產業新聞分析

資管三 黃子維 B11705016

1. 產業2-gram 20個keywords:

|  |  |  |  |
| --- | --- | --- | --- |
| * 1. 捷流 | * 1. 志聖 | * 1. 慧衣 | * 1. 智先 |
| * 1. 寬庭 | * 1. 永冠 | * 1. 越鋼 | * 1. 精測 |
| * 1. 光洋 | * 1. 味全 | * 1. 王品 | * 1. 臺銀 |
| * 1. 虎航 | * 1. 瑞軒 | * 1. 徐世 | * 1. 中租 |
| * 1. 誠品 | * 1. 友訊 | * 1. 台苯 | * 1. 廢鋼 |

* **Keywords應用**：分析產業趨勢、財經分析、投資預測、品牌與消費市場分析、輿情監測。
* **挑選方法**：首先我先觀察每個統計量排序出來的關鍵字，由於每個統計量的意義不同，比如tf-idf可以看到區分性高的關鍵字、卡方值能夠檢測異常及突出的詞彙，在產業的資料集裡tf-idf、全部tf-idf、MI和Lift是表現得較好的篩選方式，將個統計量前30個顯著的keywords計算他們的出現次數，取前20個當作這個資料的keywords，若有相同出現次數者則以tf-idf排名當作比較條件。
* **理由**：結合多種統計方法再進行篩選避免偏頗、平手的比較條件設為tf-idf是希望可以確保關鍵字的區分性。另外，可以在查看各統計量排序的結果中發現Lift和MI會出現非完整詞彙也無意義的關鍵字（光昨、海昨），可能是因為MI偏好「彼此共現機率高，但單獨出現機率低」的詞，Lift對「少見但出現時經常一起出現的詞」比較敏感，導致這樣的詞彙排名偏高。

1. 產業3-gram 20個keywords:

|  |  |  |  |
| --- | --- | --- | --- |
| * 1. 光洋科 | * 1. 徐世昌 | * 1. 台數科 | * 1. 環球晶 |
| * 1. 東森國 | * 1. 羅智先 | * 1. 東森電 | * 1. 智慧衣 |
| * 1. 林文淵 | * 1. 房養老 | * 1. 區塊鏈 | * 1. 專利申 |
| * 1. O奇摩 | * 1. 李瑞倉 | * 1. 張國煒 | * 1. 段行建 |
| * 1. 諾基亞 | * 1. 灣虎航 | * 1. 李紀珠 | * 1. 富智康 |

* **Keywords應用**：**產業分析、投資分、技術趨勢預測、行銷策略擬定、商業情報分析。**
* **挑選方法**：3-gram的部分我使用全部tf-idf作為關鍵字的挑選依據，再將明顯重複的詞彙（如：任教師和兼任教、東森電和森電視）刪除選出前20個詞彙當作keywords。
* **理由**：在3-gram的資料裡，會發現大部分的統計值做排序後的結果會出現像2-gram類似的問題，且情況較為嚴重，會讓罕見但同時出現的詞彙被異常放大導致統計值更極端，故

1. 鴻海2-gram 20個keywords:

|  |  |  |  |
| --- | --- | --- | --- |
| * 1. 權證 | * 1. 專利 | * 1. 印度 | * 1. 申請 |
| * 1. 認購 | * 1. 電信 | * 1. 賣超 | * 1. 家電 |
| * 1. 機器 | * 1. 同期 | * 1. 組裝 | * 1. 買超 |
| * 1. 出貨 | * 1. 三星 | * 1. 業者 | * 1. 工作 |
| * 1. 電腦 | * 1. 股票 | * 1. 發表 | * 1. 通訊 |

* **Keywords應用**：投資分析、供應鏈分析、產業趨勢預測、股票市場分析、專利與技術發展。
* **挑選方法**：先觀察每個統計量的排序結果，最後選用tf-idf和全部tf-idf取出前30個keywords，仿照產業2gram的方法先統計每個詞彙的出現次數，再以tf-idf作為平手抉擇條件，刪除無意義的詞彙（昨天、減少）作為最後的20個keywords。
* **理由**：在鴻海的資料集裡，鴻海相關的詞彙出現次數一定是非常多的（鴻海、夏普）等等相關詞彙，所以應該以特殊性和區分性作爲主要篩選原則，可以發現只有tf-idf有表現出篩選到不具區分性詞彙的作用，故選擇tf-idf作為主要條選的條件。

1. 鴻海3-gram 20個keywords:

|  |  |  |  |
| --- | --- | --- | --- |
| * 1. 諾基亞 | * 1. 富智康 | * 1. 呂芳銘 | * 1. 基金會 |
| * 1. 大數據 | * 1. 出貨量 | * 1. 自動化 | * 1. 太陽能 |
| * 1. 人民幣 | * 1. 面板廠 | * 1. 台指期 | * 1. 機器人 |
| * 1. 聯準會 | * 1. 戴正吳 | * 1. 日月光 | * 1. 智慧型 |
| * 1. 股東會 | * 1. 富士康 | * 1. 供應商 | * 1. 物聯網 |

* **Keywords應用**：產業發展、技術趨勢、供應鏈分析、投資決策、企業管理、決策制定。
* **挑選方法**：採用和鴻海2gram相同的做法，先觀察每個統計量的排序結果，最後選用tf-idf和全部tf-idf取出前30個keywords，仿照產業2gram的方法先統計每個詞彙的出現次數，再以tf-idf作為平手抉擇條件，刪除無意義的詞彙（昨天、減少）作為最後的20個keywords。
* **理由**：和2gram相同。