**統計計算期末報告**

暢銷財金類書籍文字探勘分析

系級/學號：計財所碩二108071601賴冠維

目錄

1. **摘要**
2. **資料清洗**
3. **探索性資料分析**
4. **情緒分析**
5. **主題模型**
6. **與其他書籍比較**
7. **結論**

一、 **摘要**

分析著名財金暢銷書：“The Most Important Thing: Uncommon Sense for the Thoughtful Investor”，由Howard S. Marks所著，作者為一位美國投資者、企業家和作家，曾預測金融海嘯和網路泡沫，被譽為「價值投資大師」。 由於高效率的投資策略，使得Howard S. Marks在財務領域擁有著強勁而可靠的信譽。

價值投資是主動投資領域中最古老的派別之一。主要根據基本面、產生現金流能力等量化因素的內在價值，以及在股價明顯低於內在價值時買入。盡可能去預測未來的現金流量，並使用折現率回推現值，折現率由當前的無風險利率(美國國庫券報酬)加上風險溢酬來補償其不確定性。

本研究對此書進行文字探勘分析，利用文字雲、二元語法(bigram)、TF-IDF、Topic Model等方式，對書中的內容進行分析，最後與其他財金書籍內容進行比較，凸顯價值投資與其他投資策略的差異。

1. **資料清洗**

本研究使用tidytext dataset內建的停止詞(stop\_words)進行資料清洗，包含像是“the”, “of”, “to”等單獨無意義的單字。

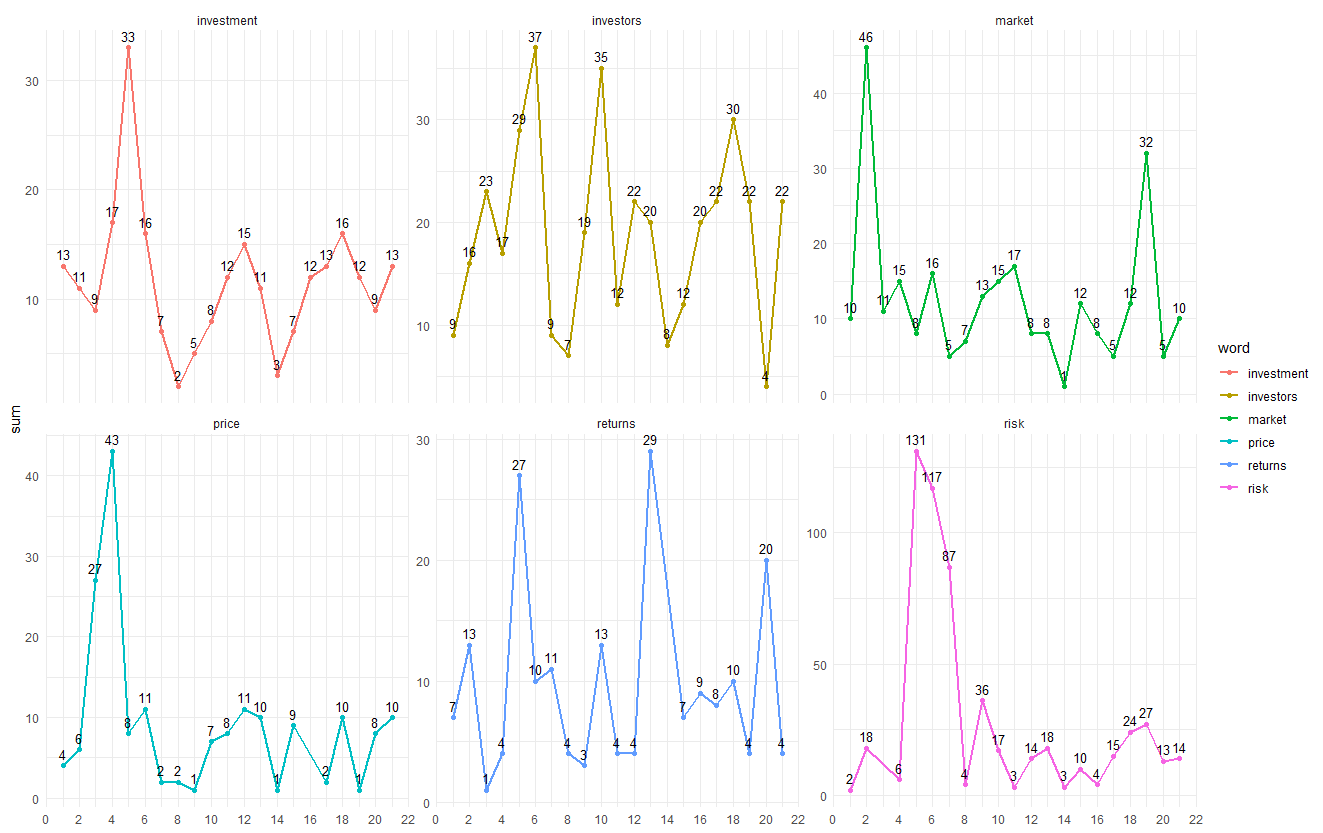
1. **探索性資料分析**
3. **文字雲**

依照字出現頻率畫出文字雲，觀察整本書提到的重點，可以看到risk,investors,returns,market…等字大量出現，推測本書內容涵蓋風險的解釋、投資人對於報酬、市場的理解及看法等內容。



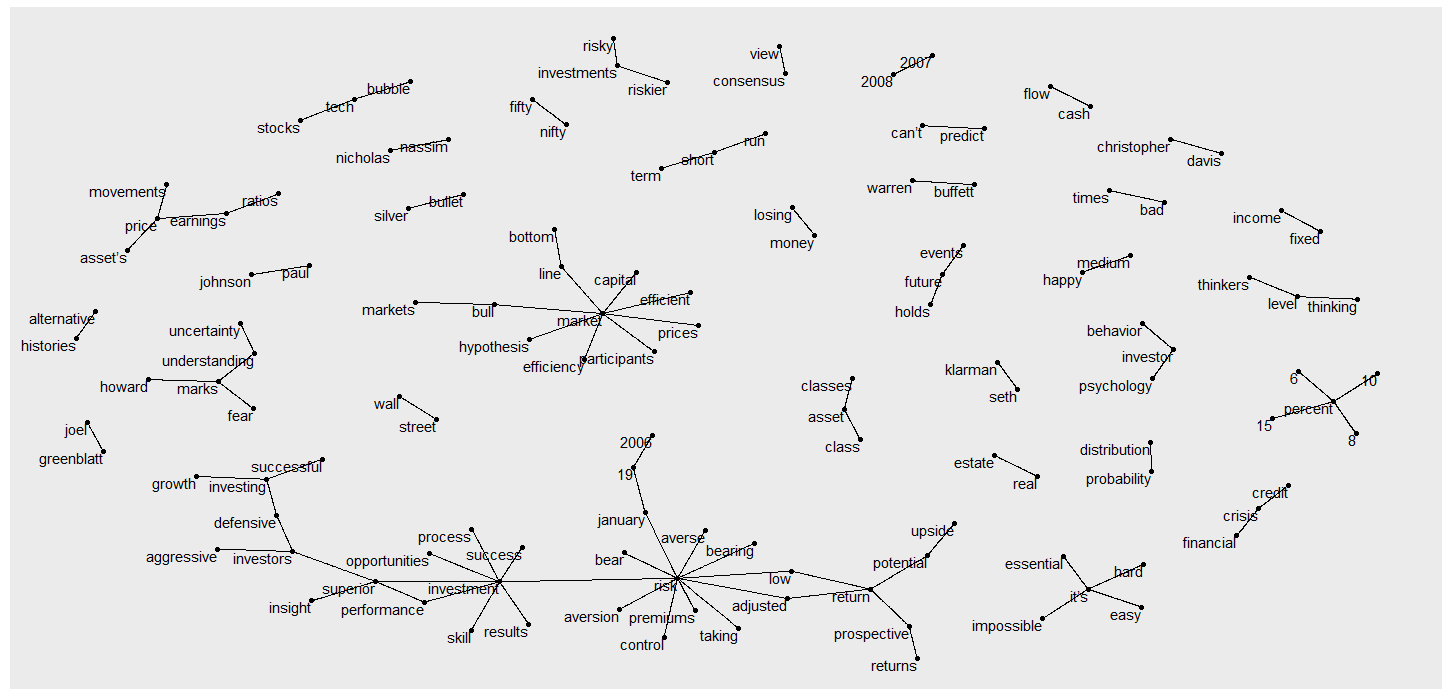
1. **熱門字在各章節出現頻率**

由下圖可見，從文字雲中發現的熱門字，在各章節中出現的頻率，price集中在第四章，Investment集中在第五章，risk集中在第五至第七章，marker集中在第二、第十九章，可由此初探各章節談論的內容大致為何。



1. **二元語法(bigram)分析**

我們將分析的字數擴增為二元語法，連續觀察兩個字聯合出現的字數，捕捉更完整的語意，如下圖可見，前面觀察到的risk與bear、premiums、averse、control相連，談論重點可能與景氣不好時的熊市、投資人的風險趨避屬性、估計承擔風險所獲得的風險溢酬等內容；再看到market的部分，可看到與capital、efficients、bull、hypothesis等字相連，代表本書可能談論到資本市場上是否符合效率市場假說，在景氣好的牛市、景氣差的熊市時，是否假設有所改變。

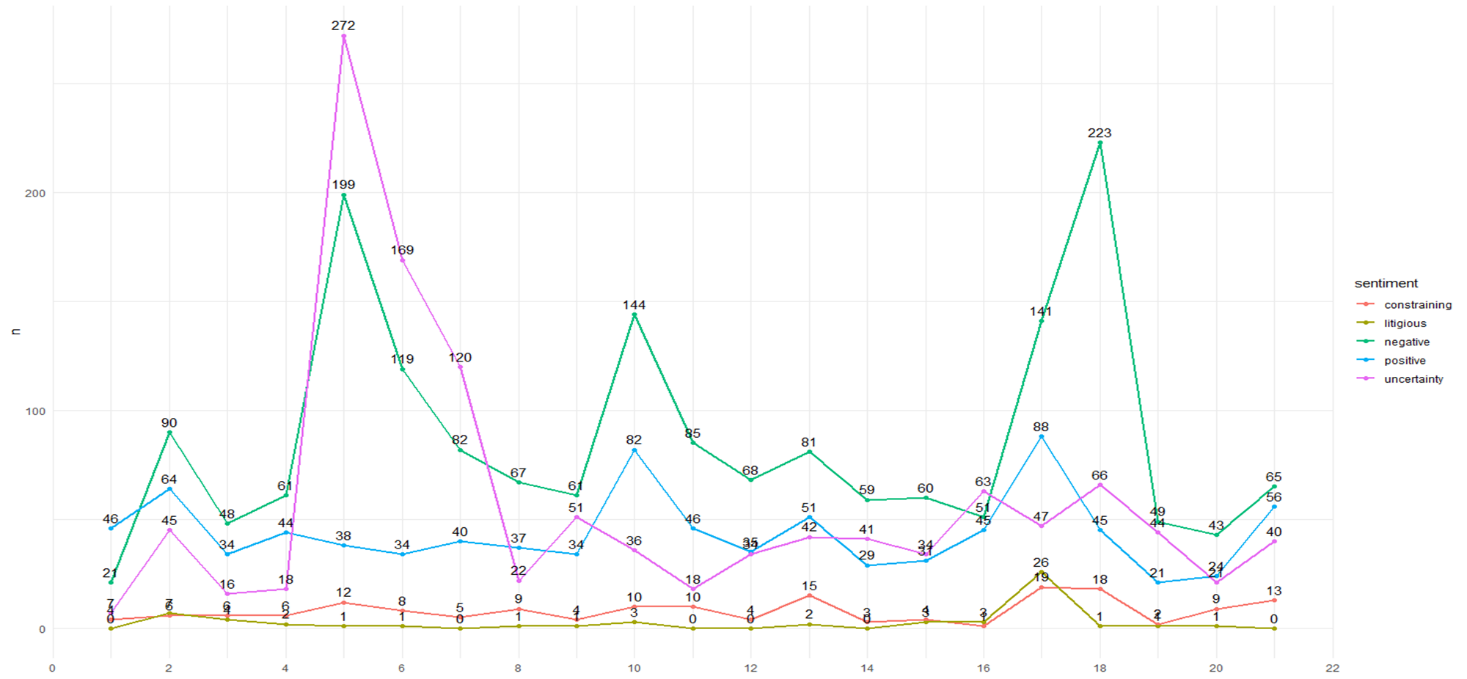


1. **情緒分析(Sentiment Analysis)**

因為本研究書籍為財金相關資料，故與一般文章所使用的情緒字典不太相容，可能會出現會錯意的情況發生，像是代表股份的share跟分享，還有風險對財金來說應該為中性詞而非負面詞，綜和上述，本研究採用Loughran and McDonald dictionary of financial sentiment terms (Loughran and McDonald 2011)，分為六種情緒: “positive”, “negative”, “litigious”, “uncertain”, “constraining”, and “superfluous”，代表“正面”、“負面”、“爭議的”、“不確定性”、“約束的”和“多餘”等六種概念。

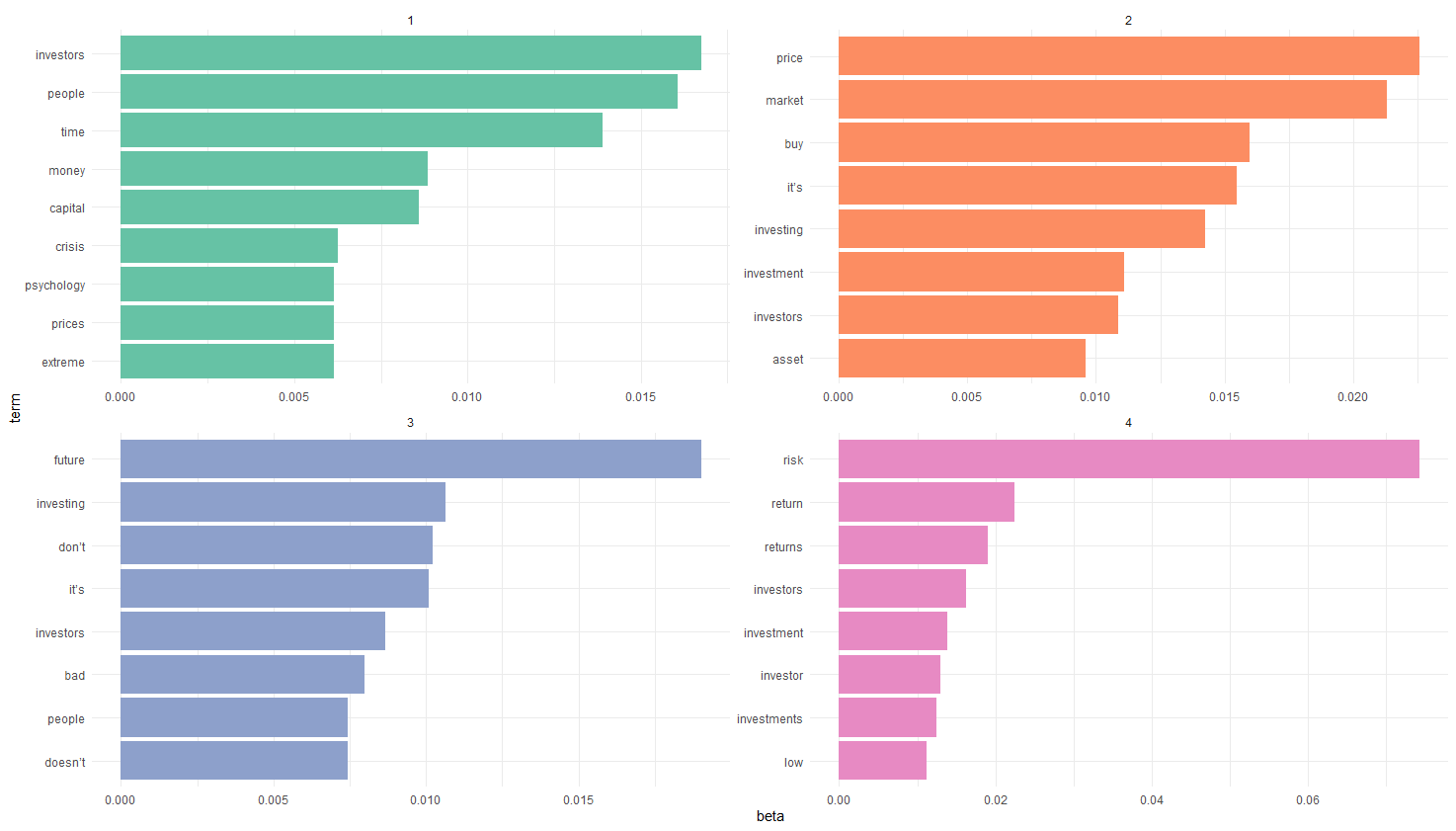
由下圖可以看到第五、六、七章的不確定性及負面傾向最為明顯，另外

第十、十七、十八章的負面也很突出。



1. **主題模型(Topic Model)**

主題模型多由Latent Dirichelet Allocation（LDA）所建立，LDA是一種主題建模演算法（topic modeling），廣泛使用於以無監督學習，探索語料庫中的隱含主題（latent topic）。需要先給定隱含主題個數，本研究假設個數為4，可得下圖，Topic 1由投資人相關字詞組成，Topic 2由價格、市場等字詞組成，Topic 3由未來等字詞組成，Topic 4由風險相關字詞組成。



接著我們以主題模型中的進行進一步分析，可視為各文件屬於各主題的機率呈現(per-document-per-topic probabilities)，此處我們將本書的21個章節視作21個文件，透過觀察其分類結果，進一步推測本書各章節之間的相關性。由各章節最高的值進行分群，可得下表：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chapter | Topic | gamma | Chapter | Topic | gamma | Chapter | Topic | gamma | Chapter | Topic | gamma |
| 8 | 1 | 0.55345 | 1 | 2 | 0.561636 | 14 | 3 | 0.641344 | 5 | 4 | 0.525942 |
| 9 | 1 | 0.582932 | 2 | 2 | 0.642615 | 16 | 3 | 0.681553 | 6 | 4 | 0.562431 |
| 10 | 1 | 0.487787 | 3 | 2 | 0.681504 | 17 | 3 | 0.626847 | 7 | 4 | 0.503331 |
| 11 | 1 | 0.532463 | 4 | 2 | 0.49014 | 21 | 3 | 0.347718 | 13 | 4 | 0.37548 |
| 15 | 1 | 0.351523 | 12 | 2 | 0.506552 |  |  |  | 19 | 4 | 0.576281 |
| 18 | 1 | 0.442804 |  |  |  |  |  |  | 20 | 4 | 0.438462 |

將各主題內被分類的章節列出標題，可得下圖，可得以下結論，Topic 1與景氣循環相關，並且進行價值投資中重要的一環，反向(逆勢)投資，Topic 2與價格有關，談論標的價值與價格的關係，好標的如果買貴了也是一筆不好的交易，Topic 3談論不確定性，講到預測、運氣、機率等因素存在，因此需要採取防禦型投資策略，最後Topic 4談論風險等內容。

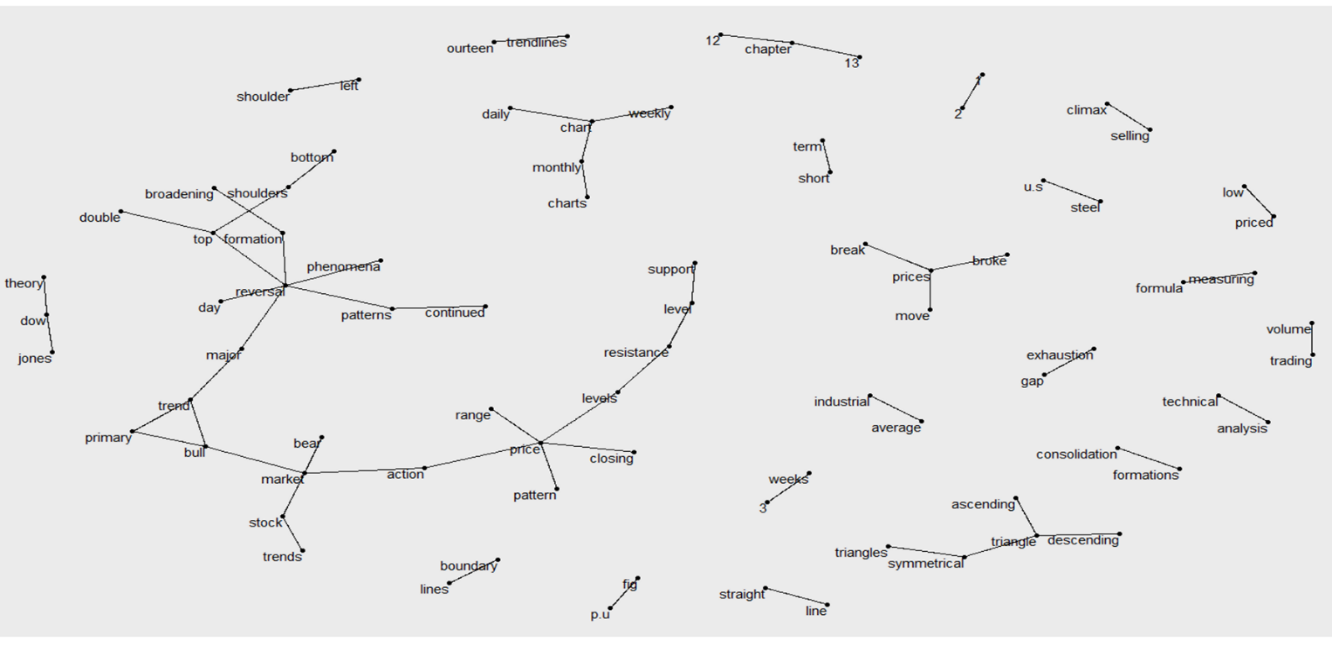


1. **與其他投資策略書籍進行比較**

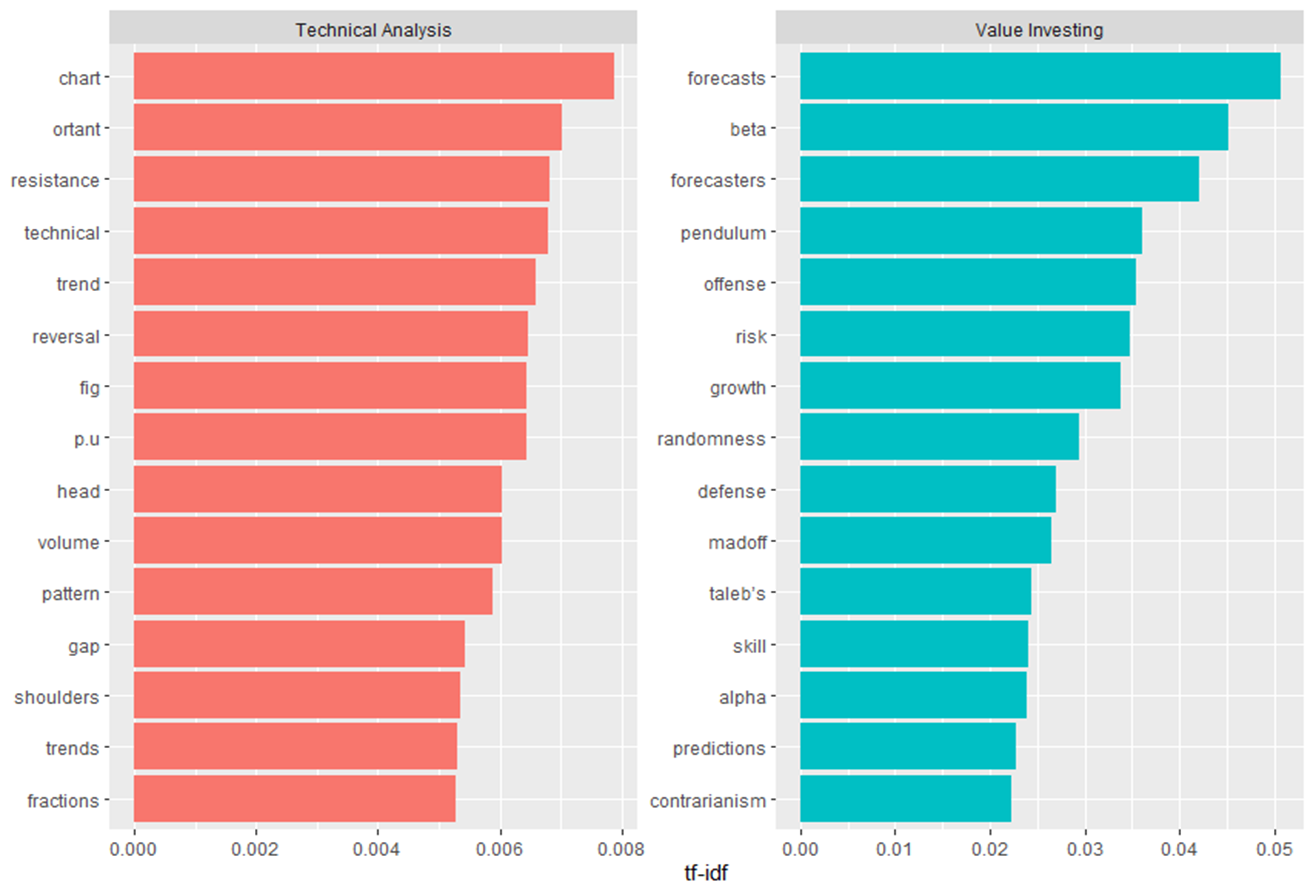
為了凸顯價值投資與其他投資策略的不同之處，本研究採用另一本談論技術分析的書籍，由Robert.D.Edwards所著,” Technical Analysis of Stock Trend ”，技術分析多用於短期內價格的波動、趨勢的推測，與中長期策略的價值投資有明顯不同，先以文字雲觀察，如下圖；可看到出現trend、chart、pattern、volume、reversal等字詞出現，明顯與上述不同。



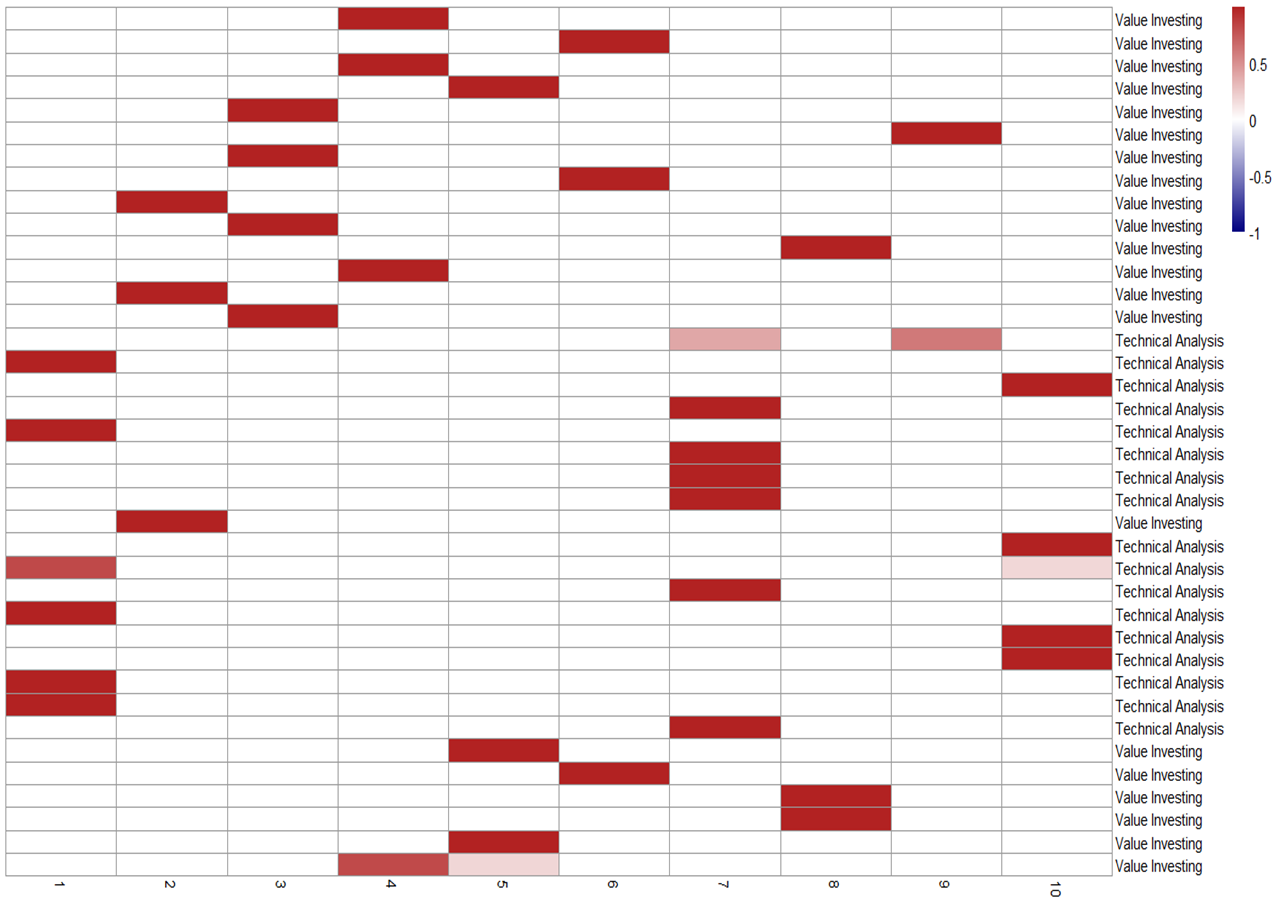
由二元語法也可發現，出現的字詞組合與上述明顯不同，談論day reversal patterns reversal等，還有daily chart、monthly charts等，都在談論短期價格的模式、反轉。



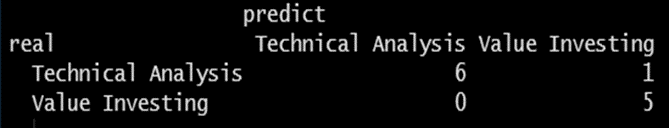
這邊列出兩本書的TF-IDF，可以看到兩者有顯著的不同。



最後同樣將各章節視作文件，將兩本書的資料混合，以主題模型進行分群，雖然我們已知由兩本書組成，但在此我們假設主題個數有10個，得到下圖，可明顯發現，為技術分析書籍的章節時，其皆在Topic 1、7、10，這三個主題中最顯著，而價值投資書籍的章節則在其他主題。



最後將主題模型以SVM模型分類，兩書共有38個章節，抽樣26個章節為訓練集建立模型，其他章節為測試集，測試模型表現，可得混淆矩陣如下，僅有一個章節分錯，代表Topic模型分群的資訊捕捉良好。



1. **結論**

我們可藉由文字探勘分析，系統性、快速地了解整本書的內容，藉由Topic模型或是TF-IDF字詞的統計結果，我們可以很快地掌握各文件、

各章節的內容，並且了解彼此之間的相關性。