Project : from Classification to Prediction

楊立偉教授
wyang@ntu.edu.tw

練習：用AI及社群數據協助投資決策

- 常見的股市分析方法，與各種內外部變數有關

基本面	例如企業的營收、獲利、本益比等
技術面	例如股票價格的走勢，價格及交易量的關係等
消息面	例如重大訊息、產業分析師撰寫之文章內容等
環境面	例如經濟景氣指標，匯率、利率等

基本面

包括企業的登記事項、營業狀況、財務報表等。

以台積電為例

(取自Yahoo!股市)

公司資料			
基本資料		股東會及105年配股	
產業類別	半導體	現金股利	7.00元
成立時間	76/02/21	股票股利	-
上市(櫃)時間	83/09/05	盈餘配股	-
董事長	張忠謀	公積配股	-
總經理	劉德音、魏哲家	股東會日期	106/06/08
發言人	何麗梅		
股本(詳細說明)	2593.04億		
股務代理	中信託02-66365566		
公司電話	03-5636688		
營收比重	晶圓95.91%、其他4.09% (2016年)		
網址	http://www.tsmc.com/		
工廠	新竹、台南、大陸上海、美國、新加坡		

獲利能力(106第2季)		最新四季每股盈餘		最近四年每股盈餘	
營業毛利率	50.85%	106第2季	2.56元	105年	12.89元
營業利益率	38.93%	106第1季	3.38元	104年	11.82元
稅前淨利率	40.27%	105第4季	3.86元	103年	10.18元
資產報酬率	3.42%	105第3季	3.73元	102年	7.26元
股東權益報酬率	4.74%	每股淨值:	51.74元		

除權資料		除息資料	
除權日期	-	除息日期	106/06/26
最後過戶日	-	最後過戶日	106/06/27

技術面

以價格及交易量的關係建立各式指標，作時間序列分析

以台積電為例 (取自XQ操盤高手)

內容分析學派的興起

BUSINESS INSIDER UK — FINANCE

The ECB says Twitter can predict the stock market

The European Central Bank (ECB) just put out an interesting study looking at whether Twitter and Google can be used to predict stock market moves — and its conclusion is, for Twitter, it can.

The ECB says: "Twitter bullishness has a statistically and economically significant

Source: http://uk.businessinsider.com/ecb-twitter-bullishness-stock-market-moves-2015-7

Microsoft Technologies Documentation Resources

DEVEL0PER B10G About Authors

Stock Market Predictions with Natural Language Deep Learning

December 4, 2017 70,219

Overview

We recently worked with a financial services partner to develop a model to predict the future stock market performance of public companies in categories where they invest. The goal was to use select text narrative sections from publicly available earnings release documents to predict and alert their analysts to investment opportunities and risks. We developed a deep learning model using a one-dimensional convolutional neural network (a

Source: https://www.microsoft.com/developerblog/2017/12/04/predicting-stock-performance-deep-learning/-7

Reference	Text type	Text source	No. of items	Prescheduled	Unstructured
Wuthrich et al. (1998)	General news	The Wall Street Journal, Financial Times, Reuters, Dow Jones, Bloomberg	Not given	No	Yes
Peramunetilleke and Wong (2002)	Financial news	HFDF93 via www.olsen.ch	40 headlines per hour	No	Yes
Pui Cheong Fung et al. (2003)	Company news	Reuters Market 3000 Extra	600,000	No	Yes
Werner and Myrray (2004)	Message postings	Yahoo! Finance, Raging Bull, Wall Street Journal	1.5 million messages	No	Yes
Mittermayer (2004)	Financial news	Not mentioned	6602	No	Yes
Das and Chen (2007)	Message postings	Message boards	145,110 messages	No	Yes
Soni et al. (2007)	Financial news	FT Intelligence (Financial Times online service)	3493	No	Yes
Zhai et al. (2007)	Market-sector news	Australian Financial Review	148 direct company news and 68 indirect ones	No	Yes
Rachlin et al. (2007)	Financial news	Forbes.com, today.reuters.com	Not mentioned	No	Yes
Tetlock et al. (2008)	Financial news	Wall Street Journal, Dow Jones News Service from Factiva news database.	350,000 stories	No	Yes
Mahajan et al. (2008)	Financial news	Not mentioned	700 news articles	No	Yes
Butler and Kešelj (2009)	Annual reports	Company websites	Not mentioned	Yes	Yes
Schumaker and Chen (2009)	Financial news	Yahoo Finance	2800	No	Yes
Li (2010)	Corporate filings	Management's Discussion and Analysis section of 10-K and 10-Q filings from SEC Edgar Web site	13 million forward-looking-statements in 140,000 10-Q and K filings	Yes (company annual report)	Yes
Huang, Liao, Yang, Chang, and Luo (2010) and Huang, Chuang, et al. (2010)	Financial news	Leading electronic newspapers in Taiwan	12,830 headlines	No	Yes
Groth and Muntermann (2011)	Adhoc announcements	Corporate disclosures	423 disclosures	No	Yes
Schumaker et al. (2012)	Financial news	Yahoo! Finance	2802	No	Yes
Lugmayr and Gossen (2012)	Broker newsletters	Brokers	Not available	No	Yes
Yu, Duan, et al. (2013)	Daily conventional and social media	Blogs, forums, news and micro blogs (e.g. Twitter)	52,746 messages	No	Yes
Hagenau et al. (2013)	Corporate announcements and financial news	DGAP, EuroAdhoc	10870 and 3478 respectively	No	Yes
Jin et al. (2013)	General news	Bloomberg	361,782	No	Yes
Chatrath et al. (2014)	Macroeconomic news	Bloomberg	Not mentioned	Yes	No
Bollen and Huina (2011)	Tweets	Twitter	9,853,498	No	Yes
Vu et al. (2012)	Tweets	Twitter	5,001,460	No	Yes

Ref. Nassirtoussi, Arman Khadjeh, et al. "Text mining for market prediction: A systematic review." *Expert Systems with Applications* 41.16 (2014): 7653-7670.

- "Twitter mood predicts the stock market"
  - We find an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error (MAPE) by more than 6%. Bollen, et. al. 2010
- 應用範圍
  - 在資訊不對稱的群眾市場，以人工智慧語意技術，分析消息情報後所做的預測
  - 對台灣上市櫃公司，預測 n 日後漲跌，出手正確率可達多少？

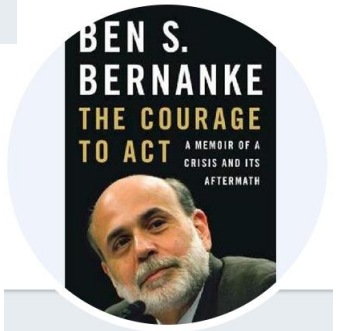
社群大數據 預測原理

基本分析 技術分析 → 內容分析

- ✓ **訊息易由社群傳播**：不論市場訊息、專業訊息、或內部訊息 (inside information)，以社群傳播最為容易，進而全面擴散。
- ✓ **社群傳播速度快**：不需編輯審核，發布速度最快
- ✓ **群眾決定市場**：當市場是由群眾決定時，了解群眾的想法，即可預測市場。例如：群眾傳言將銀行破產，可能造成擠兌而真的破產
- ✓ **消息不論真確度均值得參考**：無論是流言或假消息，皆會影響大眾及市場。

Donald J. Trump
@realDonaldTrump

Elon Musk
@elonmusk



Ben Bernanke
@benbernanke

以短期、特定個股為例

鴻海翻轉成功？夏普今年被估賺122億
新唐人亞太電視台 - 2016年10月19日
【新唐人亞太台2016 年10 月20 日訊】再來關心，鴻海集團戰略投資夏普，轉盈。《日經新聞》報導，夏普今年本業將擺脫三年來虧損，轉為獲 ...

營業利潤轉正？大漲9% 夏普澄清非公司發布之
HiNet 新聞社群 - 2016年10月18日

有些新聞或討論文章出現後，若干日後股價會漲
→收集一批成為【看漲文章】範本

攜手華為偉創力搶走鴻海生意
蘋果日報 - 2016年9月23日
【王郁倫、劉煥彥／台北報導】鴻海（2317）董事長郭台銘表示，網路與硬來相互整合將是趨勢，傳統與創新不是互相對立，鴻海雖然傳統，但不守舊維持 ...

有些新聞或討論文章出現後，若干日後股價會跌
→收集一批成為【看跌文章】範本

日經：鴻海研究在美國生產iPhone的可能性
聯合新聞網 - 2 小時前
美國準總統川普在競選期間高喊「美國第一」，盼製
事長郭台銘17日表示，鴻海會協助美國創造新機會，
家 ...

iPhone 回美製造？日媒：鴻海正在考量
科技新報 TechNews - 1 小時前

iPhone回美生產有譜？傳鴻海正在考慮但
HiNet 新聞社群 - 20 分鐘前

iPhone將大改版鴻海台積樂
聯合財經網 - 11 小時前

川普來了、iPhone變美國製造？傳鴻海評估中、和碩拒絕
MoneyDJ理財網 - 3 小時前


科技新報 Tec...


MoneyDJ理...


鉅亨網財經新...


中時電子報 (...

機器學習法為例

今天新聞和討論文章又這麼多。每篇都用相似分析看看，是比較像**【看漲文章】**、還是**【看跌文章】**。
每篇文章都有幾個選項(漲或跌、或持平)，最後機器一起投票，猜猜數天後股價會漲或跌？

Requirement (1) 基本題

- 各挑選出看漲及看跌的一批文章，從中取出關鍵字列表，建構向量空間
- 參考做法
  1. 用種子關鍵字，如「股價&下跌」、「營收&衰退」等，各挑選一批文章；或用指數或股價漲跌，例如第D+n天與第D天相比，指數或價格下跌超過特定幅度σ，則視第D天的文章合為一批看跌文件集；看漲文件集的做法類似。
  2. 類似作業1，從這兩批文章中找出具鑑別力 (扣除共通字詞) 的關鍵字列表，合起來建構向量空間。
     - n及σ為實驗參數；可自行設計或應用其他技巧。

Requirement (2) 基本題

- 將前述兩批文章作為訓練資料及測試資料，使用監督式學習之分類演算法，評估分類模型之準確率
- 參考做法
  1. 將前述兩批文章分為訓練資料及測試資料 (例如80%及20%)
  2. 在向量空間中，以kNN為例，以每篇測試文章挑出最相似 (向量夾角最小) 的5篇訓練文章，依這5篇為漲或跌之數量進行投票，預測該測試文章歸為看漲或看跌
  3. 由測試資料評估分類模型準確率 (以confusion matrix呈現)
  - 可自行替換為NB、SVM、DT等其他分類演算法

結果範例

- 實驗參數：以 xxx 演算法實作，參數如下...
- 訓練資料中標記為漲的共 a 篇、標記為跌的共 b 篇
- 測試資料中標記為漲的共 c 篇、標記為跌的共 d 篇
- 測試資料中的分布如下，分類準確率為...

	真實為漲	真實為跌
預測為漲		
預測為跌		

Requirement (3) 加分題

- 判斷 n 日後指數或股價歸類為看漲或看跌，進行移動回測
- 參考做法
  - 在36個月資料中，每次取3個月資料建立需求(1)及需求(2)的模型
  - 用該模型預測第3+1個月：於該月中依第D日之相關文章歸類為看漲或看跌的篇數，預測第D+n日為看漲或看跌；若篇數過於接近則不判別 (不出手)，紀錄出手次數、以及預測漲或跌的準確率
  - 往後移動1個月，重複以上步驟。最後評估總出手率及預測漲或跌的總準確率 (以confusion matrix呈現)

採移動式訓練及回測，每次移動一個月

1月	2月	3月	4月	5月	6月	…	…	…	…
訓練資料			測試資料						

1月	2月	3月	4月	5月	6月	…	…	…	…
	訓練資料			測試資料					

例如計算出手率 50%、準確率 66%

	4/1	4/2	4/3	4/4	4/5	4/6
預測	漲	跌	X	漲	X	X
n 日後真實	漲	跌	漲	跌	跌	漲

分布狀況如右

	真實為漲	真實為跌
預測為漲	1	1
預測為跌	0	1

Datasets 資料集

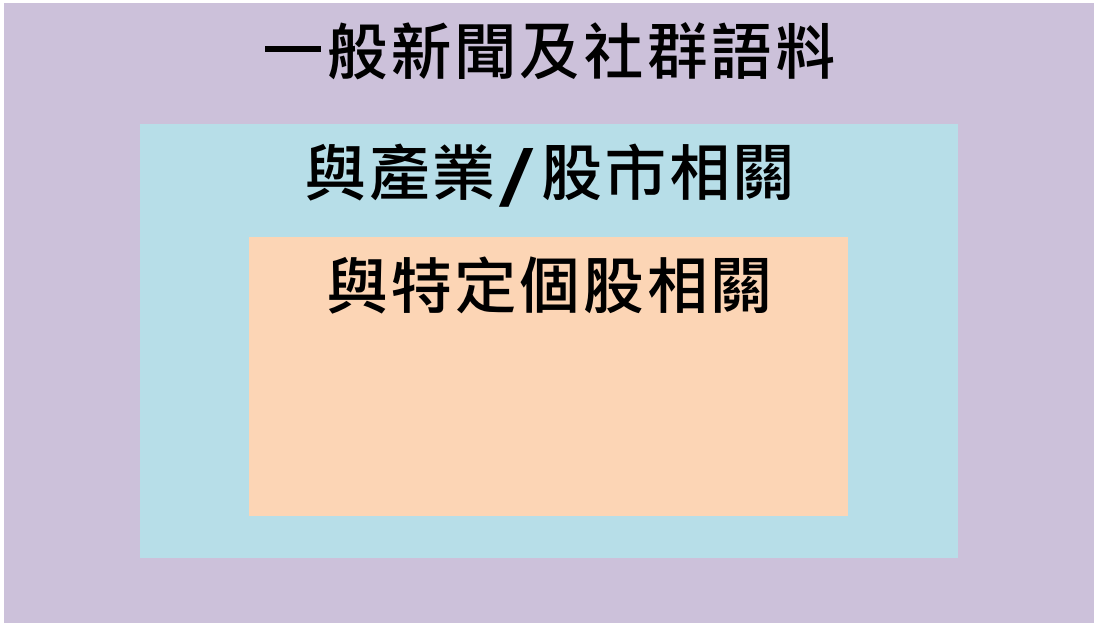
- 資料集
  - 2016~2018 網路公開之新聞、論壇、BBS、股市價量交易資訊
  - 自行過濾部分文章 (例如日常例行發文、過短內容) 及進行前處理
  - 挑選研究之公司或產業
    - 上市或上櫃之公司、類股指數、或大盤指數
    - 個股建議挑選股價活潑、討論量大者
- 資料特性
  - 包括了結構性及非結構性的資料
  - 連續資料，帶有時間性，可能與過往相關
  - 亦可自行結合其他資料或統計技巧進行實驗比較 (加分)

Deliverables

- 分組展示
  - 不限程式語言與演算法
  - 2019.05.05 繳交，於隔日展示
    - 需繳簡報檔(尾附影片連結)，另錄製八分鐘內的說明影片，解說成果及過程。影片中若能以程式化處理並實際執行 (live demo) 者加分。
    - 將簡報檔、系統擷圖、程式碼打包壓縮zip繳交 (勿附資料)

範例：以機器學習方式決定詞彙集

- 取出能反應股價漲跌的特徵詞，建立向量空間
- 使用n-gram法，實驗多種指標、多層次的語料集來挑選出特定用詞



股市用詞結果範例

詞	TF-DF卡方
台股	85054.0044
下跌	76447.66918
指數	55572.02061
股市	42265.127
損失	39626.62603
震盪	39387.04353
收購	36676.13536
股價	35145.0833
早盤	33473.65877
外資	32792.40289
跌幅	31312.76858
市場	31254.00656
上漲	29087.65011
虧損	29054.75353