運動與財經新聞

趨勢追蹤分析



應統職二 408145137 駱俊宏;應統碩一 410336130 陳冠廷

目錄

1. 資料背景： 1

1) 資料來源的取得條件： 1

2) 資料來源的取得方法： 1

2. 資料探索： 2

1) 資料集說明-原始資料： 2

2) 資料集個別變數敘述性統計 2

3. 資料預處理： 4

1) 資料清理: 4

i. 資料不純度定義調整： 4

ii. 資料清理條件調整： 5

2) 語言處理(Liguistic Processing) 6

4. 資料探勘分析： 7

1) 詞袋模型: 7

i. 計算詞彙矩陣： 7

ii. 內文相似度分析 8

2) TF-IDF模型： 9

3) K-means分析 10

5. 結論 11

1) 資料原始來源 11

2) 資料主題分析 11

## 資料背景：

自從2021年中台灣在東京奧運有了近幾十年來最好的收獲後，除了台灣典型最受歡迎運動-棒球或籃球，我們可以發現其他的單項運動選手也開始在社群媒體或者公關活動開始大為活躍，那除了短暫的風潮外，本組試著由台灣的財經新聞收集來了解最近12個月以來在台灣運動與財經新聞的趨勢。

### 資料來源的取得條件：

透過搜尋引擎Google，給予財經、運動 2個關鍵字，並且依照資料背景設定台灣在最近12個月的與運動有關財經新聞，增加更為明確的搜尋條件，選擇新聞、搜尋繁體中文網頁、過去一年的新聞與依照關聯性排序來取得我們所需要的財經新聞。



### 資料來源的取得方法：

考量人工作業尚可負擔報告需求文章數量，所以採用人工逐頁面與判斷符合財經新聞資料，人工整理出新聞標題、資料來源、作者、日期、網址與內文等6個欄位並以xlsx格式儲存，其中日期先以人工修正為相同格式。

## 資料探索：

### 資料集說明-原始資料：

由300個搜尋結果中，人工挑選出內容與運動、財經有關的文章共53篇。

|  |  |  |  |
| --- | --- | --- | --- |
| 欄位編號 | 中文說明 | 英文說明 | 說明 |
| 1 | 新聞標題 | Title | 主要標題 |
| 2 | 資料來源 | Source | 新聞媒體名稱 |
| 3 | 作者 | Author | 記者或撰稿人名稱 |
| 4 | 日期 | Time | 各種標準日期格式 |
| 5 | 網址 | Link | 原文網址 |
| 6 | 內文 | Context | 原文內容，不包含圖片、聲音與影像 |

### 資料集個別變數敘述性統計

1. 變數[新聞標題]與[內文]字數分析：

由於某些報導偏向於專題報導，所以有較長的文字內容，不過所擷取的資料中就文章長度大致上符合一般的新聞報導長度。

|  |  |  |
| --- | --- | --- |
| 單位:字元數 | 標題(Title) | 內文(Context) |
| 平均值 | 24.5283 | 984 |
| 標準差 | 9.55 | 669.24 |
| 最小值 | 10 | 287 |
| 25% | 18 | 633 |
| 50% | 23 | 770 |
| 75% | 27 | 1,054 |
| 最大值 | 51 | 3,363 |

1. 日期變數:新聞

檢驗後與在Google搜尋中的日期區間條件相符，資料區間開始於[2021-08-03],結束於 [2022-06-09]，我們可以由下表中得知每月文章的數目，基本上是報導搜尋結果日期越接近取得報導數量較多，而[2021-07]與[2021-09]2個月分沒有合乎條件的新聞。



新聞年月次數分配表

1. 類別變數[來源]與[作者]交叉分析表：

我們由主要資訊中可以了解，資料來源以自由時報(15篇)、工商時報(9篇)與ETtoday新聞雲(7篇)為排名前3的資料來源，而由作者與資料來源交叉分析可以看到自由時報與ETtoday新聞雲相對上有專門負責記者的可能性。



主要資訊(總表紅框)



原始總表

## 資料預處理：

### 資料清理:

關於資料清理，因應資料內容為中文，我們進行以下的調整：

(1)資料不純度定義調整

(2)資料清理程式修改

### 資料不純度定義調整：

本次資料來源為中文報導，由於中文報導有時會使用全形的符號，故進行資料純度評估時，增加關於全型標點符號的不純度定義來接近相對真實的不純度。

|  |  |
| --- | --- |
| 全型符號 | "！？｡＂＃＄％＆＇（）＊＋－／：；＜＝＞＠［＼］＾＿｀｛｜｝～｟｠｢｣､、〃》「」『』【】〔〕〖〗〘〙〚〛〜〝〞〟〰〾〿–—‘'‛“”„‟…‧﹏%()。， |

調整定後的不純度前3名資料如下圖，



### 資料清理條件調整：

除調整不純度定義外，資料清理條件基於本組要了解的是過去1年的台灣運動財經相關趨勢，故增加移除數字與全型符號規則，但保留英文詞彙，原因是經過複合清理條件的測試，避免移除關鍵的運動品牌或者特殊英文專有名詞，數字部分則是過程中未見到有意義的累積。

|  |  |
| --- | --- |
| 新增清除規則 | Python Code 示意 |
| 數字 | re.sub(r'\d+', ' ', text) |
| 全形符號 | re\_punctuation = 全形符號集合  re.sub(re\_punctuation, ' ', text) |

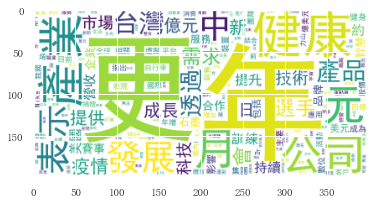
清理後的不純度皆為 0 

### 語言處理(Liguistic Processing)

採用jieba進行語言處理，斷詞後的第一次結果發現，”先前Google搜尋取得的報導，搜尋條件直接會使得條件名詞高頻出現”， 故修改stop words定義檔案，進行微幅修正。

|  |  |
| --- | --- |
| 比較結果 | 新增stop words [運動] |
| 修改前 |  |
| 修改後 |  |

文字雲並沒有明確關鍵字情形，接下來進行更進一步的文字探勘工作。



## 資料探勘分析：

### 詞袋模型:

### 計算詞彙矩陣：

提取文字特徵，經過詞頻計算後(將最小詞頻設為2)，總計有1606個詞組，長度介於2~9之間，其中長度>=5以上大多為英文單字



### 內文相似度分析

a. 以相似度門檻值 10%來觀察，大多數文章與其他文章皆有10篇以上相似度。b. 以高相似度觀察，取最高前5名為例，

|  |  |  |  |
| --- | --- | --- | --- |
| 文章索引 | 文章索引 | 相似度 | 主題 |
| 7 | 43 | 50.28% | 疫情上市櫃財經報導 |
| 26 | 39 | 50.13% | 數位化健康運動相關 |
| 14 | 28 | 48.86% | 運動與企業贊助 |
| 46 | 23 | 47.53% | 運動彩券相關 |
| 35 | 39 | 46.54% | 數位化健康運動相關 |

而排名第一的為疫情期間的上市櫃公司財經報導

|  |  |
| --- | --- |
| 文章索引 7 | 文章索引 43 |
| 馬來西亞、越南、印尼等東南亞地區新冠肺炎持續升溫，造成當地台資廠供應鏈停擺、運輸不順，也擾亂運動休閒概念股成長腳步，惟歐美經濟持續解封，相關產業業績多呈現成長，包括儒鴻（1476）等14檔上半年獲利強棒股，長線還有看頭。 運動休閒概念股已陸續公布季報，上半年每股稅後盈餘（EPS）逾2元，包括儒鴻、大田、巨大、桂盟、美利達、聚陽、明安、百和、豐泰、力山、岱宇、鈺齊-KY、鉅明、廣越等14檔。 其中，紡織成衣龍頭廠儒鴻上半年大賺10.51元，球頭代工廠大田、自行車大廠巨大上半年EPS則有9元以上，桂盟、美利達、聚陽及明安則有5元以上，表現相對突出。 群益投顧副總裁曾炎裕觀察，紡織成衣、製鞋、自行車族群、球頭代工等運動休閒概念股，皆受惠於疫後歐美經濟解封，消費需求回升，品牌經銷通路補庫存效應，使各家廠商的訂單無虞，能見度高。 但零件缺料、運輸缺櫃，以及供應鏈遭逢東南亞疫情逆風停工，則干擾相關族群營收成長動能，以致於近期股價出現修正，但法人近一周仍積極加碼巨大、美利達、明安及豐泰等個股，顯示依舊看好相關個股的後市。 以自行車雙雄為例，….. | 歐美地區經濟重啟帶動強勁需求，加上供應鏈問題可望緩解，以及大者恆大效應，法人看好，國內製鞋成衣、自行車等運動休閒概念股明年成長動能無虞，首選儒鴻（1476）、豐泰、巨大、桂盟等四檔。 受到缺櫃、塞港、東南亞疫情，以及財報等干擾因素影響，加上市場資金較為青睞元宇宙、低軌道衛星、電動車等作夢題材族群，運動休閒概念股近期走勢相對平淡，儒鴻、豐泰明顯拉回，但巨大、美利達、桂盟、聚陽等低檔有撐。 法人分析，2021年新冠肺炎擾亂東南亞供應鏈，使缺料狀況加劇，加上短期缺櫃、塞港事件，歐美地區鞋子、紡織品庫存仍維持在2013年低檔水準，自行車終端通路庫存已有一年半時間僅維持二至三周。 歐美國家自5月陸續解封後，需求迅速回升，尤以Nike、Lululemon、Adidas、UA等大型運動品牌最為強勁。法人認為，鑒於供應鏈大者恆大、高技術門檻、原物料及運輸等成本影響有限、越南勞工成本維持低檔、新台幣匯率升幅有限等有利條件，看好儒鴻、豐泰受惠最大。 自行車方面，雖然各國疫苗覆蓋率提升，中低階車種需.... |

### TF-IDF模型：

經由TF-IDF調整後的詞頻處理與BOW相似度結果比較，發現不同程度下降：

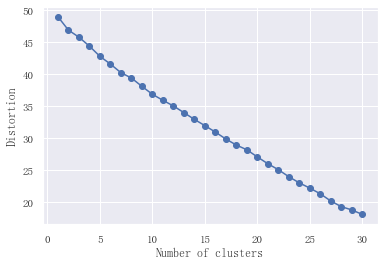
|  |  |  |  |
| --- | --- | --- | --- |
| 文章索引 | 文章索引 | 相似度 | TF-IDF相似度 |
| 7 | 43 | 50.28% | 48.60% |
| 26 | 39 | 50.13% | **30.90%** |
| 14 | 28 | 48.86% | 43.14% |
| 46 | 23 | 47.53% | 46.29% |
| 35 | 39 | 46.54% | **30.31%** |

從文字雲中發現本次的運動財經產業報導，部分關鍵字有更明顯的揭露，例如 產品、自行車、NFT、Nike、疫情、運彩、數位、運動鞋、健康

|  |
| --- |
|  |
| TF-IDF 改良後 |
|  |
| TF-IDF 改良前 |

### K-means分析

尋找最佳分群，本次的資料經過分群測試，並未找到較佳的分群結果，



在經過分群 3, 4, 5，的人工分析後，我們認為5個分群是這次分析結果比較能夠接受的結果。

|  |  |
| --- | --- |
| 分群 | 描述 |
| 1(g0) | 15篇文章，關於數位運動與健康議題 |
| 2(g1) | 3篇文章運動彩券法規相關 |
| 3(g2) | 12篇文章，運動選手贊助與運動醫學科技相關 |
| 4(g3) | 4篇文章，運動加密貨幣與NFT相關 |
| 5(g4) | 19篇文章，產業財經報導內容，關於銷售預估、獲利表現、製造等 |

## 結論

### 資料原始來源

本次透過google取得資料，驗證google最佳化搜尋結果後對資料來源的限制，-隱含google僅提供每頁10筆，最近30個頁面結果。建議可更換其他對使用者保持距離的搜尋引擎或者如果是要了解上市櫃公司與運動相關的財經新聞分析可透過專業財經網站(例如money dj)的已進行標籤化主題新聞來進行內容取得。

### 資料主題分析

就過去一年的主題分析結果來觀察，主要可以分成4個面向，第1個是台灣製造業的終端客戶是運動相關產業的報導，內容集中於業績分析報告，第2個是運動與數位活動的虛實結合，包含最新的NFT議題，第3部分則是運動行銷與贊助活動的相關企業報導，第4個則運動彩券相關法案的修正。