**網路爬蟲資料屬性、平台及處理**

**1. 爬取資料屬性 (期間、資料內容及對象):**

(1) 期間:

本計畫以上海及深圳股市五個不同財經及投資平台網站，最早從2012年2月至2018年12月底期間日資料為樣本期間。

(2) 搜集資料內容:

包括每日投資人在這兩個股市所有相關財經網站平台對分別上市ETF個股發佈的風險偏好、預期報酬、觀點及看漲及看跌等看法的充分資訊，建置完整的股票投資人個股情緒的資料庫。

(3)對象:

(a). 上證50指數為標的的ETF (上海證券交易所中，最具有市場引響力的50檔股票): ………………………………………………….

(b). 深證100指數為標的的ETF (深圳證券交易所中選出具代表性的100檔A股股票為標的): …………………………………………..

**2. 爬取資料平台**

(1) 平台簡介：

|  |  |  |
| --- | --- | --- |
| **平臺名稱** | **爬取資料內容** | **說明** |
| 1. 珞珈投資 | 網址: https://www.lj168.com/  抓取頁面: 研究院觀點 https://www.lj168.com/1008/dissort  時間: 2013-01-04 (page=561)  程式檔: crawler\_lj168.py  資料檔: news\_data/lj168.csv |  |
| 2. 中國網財經 | 網址: https://finance.china.com.cn/  抓取頁面: 財經首頁 > 證券 > 證券要聞  http://finance.china.com.cn/stock/zqyw/index.shtml  財經首頁 > 證券 > 上市公司  http://finance.china.com.cn/stock/ssgs/index.shtml  財經首頁 > 證券 > 市場風雲  http://finance.china.com.cn/stock/dp/index.shtml  時間: 2012-02-01(page=366)  程式檔: crawler\_financeChina.py  資料檔: news\_data/financeChina.csv |  |
| 3. 新浪財經 | 網址: https://finance.sina.com.cn/stock/  抓取頁面:  證券 > 個股點評  http://finance.sina.com.cn/roll/index.d.html?  cid=56588  證券 > 上市公司  http://finance.sina.com.cn/roll/index.d.html?  cid=56592&page=1  證券 > 研究報告  http://vip.stock.finance.sina.com.cn/q/go.php/vReport\_List/p=7491  時間: 2018年03月12日 page22  程式檔: crawler\_sina.py  資料檔: news\_data/crawler\_sina.csv |  |
| 4. 同花順財經 | 網址: http://stock.10jqka.com.cn/company.shtml  抓取頁面:  首頁 >股票 > 公司資訊  時間: 2019-07-02 10:06:31 index\_183  215 pages |  |
| 5. 和訊股票 | 網址: http://stock.hexun.com/  抓取頁面: 股票首頁 > 上市公司 > 公司要聞  http://stock.hexun.com/gsxw/  用json回傳的  股票首頁 > 股市報導  http://stock.hexun.com/hxjrbd/  用json回傳的 |  |

一、珞珈投資

珞珈投資成立於1996年，十六年專注於證券諮詢業務，專業和權威鑄就了在證券諮詢行業的知名品牌。珞珈投資將上市公司各階段的價值分析的研究成果快速地提供給投資者，經中國證監會批准的證券投資諮詢機構，是目前國內證券市場服務範圍最為廣泛的投資諮詢公司，能及時準確地把握最新政策動態並提供重要證券資訊分析資料，致力於為各類投資者打造學習高手的交流平臺。

二、中國網財經

中國網是國務院新聞辦公室領導，中國外文出版發行事業局管理的國家重點新聞網站。此網通過10個語種11個文版，24小時對外發佈資訊，是中國進行國際傳播、資訊交流的重要窗口。

三、新浪財經

新浪財經提供7X24小時財經資訊及全球金融市場報價，覆蓋股票、債券、基金、期貨、信託、理財、管理等多種面向個人和企業的服務。

四、同花順財經

核新同花順網路資訊股份有限公司（同花順）成立於1995年,是一家專業的互聯網金融資料服務商,為您全方位提供財經資訊及全球金融市場行情,覆蓋股票、基金、期貨等。

五、和訊股票

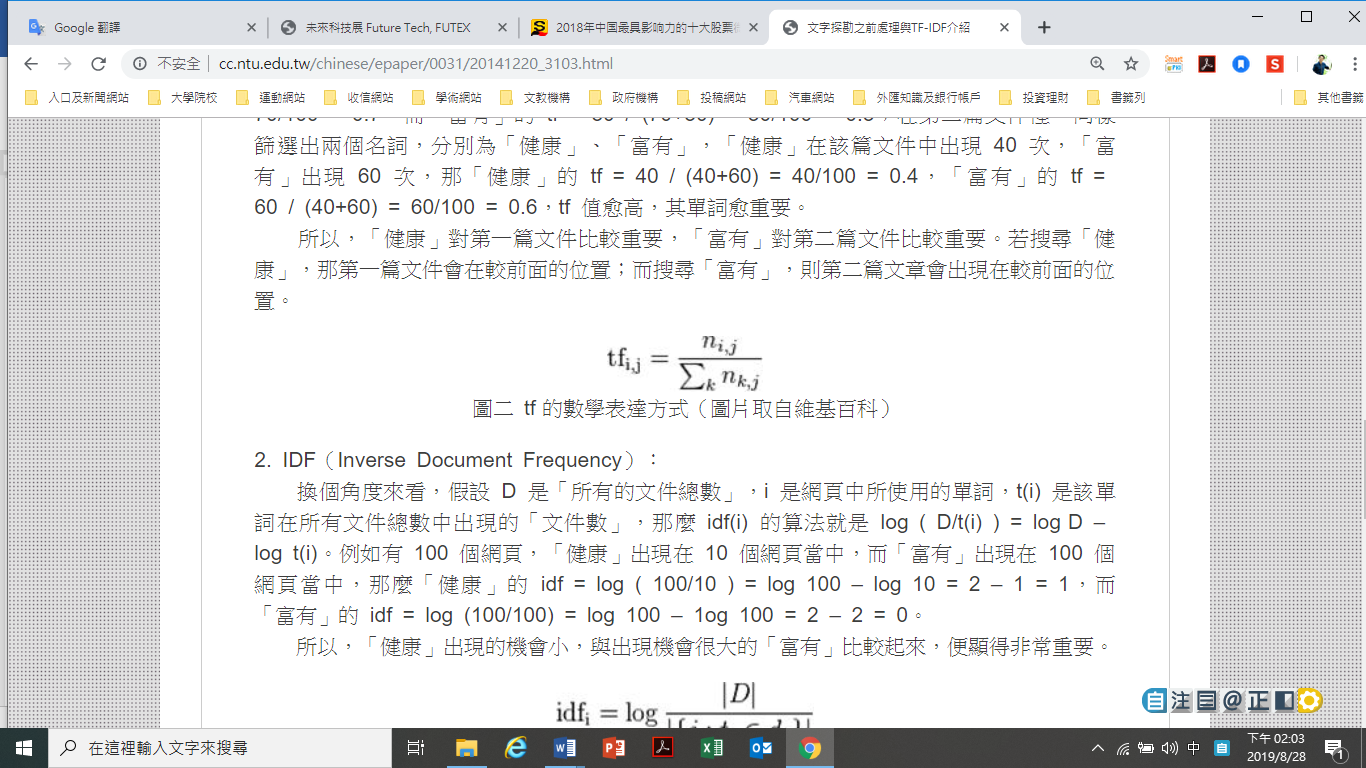
和訊股票，打造的全新手機證券綜合交易平臺。注重中國股民的看盤習慣，以實用、易用、好用為宗旨，以股票即時行情為核心，配合股票分時K線等基本資料，依託和訊強力資料分析團隊，提供最新財務評估及研究報告各種實用股票技術指標，讓客戶一鍵盡享基本資料、個股資金流向、相關公告、新聞等，更有和訊為您提供的第一手股票研究報告和財務評估。

不同平台背景及客戶群的差異

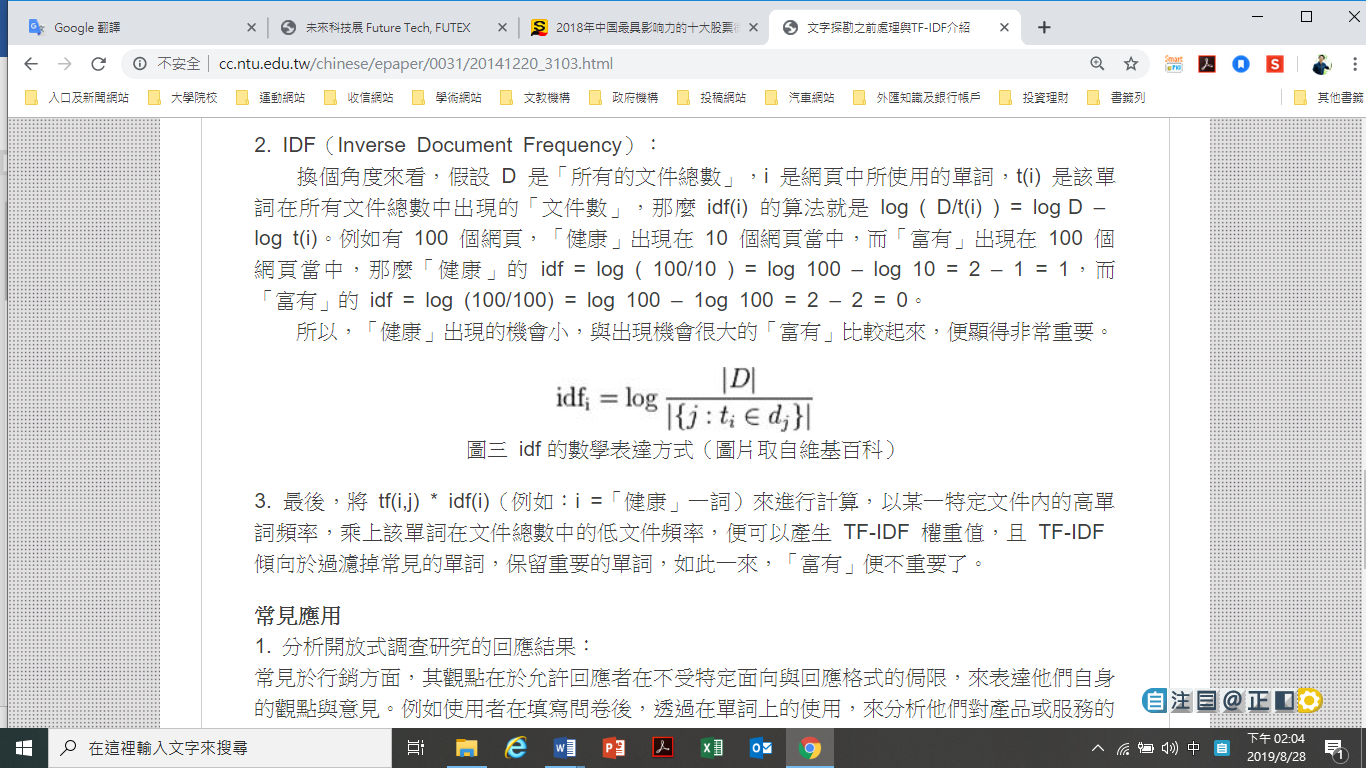
**3. 網路爬取資料處理-linguistic media**

現今應用 Linguistic Text Mining 所創造出來的產值，已超過過去以資料庫搜尋為主的價值。在網際網路以非常快速的方式，所累積愈來愈多非結構化資料的趨勢下，實際運用  Linguistic Text Mining 來探索出各種未能預見、創新、重要的資訊或知識，如商品評價、施政評價、民意調查、社群偏好等等，對政府機關、企業團體或個人而言，都是非常重要的工作，Linguistic Text Mining 的相關研究和實作，已成為晚近最為熱門的話題。隨著大量金融異常現象和非理性行為的出現，傳統金融理論已經不能做出合理解釋，學者們開始從新的角度解釋金融異常現象，由此逐漸形成了行為金融理論。投資者情緒理論成為行為金融學的重要分支之一，而投資者情緒的測量問題一直是金融學研究的重點和難點之一，投資者情緒與股票收益之間的關係研究也是行是金融學研究的熱點。本計畫研究投資者情緒指數的構建以及其對股票市場收益的影響，對投資者投資股票市場的決策、相關監管部門具有重要的現實參考意義。

**4. 網路爬取資料進階處理**-**TF-IDF（Term Frequency-Inverse Document Frequency）**  
 TF-IDF 是一種用於資訊檢索與文字探勘的常用加權技術，為一種統計方法，用來評估單詞對於文件的集合或詞庫中一份文件的重要程度，說明如下：  
(1)TF（Term Frequency）：  
　　假設 j 是「某一特定文件」，i 是該文件中所使用單詞或單字的「其中一種」，n(i,j) 就是 i 在 j 當中的「出現次數」，那麼 tf(i,j) 的算法就是 n(i,j) / (n(1,j)+n(2,j)+n(3,j)+…+n(i,j))。例如第一篇文件中，被我們篩選出兩個重要名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 70 次，「富有」出現 30 次，那「健康」的 tf = 70 / (70+30) = 70/100 = 0.7，而「富有」的 tf = 30 / (70+30) = 30/100 = 0.3；在第二篇文件裡，同樣篩選出兩個名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 40 次，「富有」出現 60 次，那「健康」的 tf = 40 / (40+60) = 40/100 = 0.4，「富有」的 tf = 60 / (40+60) = 60/100 = 0.6，tf 值愈高，其單詞愈重要。所以，「健康」對第一篇文件比較重要，「富有」對第二篇文件比較重要。若搜尋「健康」，那第一篇文件會在較前面的位置；而搜尋「富有」，則第二篇文章會出現在較前面的位置。



2. IDF（Inverse Document Frequency）：  
　　換個角度來看，假設 D 是「所有的文件總數」，i 是網頁中所使用的單詞，t(i) 是該單詞在所有文件總數中出現的「文件數」，那麼 idf(i) 的算法就是 log ( D/t(i) ) = log D – log t(i)。例如有 100 個網頁，「健康」出現在 10 個網頁當中，而「富有」出現在 100 個網頁當中，那麼「健康」的 idf = log ( 100/10 ) = log 100 – log 10 = 2 – 1 = 1，而「富有」的 idf = log (100/100) = log 100 – 1og 100 = 2 – 2 = 0。所以，「健康」出現的機會小，與出現機會很大的「富有」比較起來，便顯得非常重要。



(3) 最後，將 tf(i,j) \* idf(i)（例如：i =「健康」一詞）來進行計算，以某一特定文件內的高單詞頻率，乘上該單詞在文件總數中的低文件頻率，便可以產生 TF-IDF 權重值，且 TF-IDF 傾向於過濾掉常見的單詞，保留重要的單詞，如此一來，「富有」便不重要了。

**----------------------------------**

本計畫分別針對投資人樂觀及悲觀情緒的*lth*類別來搜集所有潛在的特徵詞，並計算*lth*類別的*jth* 特徵詞的詞頻。如 *rO*和*rP* 分別是樂觀及悲觀情緒的詞頻。在和樂觀及悲觀情緒相關的許多特徵詞中, 代表性特徵詞相較於其他者更攸關。我使用卡方檢測以決定代表性特徵詞(Yang and Pedersen, 1997)。首先劃分屬於樂觀、悲觀及其他(和前兩類無關)類的特徵詞。檢定每個特徵詞是否屬於特定類情緒，每次鑒定兩類(*l*=2), 如樂觀及非樂觀。虛無假設是*jth* 特徵詞屬於非樂觀類，而對立假設是 *jth* 特徵詞屬於樂觀類。悲觀類可設計為相似的假設作法。卡方獨立性檢定的統計量為下列公式:

 (1),

此處A和B是指在*l*th類*j*th和non-*j*th 特徵詞的詞頻; C和D是指在non-*l*th 類*j*th和non-*j*th 特徵詞的詞頻;且N是指所有特徵詞的詞頻。假設*j*th 特徵詞獨立於*l*th類。若卡方值高於臨界值，獨立于樂觀的虛無假設被拒絕；故判定*j*th特徵詞被視為樂觀(*l*th)類的代表性術語。

本计划接续采用焗权重法分别计算乐观及悲观情绪的特征词之权重。透过权重可指出特征词的强度。此方法已被广泛使用在语意分析做为所有代表性特征词相对大小的衡量。以乐观情绪为例，焗值公式如下述:

 (2)

此处*m* 是厂商的数目; 是投资人对*i*th 厂商乐观情绪中*j*th 特征词的词频; 且 是乐观情绪中*j*th 特征词的焗值。本研究接续分别计算乐观及悲观情绪的*j*th特征词的权重。

 (3)

 (4)

此處n 是投資人樂觀或悲觀情緒特徵詞的數目。權重愈大表示樂觀或悲觀的特徵詞愈重要。

此階段計算投資人在某月*t*對*ith* 廠商的“樂觀情緒密度”(*IOSit*)及“悲觀情緒密度”(*IPSit*)。藉由對樂觀情緒計算特徵詞，量化投資人對特定廠商的樂觀情緒之程度。本計畫首先將投資人每日對*i*th廠商樂觀情緒 *j*th特徵詞的詞頻乘以相對應的權重，並累積對樂觀情緒的所有代表性特徵詞，計算為每日*IOS* (*IOSidk*)。接續，我加總每日的*IOS*計算為每月的*IOS*, 並加總每日的*IPS*計算為每月的*IPS*.

 (5-1)

 (5-2)

此處下標*i, d,* 和 *k* 是指*ith* 廠商在*dth*天*kth*則網路資訊; 是指投資人對*ith* 廠商在*dth*天*kth*則網路資訊為樂觀情緒的*j*th 特徵詞的詞頻；則是指相對應悲觀情緒的詞頻；且和是投資人樂觀及悲觀情緒的*j*th 特徵詞的權重。被計算的*IOSit*和*IPSit*可隨時追蹤及更新投資人對*i*th廠商樂觀及悲觀情緒的程度。因為*K* (網路資訊的數目) 和*n* (特徵詞的數目) 是無限的, 這兩個密度變數的最大值是無限的，本計畫分別對*IOSit*和*IPSit*將非0的密度值，使用均等分配的公式[99\*(*x* – *x*min)/(*x*max-*x*min) +1](*x*是密度值)，將之標準化介於最小1% 和 最大100%之間；然而0密度值仍然為0。較高的密度值意謂投資人有較強的樂觀或悲觀情緒。

**二、語意解析解讀*——*建置投資人股票情緒的數據庫**

現有文獻對輿論的看漲或看跌情緒主要是通過帖子的字面含義來界定並計算，可利用人工判斷或輔助以電腦學習的方式來進行。本計畫主要利用五大財經及投資平台股市術語的市場信號來構建投資人股票情緒指數。股市術語指的是股票市場的專有名詞。無論是微博還是其他媒體，凡是和股市有關的討論，都必然大量使用這些術語。重要的是，不少股市術語本身帶有明顯的行情判斷信號。我們認為，對於在五大財經及投資平台上被海量熱議和大量提及的股市術語，用詞頻統計的方法可以構建整個關於股市走勢的總體判斷。當影響個股公司或股市大盤的事件發生時，財經及投資平台發表及評論文章信息產生，本計畫透過網路爬蟲程式密集採集投資人對個股漲跌看法，結合社群聆聽以語意解析來解讀投資大眾對公司發展及公司股價漲跌的評價。並採用「文本情感分析」來建構正面及負面情緒的網路聲量；設立投資人個股情緒的資料庫，並利用其來構建樂觀及悲觀情緒變數。

一個顯而易見的問題在於，某個詞彙帶有明顯利好和利空信號，並不意味著這個詞彙所在的語句和帖子本身具有完全相同的信號。例如，“現在可以建倉”和“現在不能建倉”，雖然都使用“建倉”這個關鍵字，但含義顯然完全相反。因此，詞頻似乎不能代表輿情的指向。此外，即便是用人工解讀來取代詞頻統計似乎也不能解決這個問題。由於互聯網的匿名性和虛假資訊懲戒機制的缺乏，理論上人們無法可靠地判斷發佈者本身是否帶有其他目的。例如，試圖出售手頭股票的發佈者，可能會散佈虛假的利好資訊，也因此同樣一則消息會有截然相反的市場預期解讀。因此，字面含義明確的關鍵字，其在宏觀層次上的統計量 (例如，詞頻加總) 是否仍然能構成表達同樣信號的指標？

因此我們提出分析層次、語法結構、測量方法和文獻證據等四個方面的理由。這些理由使我們可以利用被海量熱議且信號明確的股市術語詞頻構建輿情信心的指標。

第一，分析層次方面。被大量熱議的術語，本身就能夠直接體現輿情對市場的基本判斷和共識。例如，“股災”一詞，當置於單個句子的微觀層次中時，我們無法判斷其代表的信號方向。但如果“股災”被大量討論，在宏觀層次上就映射股市已經出現持續下跌。換句話說，即便是在大家紛紛爭論2015年6月份的股市動盪是不是“股災”時，不管是支持者還是不支持者，其對行情的判斷仍然具有基本的共識：起碼此時股市處於下跌的走勢。同樣，“建倉”一詞的反復出現和海量熱議意味著市場已經不再最高點。實際上，在推特中，學者發現推特語句的含義可以用關鍵字直接代替 (Yang et al., 2012)。

第二，語法結構方面。互聯網，特別是五大財經及投資平台的文本，往往在語法上具有直接和明晰的特點。這一特點主要來自五大財經及投資平台對文字長度的限制和移動互聯網設備的普及。換句話說，發文有文字限制，在移動設備上輸入文字又不方便，這導致專家或記者在五大財經及投資平台上表達一個意思的時候，更傾向於使用直接句式，而不是否定句式。為支撐這一推測，我們專門對本文所使用的五大財經及投資平台股市術語做了分析。例如，我們發現，以“抄底”為例，“不抄底”“不能抄底”“不要抄底”占“抄底”總詞頻總數的比例不到10%。其他的重要術語也都存在這一特徵。這意味著，大量的“抄底”熱議就是人們積極進場的信號。

第三，測量方法方面。即便無法完全證明具有看漲信號的術語詞頻代表真實的看漲信心，我們仍然可以通過數學方法來儘量解決這一問題。最直接的方式就是在量綱統一的情況下，把看漲術語總詞頻指數和看跌術語總詞頻指數相減或相除。如果人們對看漲和看跌兩類詞彙的否定式與肯定式使用比例是大致相當的，那麼就可以通過計算差或商的形式得到可以回避否定式表達的信心指數。基於這樣一個工作假設，我們可以獲得對輿情信心的測量並用於探索性分析。

第四，文獻證據方面。近年來，不少金融和傳播學實證研究已經證明，基於詞彙字面直接含義所構建的輿情信號指標和真實的股市漲跌有正向關聯。例如，在對Google檢索的研究中，對“蕭條”“破產”等詞彙的網路檢索和後市看空顯著相關 (Da et al., 2010)；對“債務”一詞的網路檢索和相關股票行情有重要關聯，甚至可以用來預測後市 (Preis et al., 2013)。在對推特的研究中，研究者發現情緒關鍵字的詞頻和其字面所代表的社會情緒等高度相關 (Kim et al., 2013)，人們常用語和各自在網路社區的層次身份高度相關 (Bryden et al., 2013)。在中國股市研究中，來自和訊網、新浪網等包含“債務”字樣的新聞報導會從負面影響相關股票的行情 (Zhao et al., 2011)。這些研究結果均表明，單個或一系列帶有明確市場信號的股市術語，其詞頻在宏觀層次上和代表其字面含義的經濟社會結果存在正向統計關聯。