|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| 社群媒體第二次讀書會  **主題:PTT股版-存股討論情緒分析**  **第13組**  **龔敖摩N094220008、周冠呈N114320001、黃顗哲N114320007、**  **黃善羚N114320010、洪緯哲N114320013、李儼哲N114320016、**  **郭玟慧N114320024、王宥婷N114320025、凃沛瑩N114320029。** |
| |  |  |  | | --- | --- | --- | |  |  |  | |

目錄

主題:PTT股版-存股討論情緒分析

[壹、詞彙共現關係觀察 4](#_Toc163526307)

[一、 轉換成矩陣型態 4](#_Toc163526308)

[二、 計算共現矩陣 5](#_Toc163526309)

[貳、詞彙相關性觀察 7](#_Toc163526310)

[一、 計算相關性矩陣 7](#_Toc163526311)

[二、 單中心網絡圖 9](#_Toc163526312)

[三、 字詞網絡圖 10](#_Toc163526313)

[四、 關聯式文字雲 12](#_Toc163526314)

[參、加入字典後，觀察共現性 13](#_Toc163526315)

[一、 步驟說明 13](#_Toc163526316)

[二、 行業分析與公司概覽 15](#_Toc163526317)

[三、 共現性分析 16](#_Toc163526318)

[四、 單中心網路圖 16](#_Toc163526319)

主題:PTT推文(八卦、女孩、政黑)文件分類

[壹、文件分類 20](#_Toc163526321)

[一、工作流程平台建立 20](#_Toc163526322)

[二、資料前處理 20](#_Toc163526323)

[三、分類器 23](#_Toc163526324)

[四、最佳模型流程 32](#_Toc163526325)

[五、分類視覺化 33](#_Toc163526326)

**圖目錄**

[圖 1 DTM矩陣各詞出現次數 3](#_Toc163526823)

[圖 2 共現關係熱圖 5](#_Toc163526824)

[圖 3 共現關係DTM矩陣 5](#_Toc163526825)

[圖 4 相關性熱圖 7](#_Toc163526826)

[圖 5 相關性DTM矩陣 8](#_Toc163526827)

[圖 6 單中心網絡圖 9](#_Toc163526828)

[圖 7 字詞網絡圖 10](#_Toc163526829)

[圖 8 關聯式文字雲 12](#_Toc163526830)

[圖 9 替換字串 13](#_Toc163526831)

[圖10 保留重要字 13](#_Toc163526832)

[圖11 轉為DTM 13](#_Toc163526833)

[圖12 DTM矩陣各詞出現次數(字典) 14](#_Toc163526834)

[圖 13 共現性矩陣熱圖 15](#_Toc163526835)

[圖14共現性矩陣 15](#_Toc163526836)

[圖15 參數設定 16](#_Toc163526837)

[圖16 計算共現矩陣 17](#_Toc163526838)

[圖17 單中心網路圖(字典)1 17](#_Toc163526839)

[圖18 單中心網路圖(字典)2 18](#_Toc163526840)

[圖19 單中心網路圖(字典)3 18](#_Toc163526841)

[圖20 文件分類流程圖 20](#_Toc163526842)

[圖21 抓取PTT八卦版資料 20](#_Toc163526843)

[圖22 抓取PTT女孩版資料 21](#_Toc163526844)

[圖23 抓取PTT政黑版資料 21](#_Toc163526845)

[圖 24 透過CSV來上傳資料 21](#_Toc163526846)

[圖 25 Tf-idf詞彙數量選擇1000 22](#_Toc163526847)

[圖 26 合併資料後篩選欄位 22](#_Toc163526848)

[圖 27 80%訓練20%測試 22](#_Toc163526849)

[圖 28 迴歸分析 23](#_Toc163526850)

[圖 29 篩選1000個出現次數最多的資料(迴歸分析) 23](#_Toc163526851)

[圖 30先做參數挑選，再篩選出500個資料(迴歸分析) 23](#_Toc163526852)

[圖 31 篩選500個出現次數最多的資料(迴歸分析) 24](#_Toc163526853)

[圖 32 決策樹 24](#_Toc163526854)

[圖 33 篩選1000個出現次數最多的資料(決策樹) 25](#_Toc163526855)

[圖 34 先做參數挑選，再篩選出500個資料(決策樹) 25](#_Toc163526856)

[圖 35 篩選500個出現次數最多的資料(決策樹) 25](#_Toc163526857)

[圖 36 Naive Bayes 26](#_Toc163526858)

[圖 37 篩選1000個出現次數最多的資料(Naive Bayes) 26](#_Toc163526859)

[圖38先做參數挑選，再篩選出500個資料(Naive Bayes) 27](#_Toc163526860)

[圖39篩選500個出現次數最多的資料(Naive Bayes) 27](#_Toc163526861)

[圖40 KNN 28](#_Toc163526862)

[圖41 篩選1000個出現次數最多的資料(KNN) 28](#_Toc163526863)

[圖42 先做參數挑選，再篩選出500個資料(KNN) 28](#_Toc163526864)

[圖43 篩選500個出現次數最多的資料(KNN) 29](#_Toc163526865)

[圖 44 SVC 29](#_Toc163526866)

[圖 45 篩選1000個出現次數最多的資料(SVC) 30](#_Toc163526867)

[圖 46先做參數挑選，再篩選出500個資料(SVC) 30](#_Toc163526868)

[圖 47 篩選500個出現次數最多的資料(SVC) 31](#_Toc163526869)

[圖 48 最佳模型流程圖 32](#_Toc163526870)

[圖 49 分類後的保留詞彙 32](#_Toc163526871)

[圖 50 迴歸分析視覺化(八卦版) 33](#_Toc163526872)

[圖 51 迴歸分析視覺化(女孩版) 34](#_Toc163526873)

[圖 52 迴歸分析視覺化(政黑版) 34](#_Toc163526874)

[圖 53 決策樹視覺化 35](#_Toc163526875)

# 詞彙共現關係觀察

1. 轉換成矩陣型態

觀察詞彙之間的關係，已篩選詞彙數量200為例，使用DTM功能轉為矩陣型態。

**A screenshot of a computer

Description automatically generated**

圖 1 DTM矩陣各詞出現次數

說明:

1. 市場波動性指示詞:「上漲」與「下跌」這類詞彙直接反映了市場的波動性，這些詞彙的頻繁出現表示市場在某一時期內可能經歷了顯著的價格變動。
2. 市場活躍度指標: 「交易量」和「人數」代表了市場活躍度，其中「交易量」的增加通常與市場活躍度提高相關，而「人數」的增加則可能指市場參與度的提升。
3. 特定股票關注度:「中信金」、「中華電」、和「中鋼」作為關鍵字出現，顯示了市場對這些特定股票的高度關注。這可能與它們在市場中的表現、財報發布或是其他重要事項公告有關。
4. 市場情緒反映: 像是「一堆」這類相對非正式的表達可能用於描述市場中的某種混亂或不確定性，而「不用」可能暗示對某些策略或資產的看衰。
5. 企業與個股分析: 「企業」和「個股」的討論反映了分析的焦點不僅僅在於整體市場，還包括對單一公司或股票的深入研究。這類分析對於識別具體投資機會至關重要。
6. 計算共現矩陣

透過分析共現矩陣中頻繁出現的詞彙如「美國」、「升息」、「金融」、「外資」、「金控」、「風險」、「企業」、「產業」、「價值」、「標的」、「資產」、「基金」、「買進」、「長期」、「穩定」等，探討了當前金融市場的主要趨勢和投資方向。

1. 美國升息與金融市場：美國的升息行為對全球金融市場產生深遠影響，包括增加借貸成本、影響匯率波動，以及對投資回報率的潛在壓力。金融機構和投資者需關注美聯儲的政策動向，並調整投資策略。
2. 外資動向：「外資賣超」與「外資買超」的提及顯示了外資在市場中扮演重要角色。外資流入與流出反映了國際投資者對本地市場的信心與評估，是影響市場情緒與趨勢的關鍵因素。
3. 金融機構與風險管理：「金控」與「金融」的聯繫突顯了金融控股公司在市場中的影響力。「風險」、「企業」、「產業」、「價值」和「標的」的共現指向了市場對於風險評估與價值投資的重視，特別是在不確定的經濟環境中。
4. 投資策略與方向：「資產」、「長期」與「金融」的共現性表明了投資者對長期金融資產持有的偏好。
5. 「基金」、「美國」和「金融」的關聯性顯示美國金融市場對全球基金投資者的吸引力。
6. 「買進」、「長期」和「標的」以及「長期」與「穩定」的結合，強調了市場對長期且穩定投資標的的追求。
7. 「穩定」和「標的」、「產業」、「價值」與「企業」的關聯分析，反映了投資者對於穩定性和內在價值的重視。

**A screen shot of a graph

Description automatically generated**

圖 2 共現關係熱圖

**A white rectangular grid with small squares

Description automatically generated with medium confidence**

圖 3 共現關係DTM矩陣

# 詞彙相關性觀察

1. 計算相關性矩陣

**A graph with a line

Description automatically generated**

圖 4 相關性熱圖

相關性DTM矩陣說明:

1. 報酬與長期投資：長期視角下的市場報酬與企業價值緊密相關。投資人應關注企業的長期增長能力和價值創造，而非僅僅追求短期的獲利。
2. 市場波動與機會：市場波動與投資機會之間存在著相關性。在市場波動中尋找穩定的投資標的，需要投資人有較強的市場觀察能力和產業趨勢分析能力。
3. 企業研究與投資決策：對企業進行深入研究，了解其核心價值和盈利能力，對於制定有效的投資策略至關重要。這需要投資人評估企業在面臨市場衰退時的抗風險能力。
4. 投資人觀察與策略調整：投資人應密切觀察市場趨勢和產業變化，及時調整投資組合，以適應市場的變化。靈活的策略能夠幫助投資人在市場衰退中尋找到反轉的機會。
5. 風險管理：了解各因素如「獲利」、「影響」和「衰退」之間的相關性，對於風險管理而言十分重要。這有助於投資人在面對不確定性和市場波動時，作出更加謹慎的決策。

**A white background with black and white text

Description automatically generated with medium confidence**

圖 5 相關性DTM矩陣

1. 單中心網絡圖

單中心網路圖關鍵詞中，我們可以看到「台積電」位於核心位置，周圍則是與其相關的因素，如「外資」、「大盤」、「市場」、「獲利」、「金融股」、「長期」、「美國」、「資金」、「風險」、「殖利率」、「股息」、「存股」等。這樣的配置暗示著台積電在市場中的重要性及其與多個市場因素之間的關聯，作為全球半導體產業的領先企業，台積電不僅對台灣股市有著舉足輕重的影響，也是全球投資者密切關注的對象。透過對相關市場因素的分析特別已外資流向說明：外資對台積電的投資動態，反映了國際市場對台積電以及整個台灣股市的信心和評價。外資的流入通常與股價上漲相關，反之亦然。

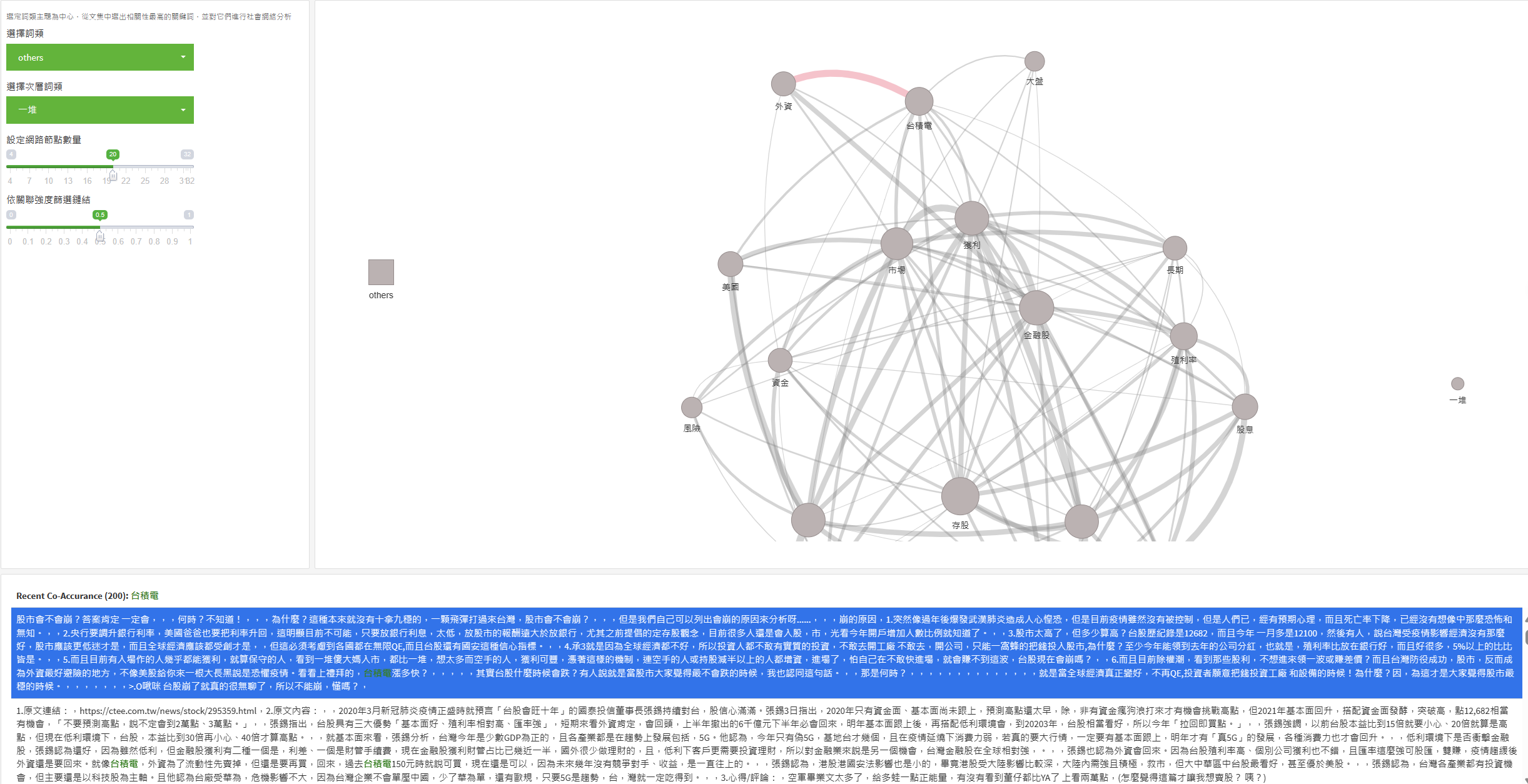
****

圖 6 單中心網絡圖

1. 字詞網絡圖

這些分群涵蓋了從金融市場的信息搜集與解讀、企業分析、財務狀況，到金融機構和貸款等多個方面。這些群體代表了金融市場分析的不同維度和相關參與者的活動範圍。

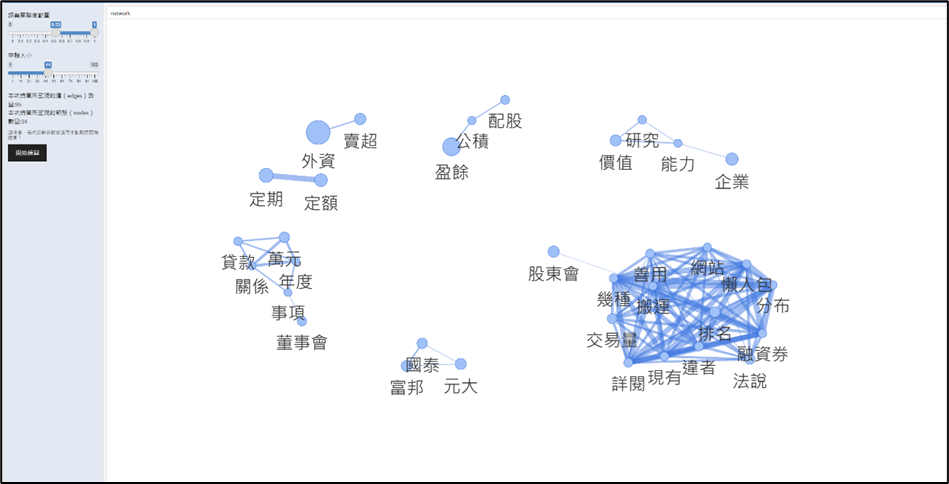


圖 7 字詞網絡圖

可以大致分為以下幾個群體說明:

1. 市場訊息與資源利用群：這一群體涵蓋了「網站」、「懶人包」、「分布」、「排名」等關鍵詞，指向了市場信息收集和解讀的重要渠道及工具。這提示投資者和分析師需要善用現有的訊息資源，如融資券、法說等資料的詳細閱讀和解析，來把握市場動向和交易機會。
2. 企業分析群：包括「企業」、「能力」、「研究」、「價值」等詞，強調對企業的深入研究及其核心價值和能力的評估。
3. 財務狀況群：圍繞「盈餘」、「公積」、「配股」等財務指標，這一群體聚焦於企業的財務狀況和投資者回報，這些指標是評估企業健康度和投資吸引力的重要依據。
4. 金融機構群：以「國泰」、「富邦」、「元大」等金融機構為代表，這些機構在市場中扮演著資金提供和投資服務的角色，對市場流動性和投資氛圍有重要影響。
5. 貸款與融資活動群：此群涵蓋「貸款」、「萬元」、「年度」等詞，關注企業融資活動及其與金融機構之間的關係。這反映了融資條件、貸款政策等對企業經營和擴張計劃的重要性。
6. 關聯式文字雲

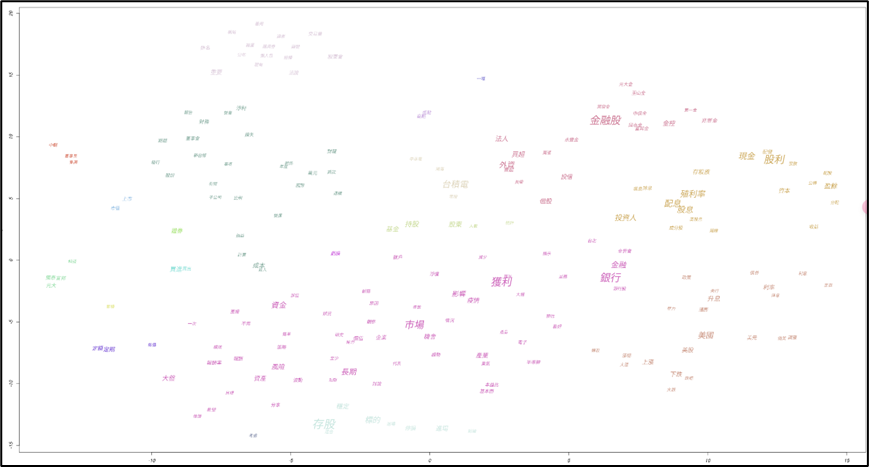


圖 8 關聯式文字雲

說明:

1. 顏色區分：使用不同的顏色來標示不同的關鍵詞群體。例如，市場信息群體使用桃紅色，財務狀況群體使用紅色，金融機構群體使用暗紅色，股東散戶投資者則用紫色，最後貸款與融資活動群體使用紫色。
2. 市場信息與資源利用：「機會」、「趨勢」、「長期」、「風險」等詞彙圍繞著如何有效收集和解讀市場信息
3. 企業分析：以「企業」為中心，「能力」、「研究」、「價值」等詞彙展示了評估企業時考慮的要素。這些關鍵詞之間的連線強調了深入了解一家企業的重要性。
4. 金融機構：「第一金」、「兆豐金」、「元大金」等金融機構的名稱表明了它們在金融市場中的角色和重要性。

# 加入字典後，觀察共現性

1. 步驟說明

* 步驟一 : 替換字串

A screenshot of a computer

Description automatically generated

圖 9 替換字串

* 步驟二：保留重要字

A screenshot of a computer

Description automatically generated

圖 10 保留重要字

* 步驟三：轉為DTM

A screenshot of a computer

Description automatically generated

圖 11 轉為DTM

A screenshot of a computer

Description automatically generated

圖 12 DTM矩陣各詞出現次數(字典)

市場與產業趨勢說明:

1. 市場與產業趨勢

* **多元產業聚焦：從關鍵詞中可見，分析覆蓋範圍廣，從傳統製造業如三陽工業、世紀鋼，到金融服務業如上海商銀與中信金，展示了市場關注的多元化。**
* **上市與上櫃公司表現：關鍵詞中的「上市」與「上櫃」反映出對於不同市場層級公司表現的關注，尤其是對於其財務表現、股價波動和市場情緒的分析。**

1. 金融**產品**

* **金融產品多樣性：關鍵詞如:「中信優先積融債」與「中信高評級公司債」指出了市場上金融產品的多樣化，以及投資者對於不同風險等級和收益潛力金融產品的興趣。**

1. 行業**與公司**

* **如「三陽工業」與「世紀鋼」代表了製造業的關注焦點，而「上海商銀」和「中信金」則代表了金融行業的重要性。**

A screen shot of a graph

Description automatically generated

圖 13 共現性矩陣熱圖

A group of small black and white text

Description automatically generated with medium confidence

圖 14共現性矩陣

1. 行業分析與公司概覽
2. 科技製造業的領頭羊：緯創、廣達、群創均屬於科技製造業，專注於電子產品的設計與製造服務（ODM）和原始設備製造（OEM）。它們在全球供應鏈中扮演著關鍵角色，尤其是在電腦、智慧設備等領域。
3. 電信服務巨頭：台灣大和中華電則代表了台灣的電信服務行業，兩者在提供移動通訊、寬頻網路服務方面佔據市場領導地位，是連接消費者和企業數位經濟的重要橋樑。
4. 石化行業的重量級玩家：台塑化與台塑是石化行業的兩大巨頭，涉及化工原料、塑膠製品等多個領域，在全球化工供應鏈中佔有重要地位。
5. 共現性分析
6. 供應鏈協同：科技製造業（緯創、廣達、群創）與電信服務業（台灣大、中華電）之間可能存在較高的共現性，反映出科技產品的製造與電信服務的提供之間存在著緊密的合作與依賴關係。
7. 跨行業合作：石化產業（台塑化、台塑）與科技製造業的共現性分析可能揭示出原料供應與製造業務之間的聯繫，尤其是在原料採購和環保材料使用等方面。
8. 市場動態：不同行業之間的共現性也可能反映市場對於跨行業合作、並購或戰略夥伴關係的關注度。
9. 單中心網路圖

利用單一中心網路圖的互動性功能，可以在次級詞類中選擇特定詞彙，進而深入探索與該詞彙相關的詞彙網絡。此外，還能進一步檢索出含有所選詞彙的相關文章。



圖 15 參數設定

A screenshot of a computer

Description automatically generated

圖 16 計算共現矩陣

A screenshot of a computer

Description automatically generated

圖 17 單中心網路圖(字典)1

A close-up of a computer screen

Description automatically generated

圖 18 單中心網路圖(字典)2

A screenshot of a computer

Description automatically generated

圖 19 單中心網路圖(字典)3

|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| 社群媒體第二次讀書會  **主題:PTT推文(八卦、女孩、政黑)文件分類** |

# 文件分類

1. 工作流程平台建立



圖 20 文件分類流程圖

1. 資料前處理

我們取PTT三個主要的閒聊版(八卦、女孩、政黑)的推文資料下來，並依此分類，希望能夠由推文內容辨別是哪個版的推文抓取。

A screenshot of a computer

Description automatically generated

圖21 抓取PTT八卦版資料

A screenshot of a computer

Description automatically generated

圖22 抓取PTT女孩版資料

A screenshot of a computer

Description automatically generated

圖23 抓取PTT政黑版資料

因資料量太大，且推文字數較少會不好分辨，於是我們取字串長度大於20的推文並給予分類，將三個版的推文資料匯出整理後用csv檔匯入。

A screenshot of a computer

Description automatically generated

圖24 透過CSV來上傳資料

A screenshot of a computer

Description automatically generated

圖25 Tf-idf詞彙數量選擇1000

A screenshot of a computer

Description automatically generated

圖 26 合併資料後篩選欄位

A screenshot of a computer

Description automatically generated

圖 27 80%訓練20%測試

1. 分類器

A screenshot of a computer

Description automatically generated

圖28 迴歸分析



圖29 篩選1000個出現次數最多的資料(迴歸分析)

A screenshot of a computer

Description automatically generated

圖30先做參數挑選，再篩選出500個資料(迴歸分析)

A screenshot of a cell phone

Description automatically generated

圖31 篩選500個出現次數最多的資料(迴歸分析)

我們可以取測試資料準確度跟測試資料加權精確度兩個數字，然後用這兩個數字來判斷我們這個模型表現的好壞。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序 | 參數值 | 測試資料準確度 | 測試資料加權精確度 | 選擇 |
| 1 | 1000 | 0.569 | 0.577 |  |
| 2 | 500(參數選擇) | 0.584 | 0.625 | V |
| 3 | 500 | 0.599 | 0.617 |  |



圖32 決策樹

A screenshot of a computer

Description automatically generated

圖33 篩選1000個出現次數最多的資料(決策樹)

A screenshot of a computer

Description automatically generated

圖34 先做參數挑選，再篩選出500個資料(決策樹)

A screenshot of a computer

Description automatically generated

圖 35 篩選500個出現次數最多的資料(決策樹)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序 | 參數值 | 測試資料準確度 | 測試資料加權精確度 | 選擇 |
| 1 | 1000 | 0.496 | 0.45 |  |
| 2 | 500(參數挑選) | 0.511 | 0.656 | V |
| 3 | 500 | 0.508 | 0.646 |  |



圖 36 Naive Bayes

A screenshot of a computer

Description automatically generated

圖 37 篩選1000個出現次數最多的資料(Naive Bayes)

A screenshot of a computer

Description automatically generated

圖38先做參數挑選，再篩選出500個資料(Naive Bayes)

A screenshot of a computer

Description automatically generated

圖39篩選500個出現次數最多的資料(Naive Bayes)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序 | 參數值 | 測試資料準確度 | 測試資料加權精確度 | 選擇 |
| 1 | 1000 | 0.431 | 0.542 |  |
| 2 | 500(參數挑選) | 0.379 | 0.556 |  |
| 3 | 500 | 0.403 | 0.559 | V |

A screenshot of a computer

Description automatically generated

圖40 KNN

A screenshot of a phone

Description automatically generated

圖41 篩選1000個出現次數最多的資料(KNN)

A screenshot of a computer

Description automatically generated

圖42 先做參數挑選，再篩選出500個資料(KNN)

A screenshot of a computer

Description automatically generated

圖43 篩選500個出現次數最多的資料(KNN)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序 | 參數值 | 測試資料準確度 | 測試資料加權精確度 | 選擇 |
| 1 | 1000 | 0.5 | 0.508 |  |
| 2 | 500(參數挑選) | 0.553 | 0.56 | V |
| 3 | 500 | 0.547 | 0.545 |  |

A screenshot of a computer

Description automatically generated

圖44 SVC

A screenshot of a computer

Description automatically generated

圖45 篩選1000個出現次數最多的資料(SVC)

A screenshot of a computer

Description automatically generated

圖 46先做參數挑選，再篩選出500個資料(SVC)

A screenshot of a computer

Description automatically generated

圖 47 篩選500個出現次數最多的資料(SVC)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序 | 參數值 | 測試資料準確度 | 測試資料加權精確度 | 選擇 |
| 1 | 1000 | 0.484 | 0.4 |  |
| 2 | 500(參數挑選) | 0.506 | 0.738 | V |
| 3 | 500 | 0.503 | 0.727 |  |

其中表現最好的是SVC，但他在1000個參數表現的很差，SVC受到關鍵字的雜訊太多影響，但經過參數挑選後測試精確度有顯著的提高，我們最後是用參數挑選這個模型為我們的最佳模型。

1. 最佳模型流程

A diagram of a flowchart

Description automatically generated

圖 48 最佳模型流程圖

A screenshot of a computer

Description automatically generated

圖 49 分類後的保留詞彙

1. 分類視覺化

A screenshot of a computer

Description automatically generated

圖 50 迴歸分析視覺化(八卦版)

因為svc模型是黑箱的，我們無法知道他是用什麼東西去判斷，但是迴歸跟決策樹是可以看得到它是用哪些參數去判斷的，從上圖我們可以看到:

1. 正面係數的關鍵字:收到、以色列、吸管…等是判斷八卦版的關鍵字。

* 「收到」 : 因收到國家級警報簡訊，所以有收到這個關鍵字。
* 「以色列」: 因戰爭議題，有比較國際化的新聞訊息。

1. 負面係數的關鍵字:國民黨、民眾黨、政黑..等，這些都會在政黑版出現所以會造成誤判。

A screenshot of a computer

Description automatically generated

圖 51 迴歸分析視覺化(女孩版)

女孩版關鍵字有line、女生、公分、房租、有錢人…等，比較屬於女孩子能辨識的話題，至於中國、小草、中央、kmt、柯文哲容易被判辨成政黑或八卦。

A screenshot of a computer

Description automatically generated

圖 52 迴歸分析視覺化(政黑版)

政黑的關鍵字就是一些政治團體，食材因為食安的問題所以出現，負面係數如Line、pr、房子、小孩容易判別到女孩版。

A screenshot of a computer

Description automatically generated

圖 53 決策樹視覺化

決策樹出現重要性明顯最高四個食材、國民黨、社工、中華民國，這四個是決策樹模型拿來判別最重要的參數。