特斯拉新聞主題分類

讀書會 第二次報告

第十二組
N124020003 林廷祐
N124020009 陳俐錚
N124020013 胡馨予
N124020002 陳威良
N124020008 朱浚騰
N124020010 張世安

資料來源

資料範圍：

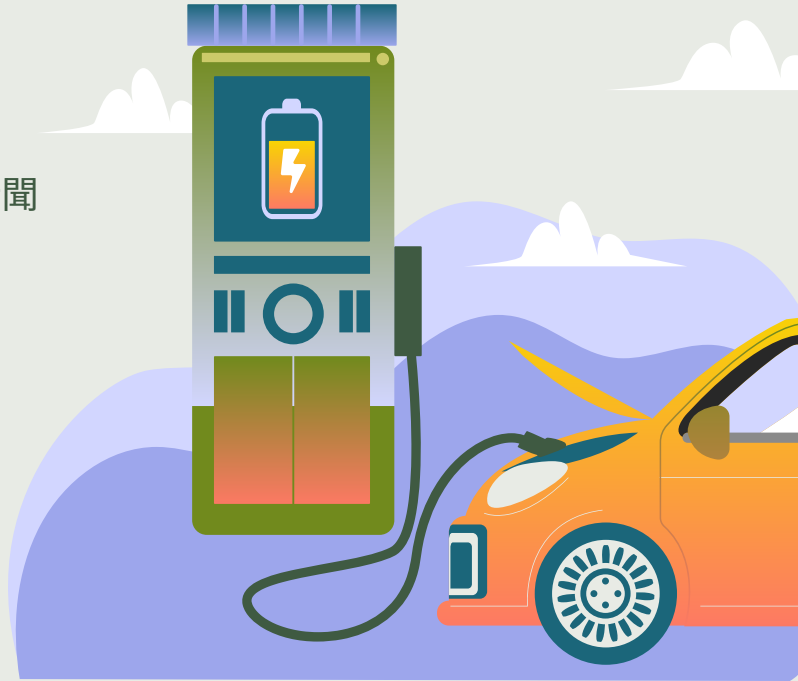
抓取東森新聞的[國際]、[汽車]、[財經]三的分類的新聞

資料時間：

2019/01/01 ~ 2025/03/31

關鍵字：

特斯拉



資料來源

分類	條數
國際	218
汽車	73
財經	144

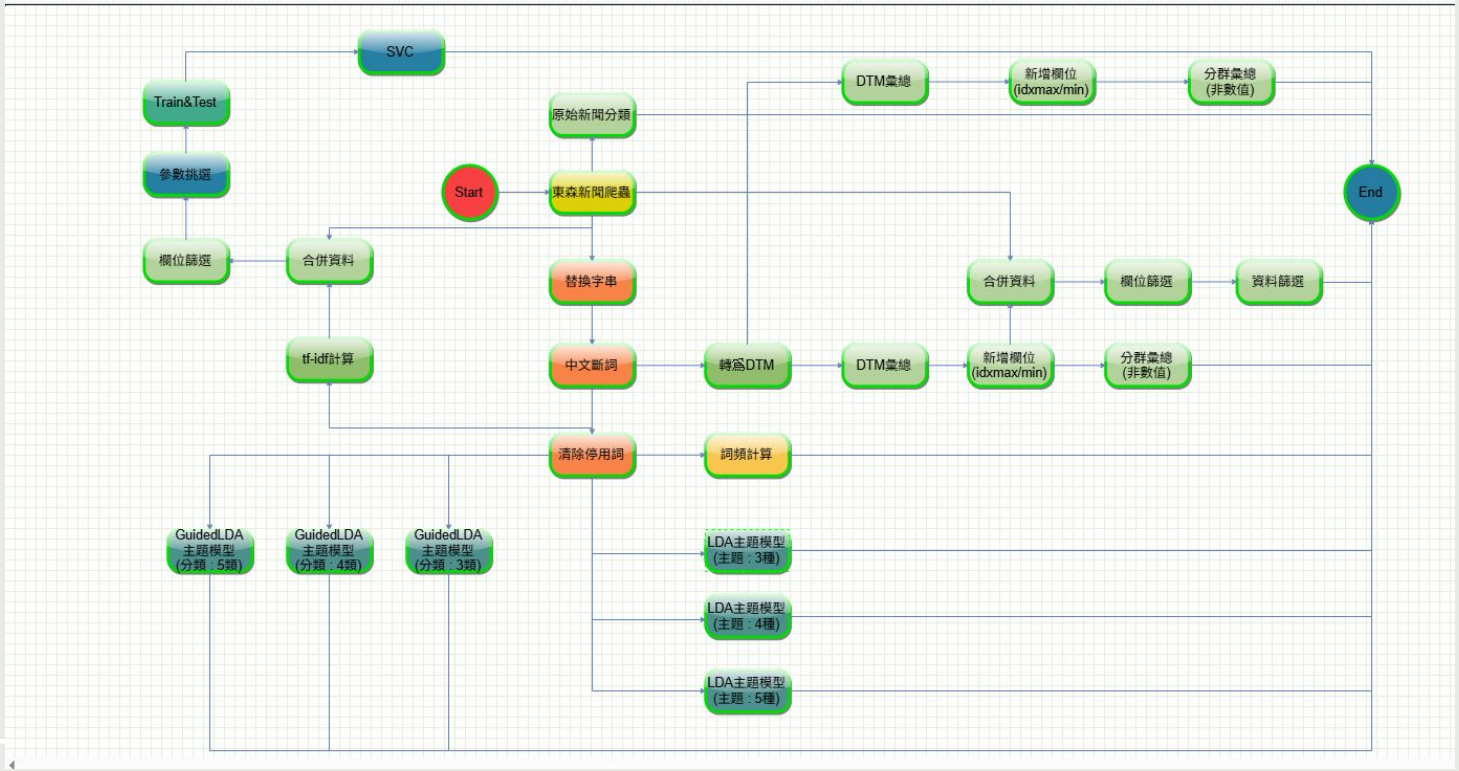
研究目的

1. 分析特斯拉新聞在財經、汽車與國際三大類別中的分類挑戰，並評估不同文本分析方法的適用性
2. 比較監督式與非監督式學習方法（SVC+DTM、LDA、Guided LDA）在特斯拉新聞分類中的效果差異
3. 探索東森新聞（2019-2025年）中特斯拉報導的潛在主題結構，揭示跨類別報導的內容特徵

研究方法

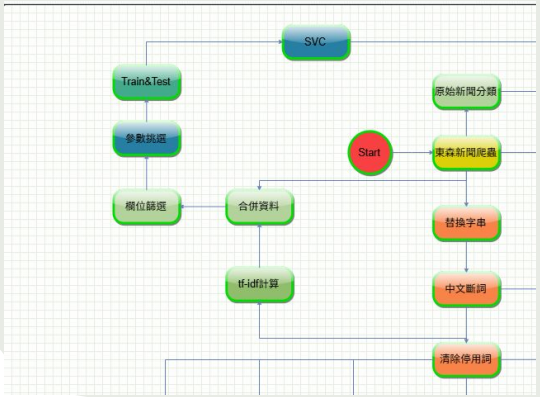
- 使用SVC分類器結合DTM分析特斯拉新聞的詞彙特徵，並與原始分類進行比較
- 潛在狄利克雷分配（LDA）主題模型：應用無監督學習方法，自動發現特斯拉新聞中的潛在主題結構，探索原始類別無法捕捉的報導面向
- 引導式LDA（Guided LDA）分析：結合專家先驗知識和數據驅動的主題發現，評估其在提高主題離散性和減少混淆方面的優勢
- 人工輔助評估：結合DTM結果與人工審核，建立更平衡的分類基準，解決特斯拉新聞跨領域特性帶來的自動分類困難
- 分類效果評估：透過主題離散性指標、混淆度分析及類別分布比較，全面評估各方法在特斯拉新聞分類中的優劣勢

使用Tarflow平台



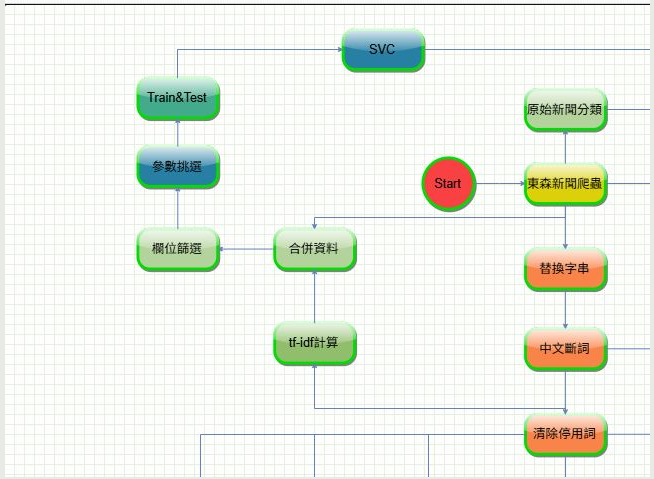
SVC+DTM分類


Train&Test（111）
參數設定
Input - 109
任務結果
目標欄位 *
artCatagory
測試資料切割比率 *
0.2
是否隨機排序資料
是
亂數種子
1314
儲存更改


SVC
Train&Test
參數挑選
欄位篩選
合併資料
tf-idf計算
清除停用詞
Start
原始新聞分類
東森新聞爬蟲
替換字串
中文斷詞


SVC（114）
參數設定
Input - 111
任務結果
統計資訊
0.007
訓練時間
0.009
推論時間
0.644
測試資料準確度
0.644
測試資料micro-F1
0.613
測試資料macro-F1
0.622
測試資料加權F1
0.644
測試資料micro精確率
0.73
測試資料macro精確率
0.725
測試資料加權精確率
0.644
測試資料micro召回率
0.608
測試資料macro召回率
0.644
測試資料加權召回率
1
懲罰係數
rbf
核函數
3
維度

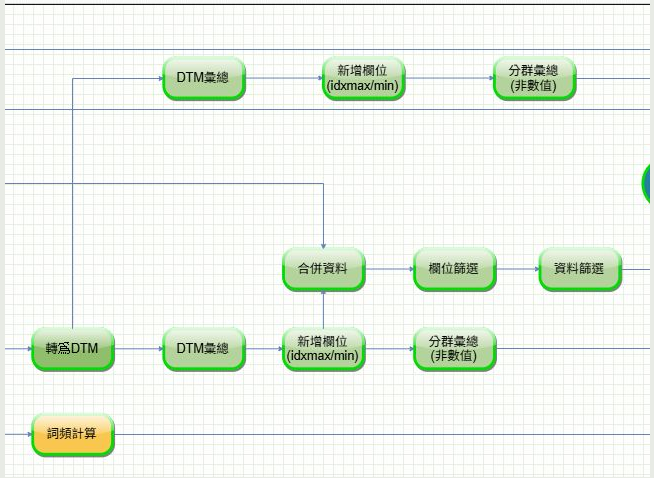
SVC+DTM分類




```
name,class,alias
台積電,財經,台積電
台股,財經,台股
新台幣,財經,新台幣
股價,財經,股價
股市,財經,股市
股票,財經,股票
功能,汽車,功能
續航,汽車,續航
車型,汽車,車型
體驗,汽車,體驗
國家,國際,國家
政府,國際,政府
總統,國際,總統
美國,國際,美國
馬斯克,國際,馬斯克
```

分類	條數
國際	251 (+33)
汽車	47 (-26)
財經	137 (-7)

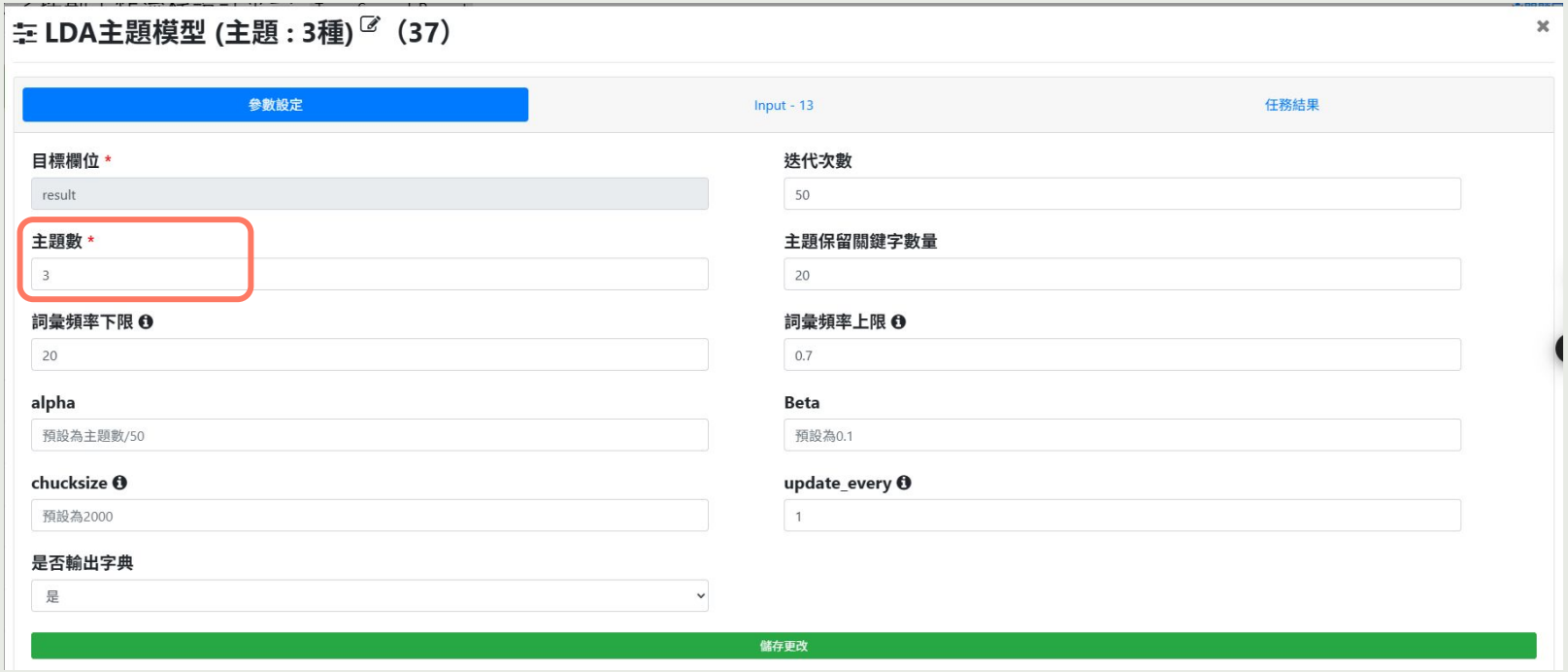
DTM+人工分類



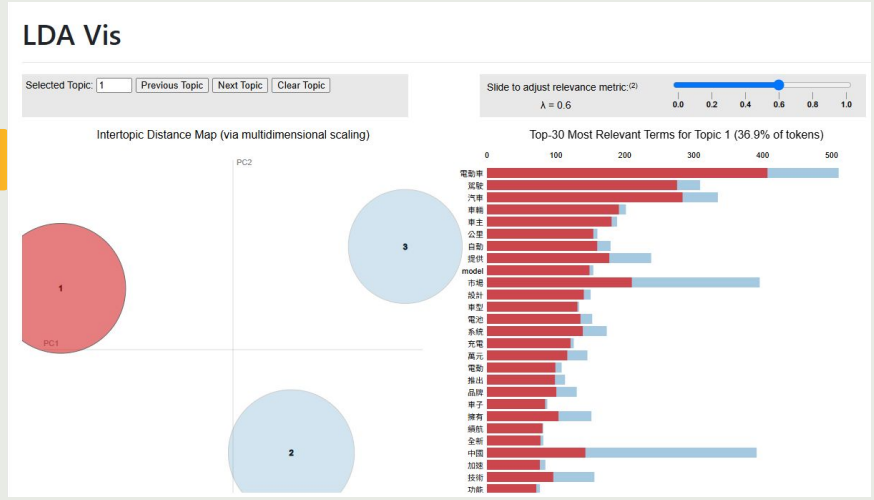

```
name,class,alias
10,汽車,10
100,財經,100
11,汽車,11
12,汽車,12
20,汽車,20
2024,汽車,2024
30,財經,30
500,財經,500
ai,汽車,ai
elonmusk,國際,elonmusk
model,汽車,model
spacex,國際,spacex
tesla,汽車,tesla
一個,國際,一個
一名,國際,一名
上漲,財經,上漲
不少,國際,不少
不斷,國際,不斷
不過,國際,不過
世界,國際,世界
中國,國際,中國
主要,汽車,主要
之後,國際,之後
今年,國際,今年
他們,國際,他們
以來,國際,以來
...
```

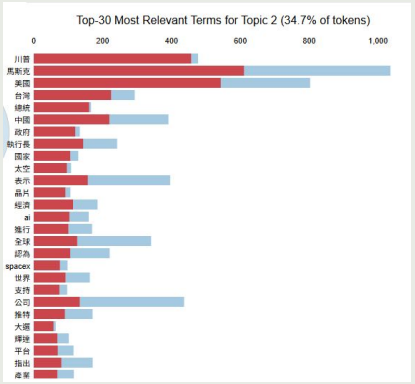
分類	條數
國際	186 (-32)
汽車	132 (+59)
財經	117 (-27)

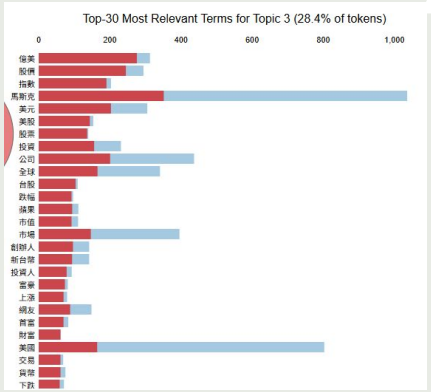
使用LDA主題模型


LDA主題模型 (主題：3種)（37）
參數設定
Input - 13
任務結果
目標欄位 *
result
迭代次數
50
主題數 *
3
主題保留關鍵字數量
20
詞彙頻率下限
20
詞彙頻率上限
0.7
alpha
預設為主題數/50
Beta
預設為0.1
chucksize
預設為2000
update_every
1
是否輸出字典
是
儲存更改

使用LDA主題模型





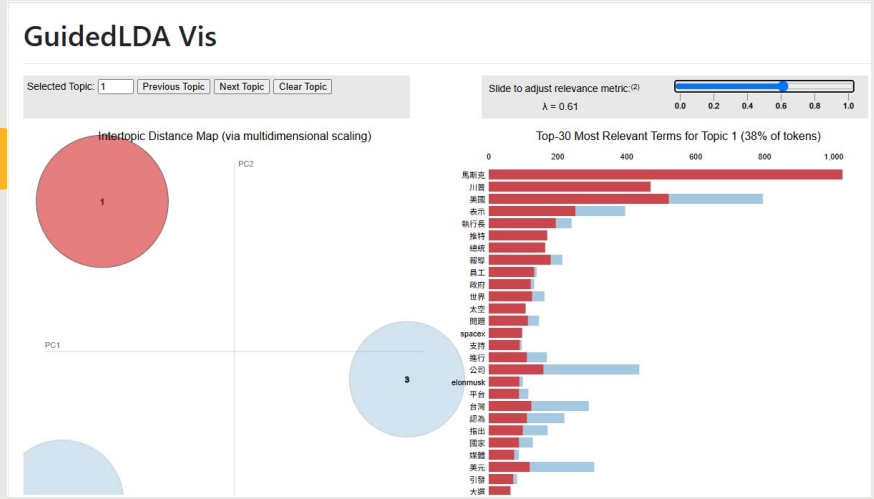


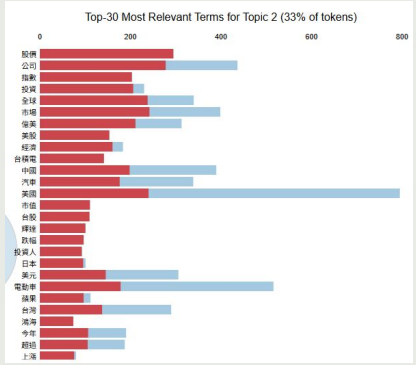
統計資訊

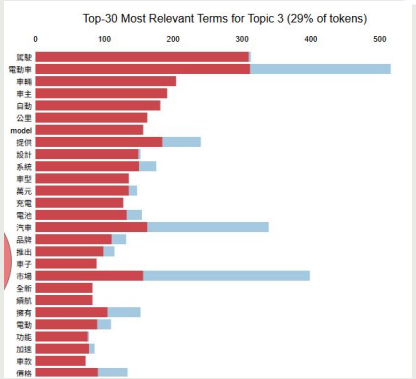
60 字數	**3** 主題數	**-1.506** 主題連貫性(UMass)	**-0.529** 主題連貫性(PMI)
0.478 主題連貫性(Cv)	**370.20** 混淆度		

使用GuildedLDA主題模型

使用GuildedLDA主題模型







統計資訊

60 字數	**3** 主題數
-1.341 主題連貫性(UMass)	**-0.226** 主題連貫性(PMI)
0.519 主題連貫性(Cv)	**465.91** 混淆度

結論

整理結論

- **分類偏差傾向：**原始分類和SVC+DTM都傾向將特斯拉新聞歸類為國際新聞，而DTM+人工和主題模型方法則識別出更多汽車相關內容
- **分類方法差異：**人工干預和先驗知識能顯著影響分類結果，表明特斯拉新聞具有跨領域特性，難以僅通過自動方法準確分類
- **最佳實踐建議：**考慮到各方法的優缺點，建議採用GuidedLDA作為主要分類方法，但需謹慎選擇種子詞彙；同時結合DTM+人工審核可獲得更平衡的類別分布