374.0	561.0	207.0	242.0	0.0	445.0	287.0	97.0	95.0	167.0	203.0	107.0	153.0	88.0	105.0	205.0	160
成長	886.0	540.0	936.0	274.0	403.0	445.0	0.0	376.0	326.0	211.0	335.0	286.0	212.0	219.0	144.0	239.0	198.0	279
持續	914.0	554.0	449.0	290.0	402.0	287.0	376.0	0.0	257.0	190.0	219.0	282.0	254.0	397.0	274.0	322.0	98.0	337
供應鏈	1022.0	327.0	273.0	202.0	144.0	97.0	326.0	257.0	0.0	312.0	72.0	100.0	66.0	441.0	78.0	174.0	139.0	241
市值	761.0	142.0	109.0	79.0	79.0	95.0	211.0	190.0	312.0	0.0	105.0	95.0	58.0	282.0	122.0	20.0	29.0	11.

Showing 1 to 10 of 100 entries　　Previous 1 2 3 4 5 ... 10 Next

我們先編輯了關鍵字點檔案，並選擇<IC 製造中>的<世界>來探索哪些字最常一起出現。其中<台積電>與半導體時常於文章中一起出現，也可於下方文章列表中找尋。



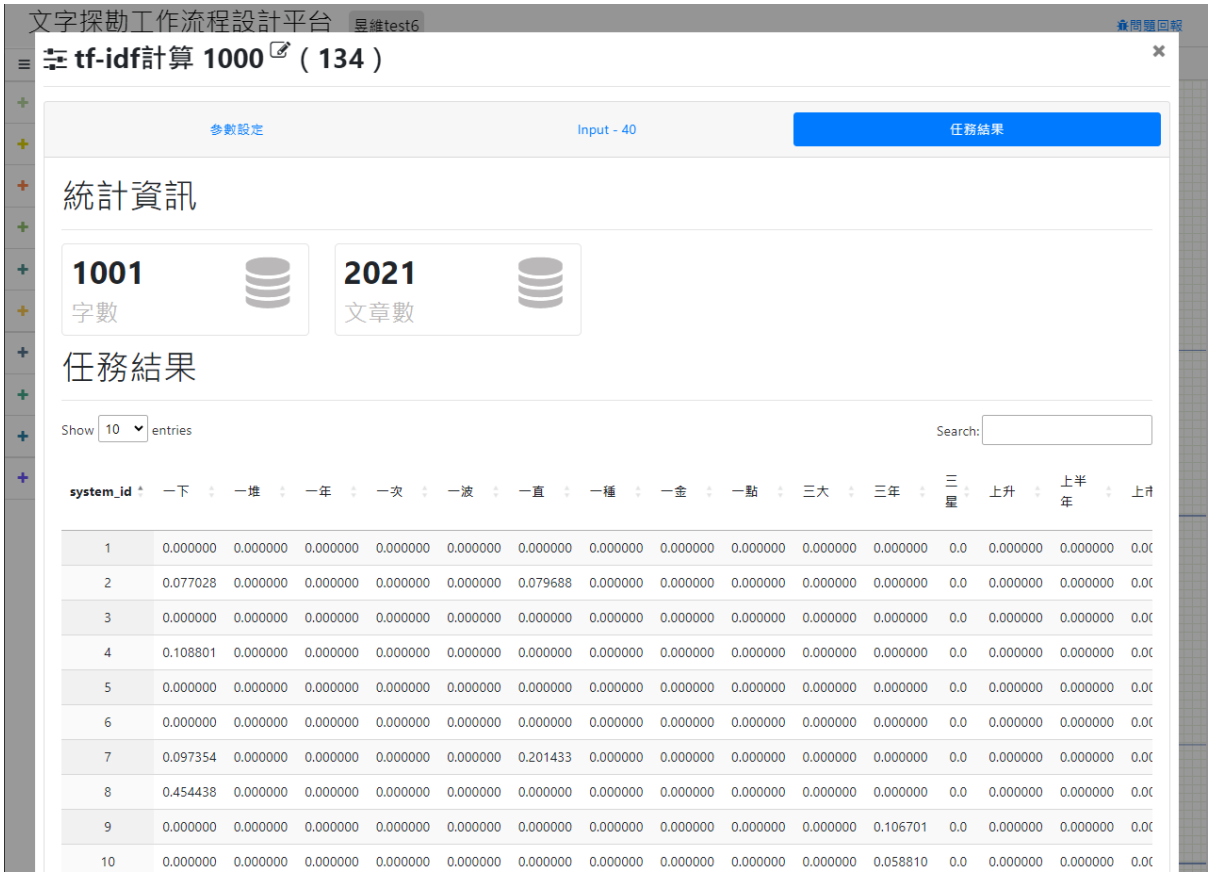
我們要去把相同的字放在一起，並做一定的分類分群，下圖結果會將不同顏色代表不同分群，相比傳統文字雲，更傾向相同類別的字會放在一起


台積電
全球
晶片
日本
客戶
製程
奈米
先進
設備
生產
台北
世界
成為
繼續
漲幅
上漲
大漲
亞洲
進入
決定
綜合
聯電
英特爾
晶圓代工
三星
韓國
預計
產能
將在
擴大
規劃
布局
法說
量產
先前
計算
比重
董事長
宣布
緯創
傳出
尾牙
日前
退休
消息
利多
使宜
旺宏
漲價
相比
調漲
貿易
地區
安全
經濟日報
即時
健身
課程
費用
最大
魔達
努力
時代
謝金河
黃仁
財經
伺服器
輝達
英業達
超微
事業
人才
大學
海外
製造商
教育
老師
學生
畢業
越來越
也將
工廠
台積
之間
建設
日圓
熊本
附近
業界
成熟
評估
收購
面臨
挑戰
壓力
加碼
經濟部
高雄
完整
神盾
滿足
算能夠

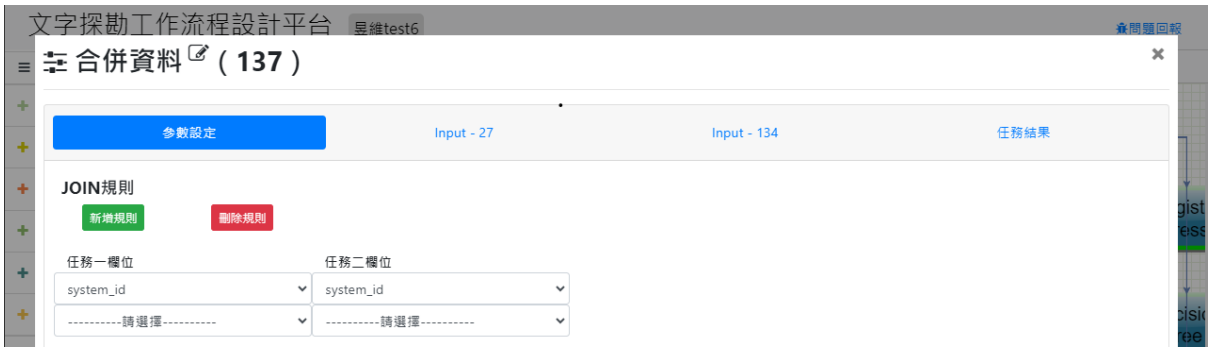

訂單
台廠
聯發科
全新
這波
定期
個股
定額
大摩
世芯
主力
鴻海
通常
通知
收到
槓桿
利息
國泰
元大
富邦
美信
超過
買超
賣超
排名
群益
上述
中信
國泰金
富邦金
永豐金
中信金
日月光
投控
南亞科
統一
宏碁
華城
大聯大
文曄
智原
佳世達
華通
力成
收盤價
傳統

分類器

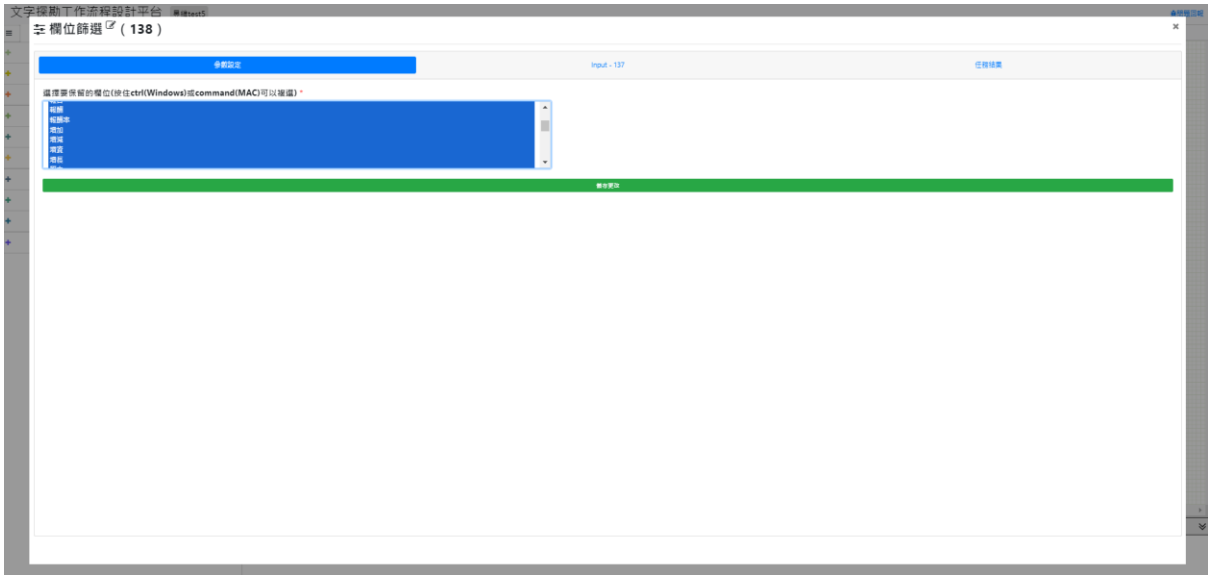
我們使用 tf-idf，用來評估一字詞對於一個文件集重要程度，字詞的重要性隨著它在文件中出現的次數成正比增加，但同時會隨著它在語料庫中出現的頻率成反比下降，我們選用 1000 個不同詞彙。

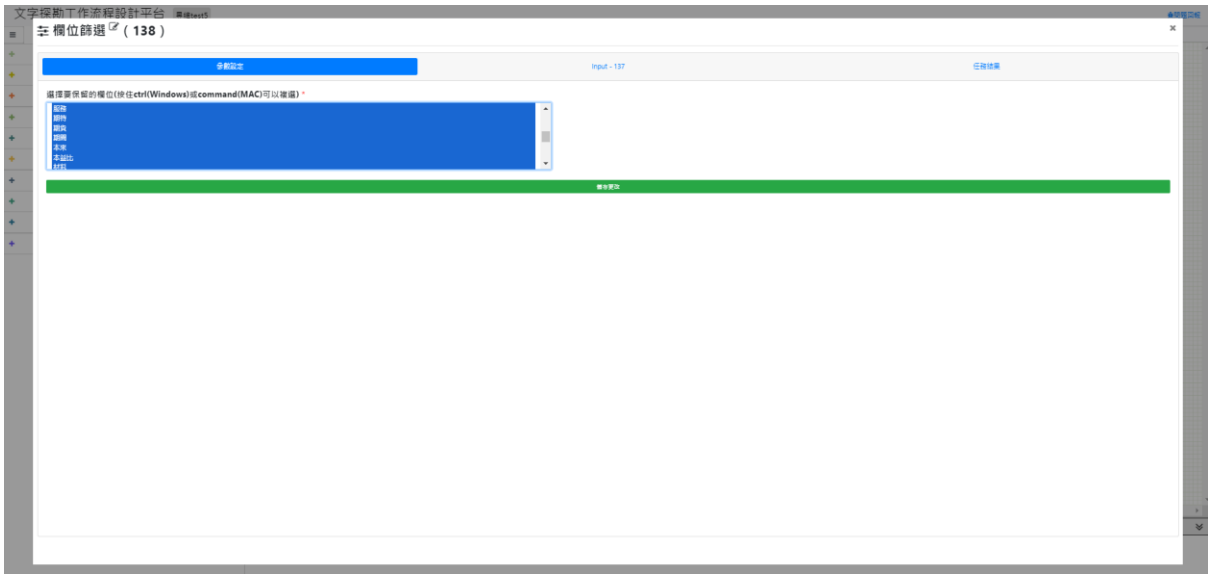


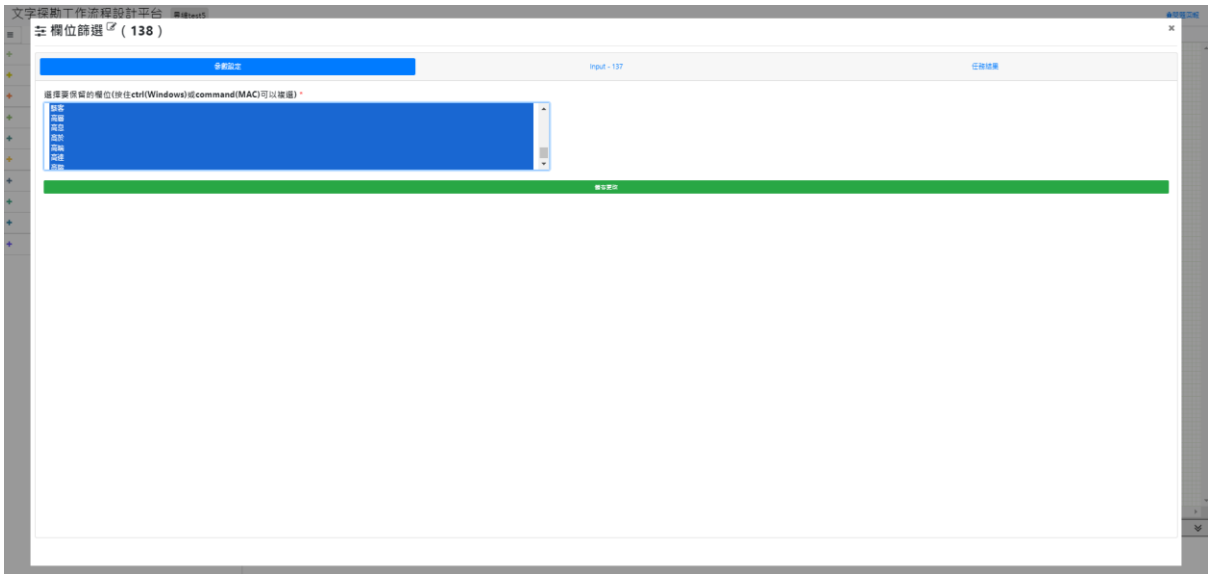
我們將原始資料與 TF-idf 跑出之特徵結果進行結合



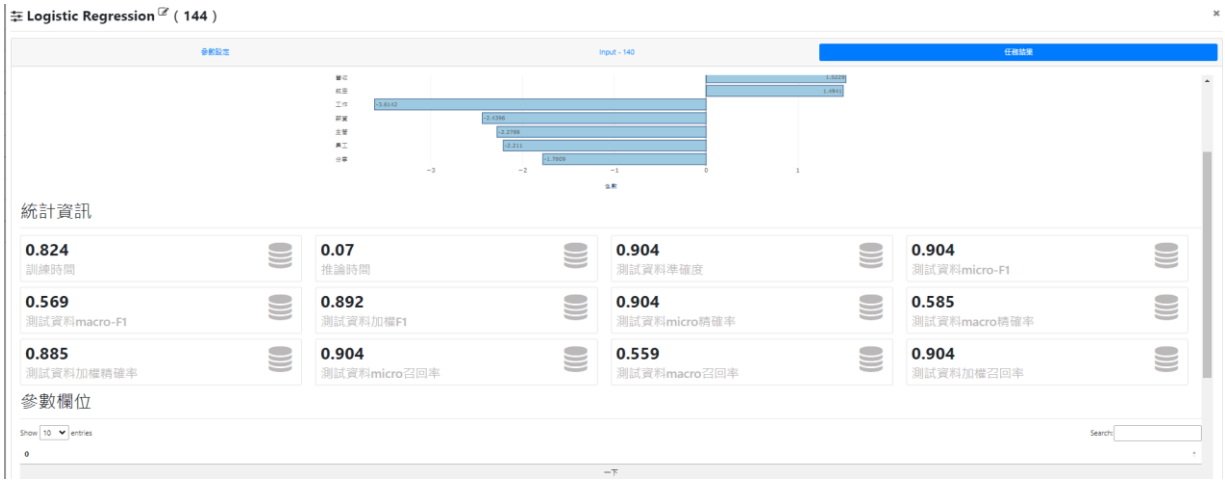
並利用欄位篩選僅保留 System_id、ArtCcatogory、其他的保留字詞與 TF-idf 跑出之特徵


文字探勘工作流程設計平台
欄位篩選（138）
參數設定
Input - 137
任務結果
選擇要保留的欄位(按住ctrl(Windows)或command(MAC)可以複選) *

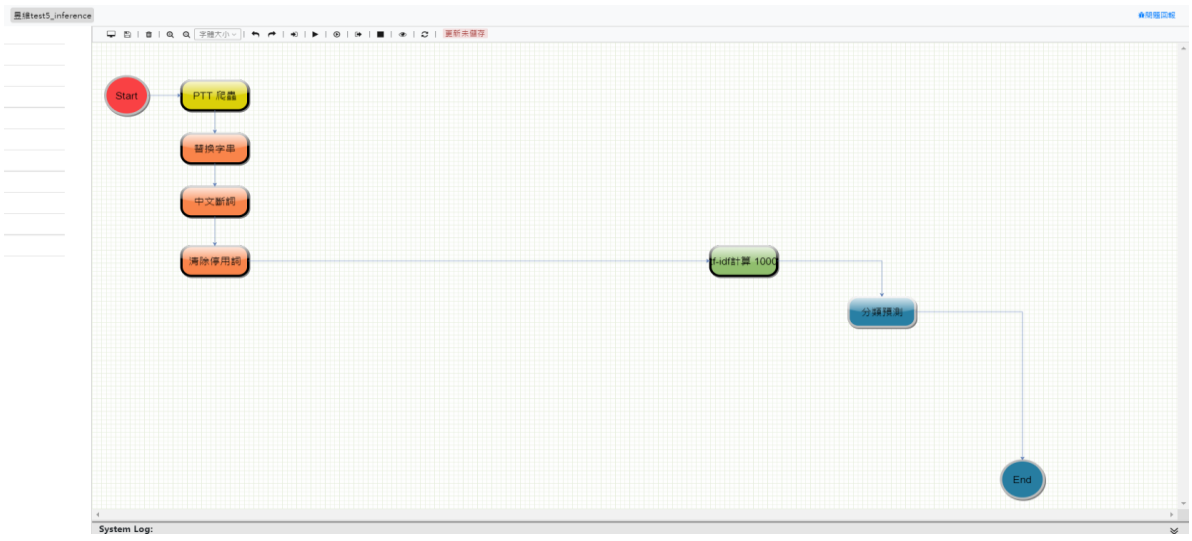

文字探勘工作流程設計平台
欄位篩選（138）
參數設定
Input - 137
任務結果
選擇要保留的欄位(按住ctrl(Windows)或command(MAC)可以複選) *

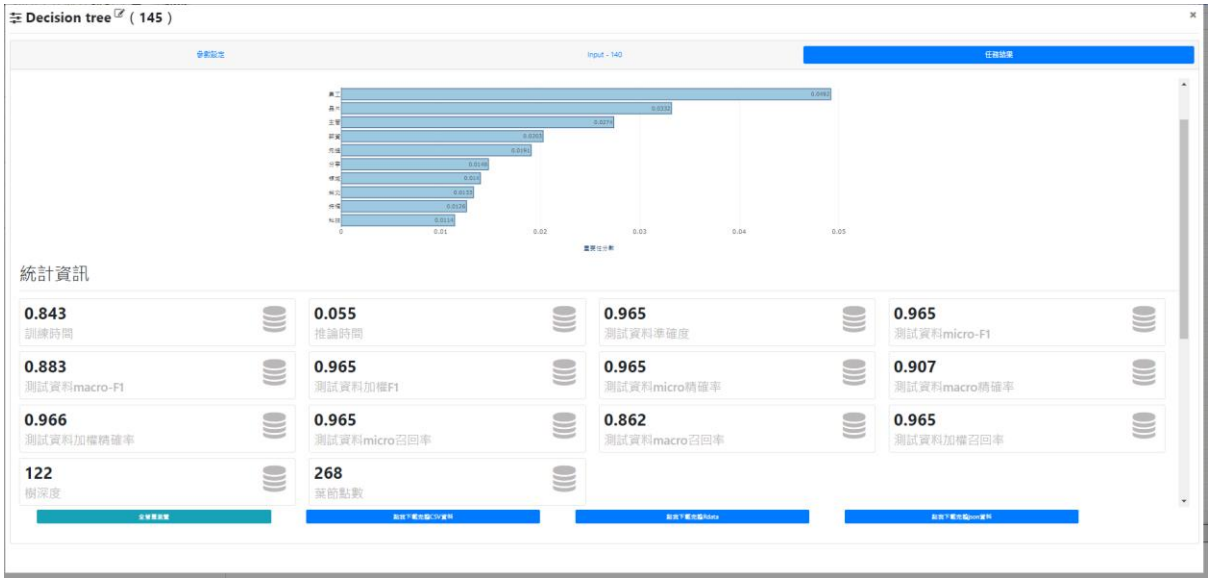

文字探勘工作流程設計平台
欄位篩選（138）
參數設定
Input - 137
任務結果
選擇要保留的欄位(按住ctrl(Windows)或command(MAC)可以複選) *

我們依目標欄位去選擇切割比例，因原始資料是依序撈取，故將資料打散，並隨機抽取 20%的資料

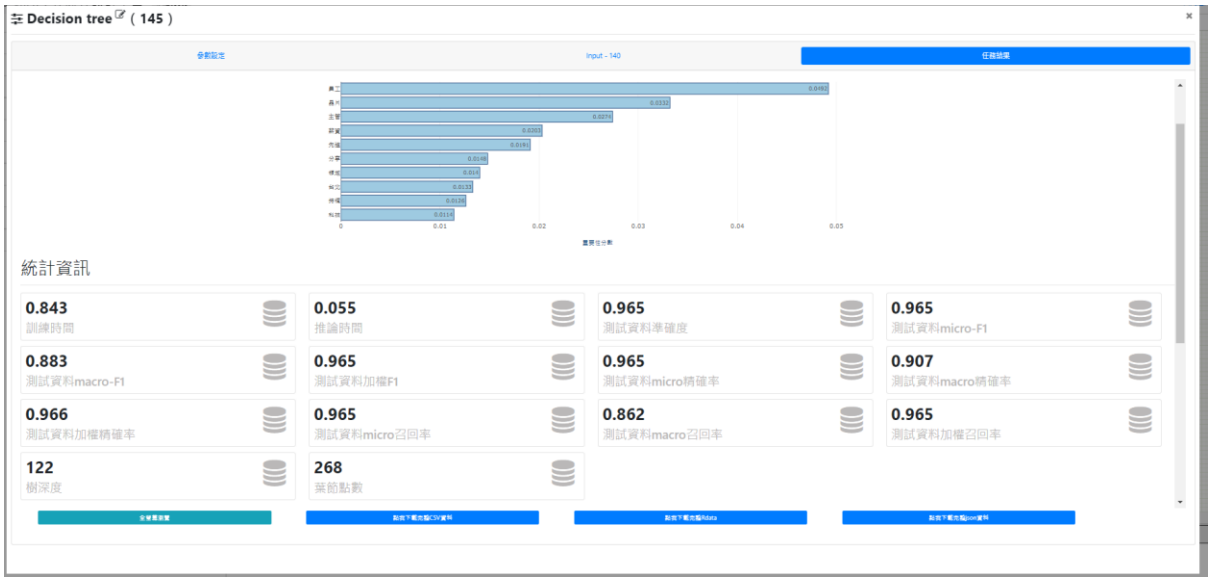


在 Logistic Regression 任務結果中，產出 inference workflow 的分類預測圖表功能

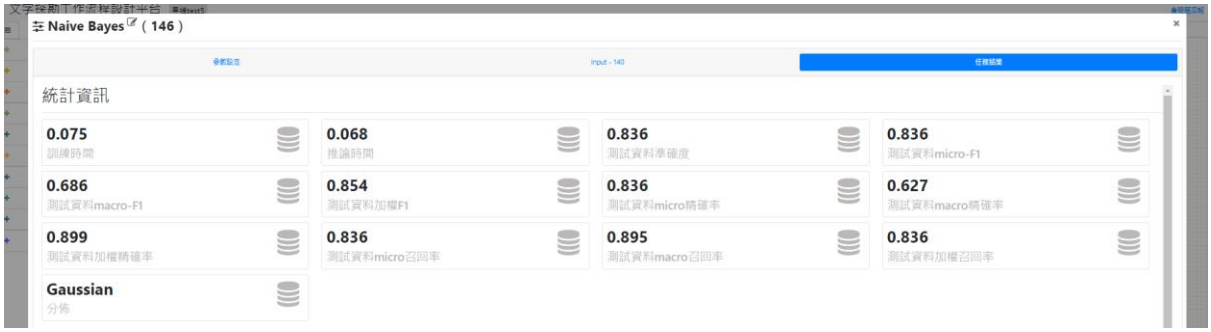




決策樹



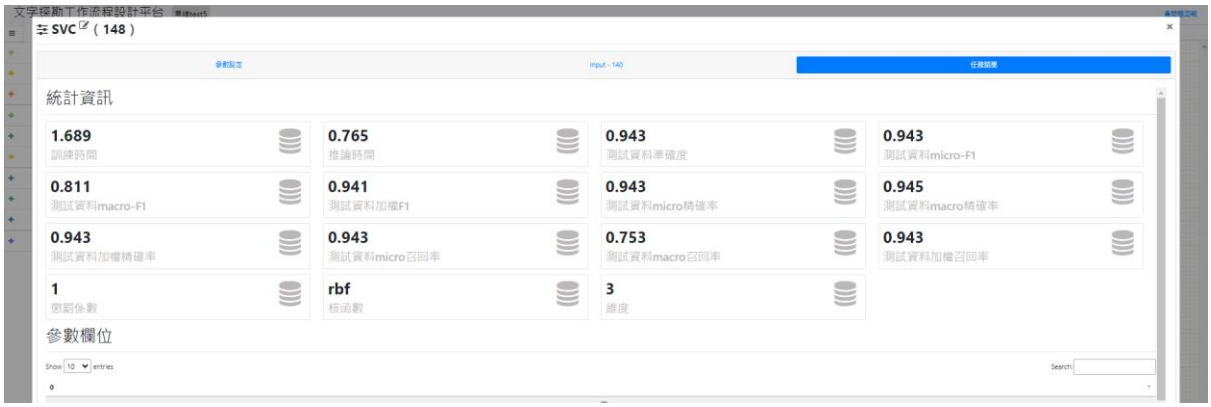
Navie Bayes



KNN


KNN（147）
統計資訊
0.033
訓練時間
0.346
推論時間
0.847
測試資料準確度
0.847
測試資料micro-F1
0.49
測試資料macro-F1
0.825
測試資料加權F1
0.847
測試資料micro精確率
0.551
測試資料macro精確率
0.831
測試資料加權精確率
0.847
測試資料micro召回率
0.468
測試資料macro召回率
0.847
測試資料加權召回率
5
鄰居數
euclidean
距離公式

SVC


SVC（148）
統計資訊
1.689
訓練時間
0.765
推論時間
0.943
測試資料準確度
0.943
測試資料micro-F1
0.811
測試資料macro-F1
0.941
測試資料加權F1
0.943
測試資料micro精確率
0.945
測試資料macro精確率
0.943
測試資料加權精確率
0.943
測試資料micro召回率
0.753
測試資料macro召回率
0.943
測試資料加權召回率
1
懲罰係數
rbf
核函數
3
維度
參數欄位