大數據與商業分析 – 金融業為例

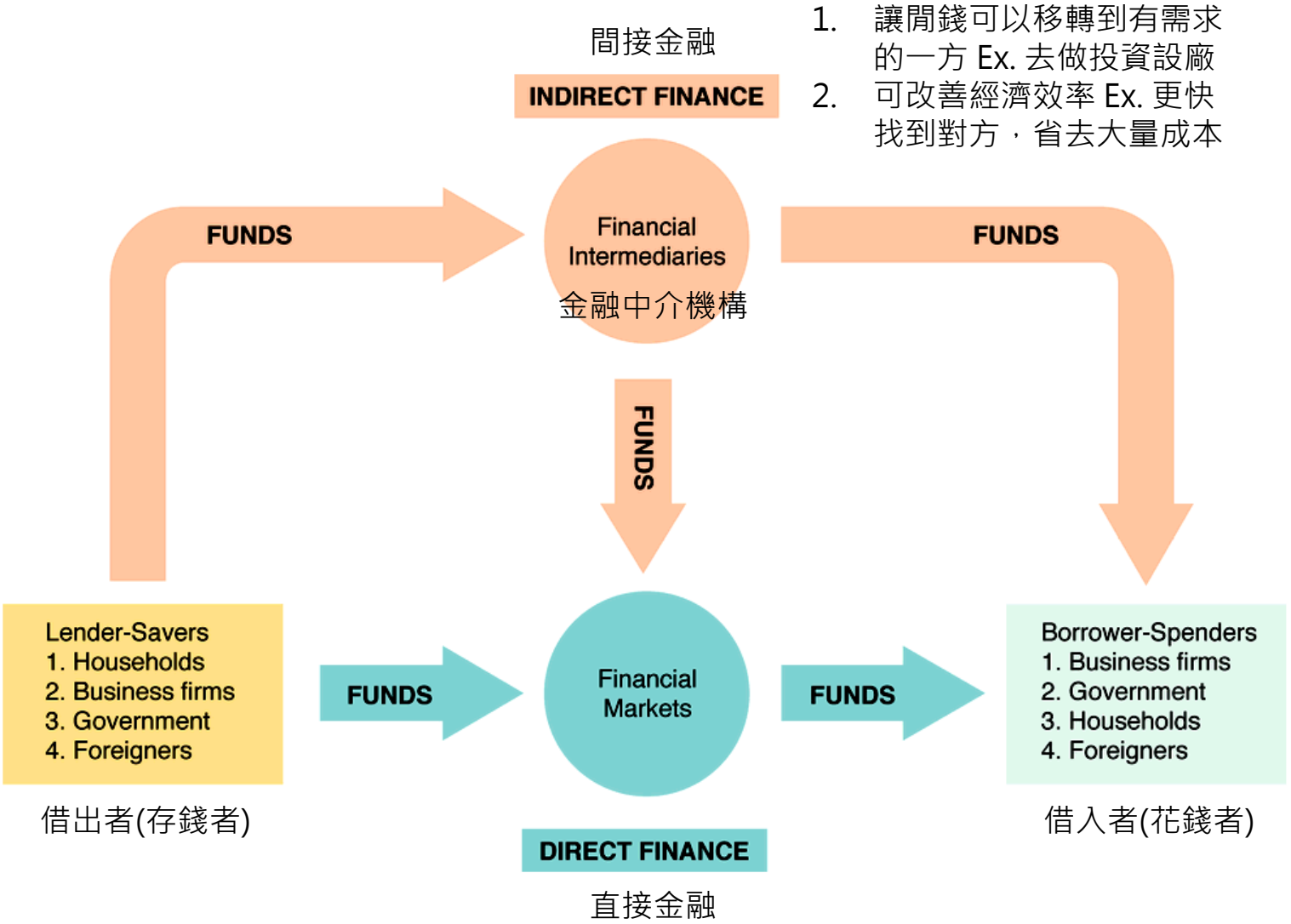
簡介金融市場、機構、商品、及金融數據

楊立偉教授

References:
Chapter 2, The Economics of Money, Banking and Financial Markets,
Frederic Mishkin, Pearson (12版/2019)
中譯本: 貨幣銀行學, Frederic Mishkin, 陳思寬, 華泰文化 (11版/2017)

Function of Financial Markets
金融市場的功能

1. 讓閒錢可以移轉到有需求的一方 Ex. 去做投資設廠
2. 可改善經濟效率 Ex. 更快找到對方，省去大量成本



Classifications of Financial Markets
金融市場的分類

1. Debt Markets 債務市場: 借貸、票券 (會還的)
   通常可依短期 (少於1年) 或長期做區分，可再分為
   Money Market 資金市場 或 Capital Market 資本市場
2. Equity Markets 權益市場: 投資 (不還的)
   Stock Market 股票市場

1. Primary Market 初級市場
   首次發行賣出的證券 (股票、債券等)
2. Secondary Market 次級市場
   之前發行過的再次買或賣

1. Exchanges 交易所
   在集中在一起交易買賣的地方 Ex. 紐約證交所
2. Over-the-Counter (OTC) Markets 櫃檯買賣
   不同地區經銷商自己做的買賣 (通常規模比交易所小)

Financial Intermediaries
認識金融中介機構

TABLE 2-2 Primary Assets and Liabilities of Financial Intermediaries

Type of Intermediary	Primary Liabilities (Sources of Funds) 資金來源	Primary Assets (Uses of Funds) 資金用途
存款機構 **Depository institutions (banks)**		
特許銀行 Chartered banks	Deposits	Loans, mortgages, government bonds
信託/抵押/借貸公司 Trust and mortgage loan companies	Deposits	Mortgages
信用合作社 Credit unions and *caisses populaires*	Deposits	Mortgages
契約性儲蓄機構 **Contractual savings institutions**	保單之保金	
人壽保險公司 Life insurance companies	Premiums from policies	Corporate bonds and mortgages
財產及損失保險公司 Property and Casualty P and C insurance companies	Premiums from policies	Corporate bonds and stocks
退休基金 Pension funds	Retirement contributions 退休提撥	Corporate bonds and stocks
投資機構 **Investment Intermediaries**		
財務公司 Finance companies	Finance paper, stock, bonds	Consumer and business loans
共同基金 Mutual funds	Shares	Stocks and bonds
資金市場共同基金 Money market mutual funds	Shares	Money market instruments

Financial Instruments
金融工具

- Financial Instruments: The written legal obligation of one party to transfer something of value, usually money, to another party at some future date, under certain conditions.
- 金融工具又稱信用工具、交易工具等，是合約雙方之間的有價契約。它們可以被創建、交易、修改和結算。它們可以是現金、有價證券、所有權憑證 (股份)，或是收取或交付某種權利 (改寫自wikipedia)
    - Ex. Ethereum上的智能合約或NFT也都算金融工具

Uses of Financial Instruments: 3 functions
金融工具的三個功能

- Financial instruments act as a means of payment
  - 類似現金，可以拿來支付，例如員工拿認股權當作薪資的一部分
- Financial instruments act as stores of value
  - 類似現金，是一種未來的購買力
  - 擺著可能會增值
- Financial instruments allow for the transfer of risk.
  - 這點比較不像現金，例如期貨和保險是一種移轉風險的合約.

Underlying vs Derivative Instruments

- **兩種基本的金融工具: 標的工具及衍生工具**
  - **Underlying instruments: 借出者(存錢者)直接對借入者(花錢者)**
    - Ex. 股票及債券
    - 這種通常移轉資源比較有效率.
  - **Derivative instruments: 將標的工具的價值或報酬近一步切割**
    - Ex. futures 期貨, options 選擇權
      - 例如以保證金60萬買下個月台股指數20000點, 或買入以600元買台積電1000股的權利
    - 通常主要用來移轉風險
      - 需要有對應的一方

金融數據

- 金融市場、金融機構、監管機關等，時時刻刻都產生許多金融數據
  - 與主體相關，例如客戶 (消費者及企業)、商品、交易等
  - 包括了結構性及非結構性的資料
  - 包括了個體資料及總體資料
  - 與時間相關，帶有時間性，可能與過往相關

金融業運用數據現況

借款、逾期、催收、呆帳、債務轉讓、授信保證人、退票、拒往、信用卡等

聯徵中心

公司基本資料、財稅資料、公發公司資料、金融機構資料、法院及主計資料等

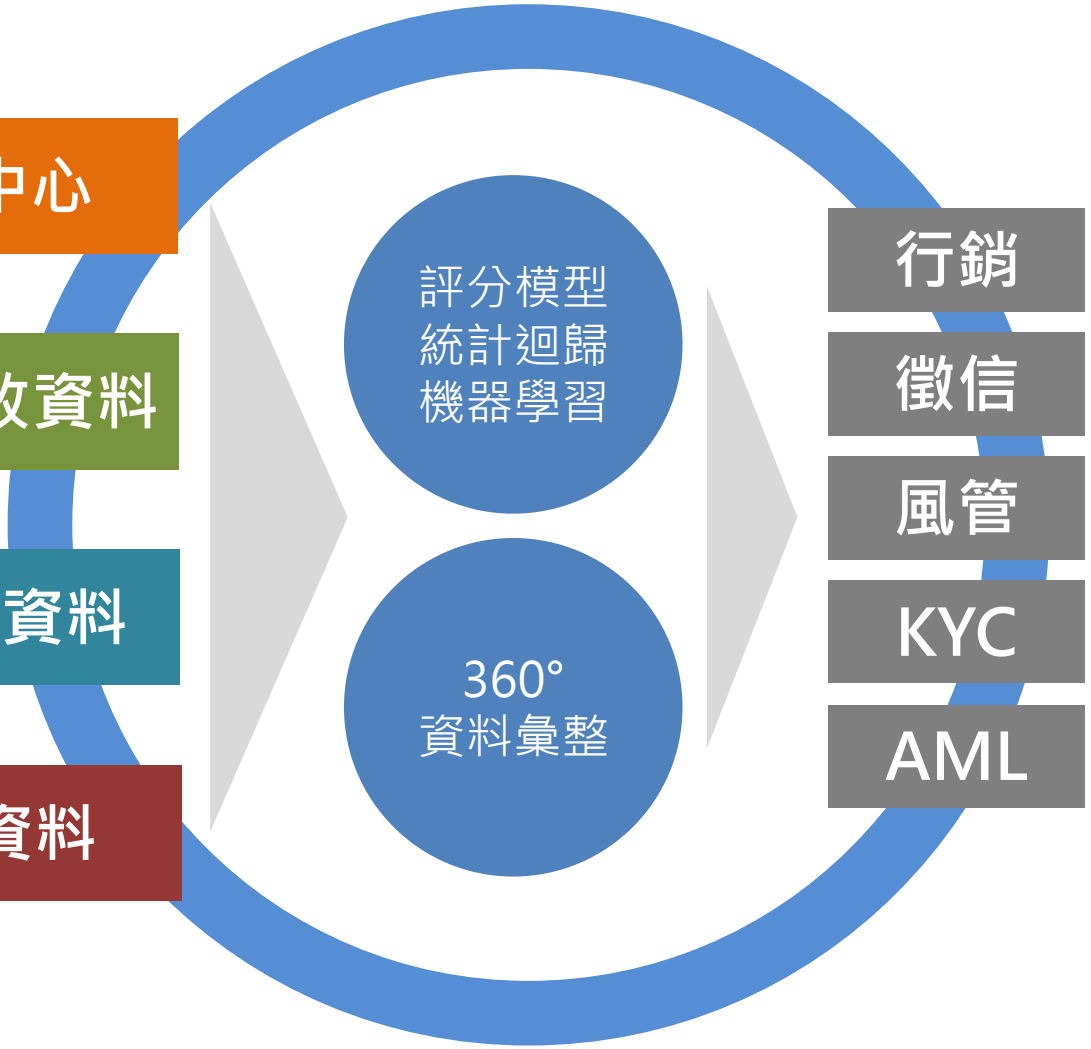
政府開放資料

社群資料、人群資料、電信資料、市場資料、調研資料、徵信資料、醫療資料等

第三方資料

客戶基本資料、交易資料、往來互動紀錄、客服資料、行為資料

自有資料



Project : from Classification to Prediction

楊立偉教授

wyang@ntu.edu.tw

專題：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、股東權益報酬率等
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？

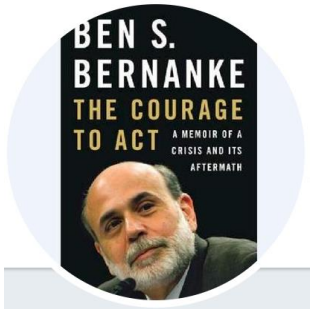
社群大數據 預測原理

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)

- 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。參考做法如下
  1. 用種子關鍵字，如 (股價&下跌)|(營收&衰退)、(股價&上漲)|(營收&成長) 等，各挑選一批文章 (不建議此方法)
  2. 或用自編碼的方式，以特定股價或指數漲跌，例如第D+n天與第D天相比，股價或指數下跌超過特定幅度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。
  3. 類似作業2，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 (若有1000詞即1000維之空間)
  - n及σ為實驗參數 (如3天及5%)；亦可自行設計或應用其他技巧。

Requirement (2)

- 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。參考做法如下
  1. 將前述看漲及看跌兩批文章合起來後，隨機打散分為訓練資料及測試資料 (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篇分別來自看漲或看跌文章集之數量進行投票，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：以〇〇 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... (i.e. confusion matrix)

	預測為漲	預測為跌
真實為漲		
真實為跌		

Requirement (3) 移動回測

- 判斷 n 日後指數或股價歸類為看漲或看跌，進行移動回測
  - 參考做法如下
    - 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)

* k-fold cross validation的方式，將資料隨機分成訓練及測試的作法，有"偷看答案"之嫌，因此時序類型的資料，應該模擬隨著時間"重播"，測試每一時段的預測結果，再一起評估模型準確率

採移動式訓練及回測，每次移動一個月 亦可自行實驗不同的移動月份長度

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					

計算出手率 50%、準確率 75%

	4/1	4/2	4/3	4/4	4/5	4/6	4/7	4/8
預測	漲	跌	X	漲	X	X	漲	X
n 日後真實	漲	跌	漲	跌	跌	漲	漲	漲

分布狀況如右

	真實為漲	真實為跌
預測為漲	2	1
預測為跌	0	1

Datasets 資料集

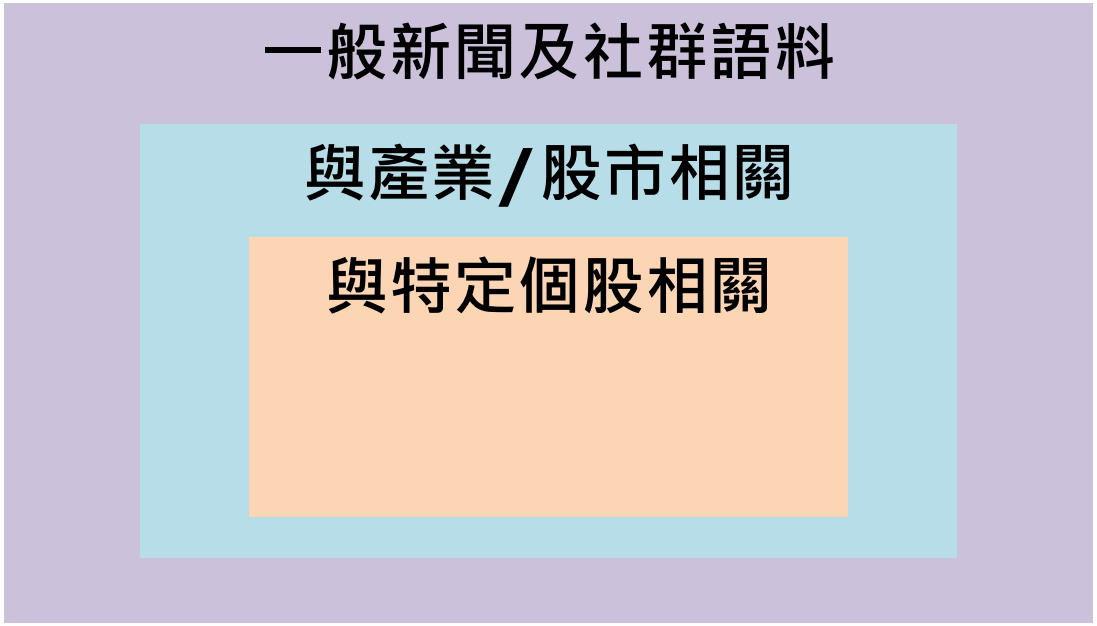
- **資料集 (下載)** 1.04GB Zip file
  - **2019~迄今網路公開之新聞、論壇、BBS、股市價量交易資訊**
  - **自行過濾部分文章 (例如日常例行發文、過短內容) 及進行前處理**
  - **挑選研究之公司或產業**
    - 若干上市或上櫃之公司、類股指數 (如全電子股)、或大盤指數 (或ETF如台灣50等)
    - 若挑選個股，建議挑選股價活潑、討論量大者
- **資料特性**
  - **包括了結構性及非結構性的資料**
  - **連續資料，帶有時間性，可能與過往相關**
  - **亦可自行結合其他資料或統計技巧進行實驗比較 (加分)**

Deliverables

- 分組展示
  - 不限程式語言與演算法
  - 於報告前 3 天繳交 (由助教公告)
    - 每組需繳簡報檔 (尾附影片連結)，另錄製10分鐘內的說明影片，解說成果及過程。影片中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 (限100M以下，勿附影片及資料)
    - 於報告當周挑選八組上台解說

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
- 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

延伸自學

- 可以比較看看下列處理單位的差異
  - BOW+機器學習方法 : 以斷詞或 n-gram 作分析
  - 詞向量+深度學習方法 : 以 character 或 subword 做分析，並使用預訓練模型
  - 其他方法

附錄

將資料匯入MySQL資料庫的參考表格結構

```
CREATE TABLE `mid_text` (
        `no` int NOT NULL AUTO_INCREMENT,
        `id` varchar(32),
        `p_type` varchar(10),
        `s_name` varchar(40),
        `s_area_name` varchar(40),
        `post_time` datetime,
        `title` text,
        `author` text,
        `content` longtext,
        `page_url` longtext,
        `content_type` varchar(40),
        `comment_count` int,
        `sentiment` varchar(1),
        PRIMARY KEY (`no`),
        KEY `idx` (`post_time`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_unicode_ci;
```