南山人壽

# 投資組合之新聞推薦智慧助理

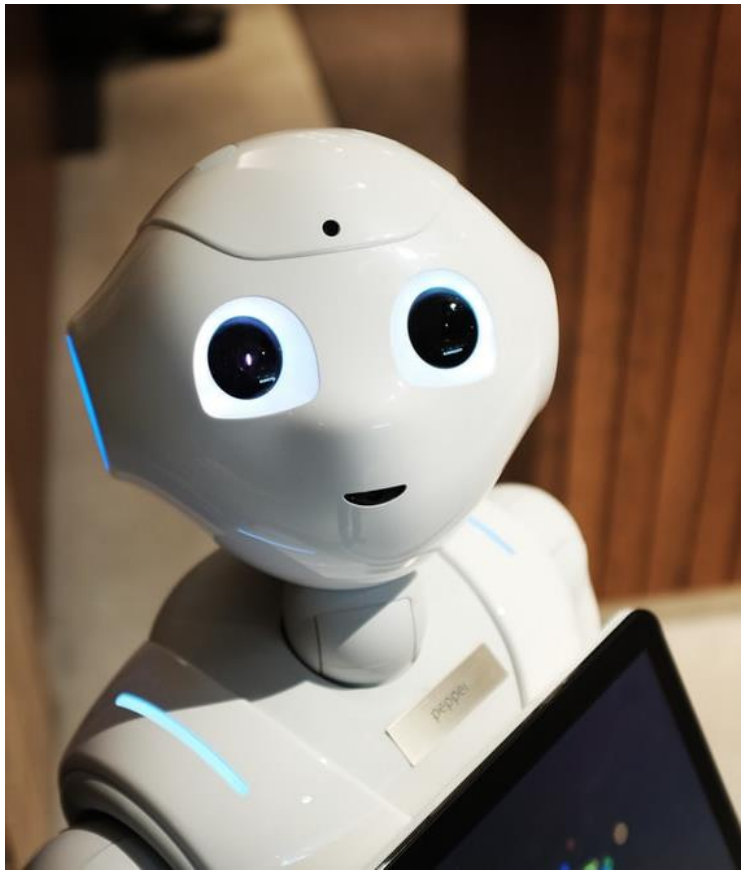
台大：蕭湧 台大：陳家偉 台大：劉柏彥 台大：吳悅寧

台大：吳育嘉 台大:吳天友 政大: 楊瑾容 政大: 楊承羲

指導老師: 石百達教授

Mentor: 張明淇學長

# 智慧助理



新聞

CNBC

Bloomberg

THE WALL STREET JOURNAL. WSJ

MarketWatch

REUTERS®

TIME

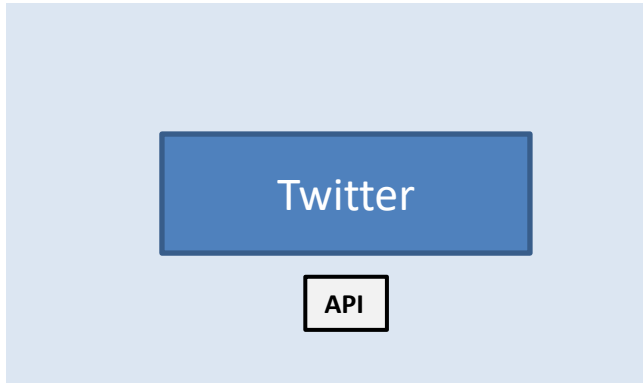
社群

# 解決使用者痛點

1

# 解決使用者痛點 – 蒐集資料

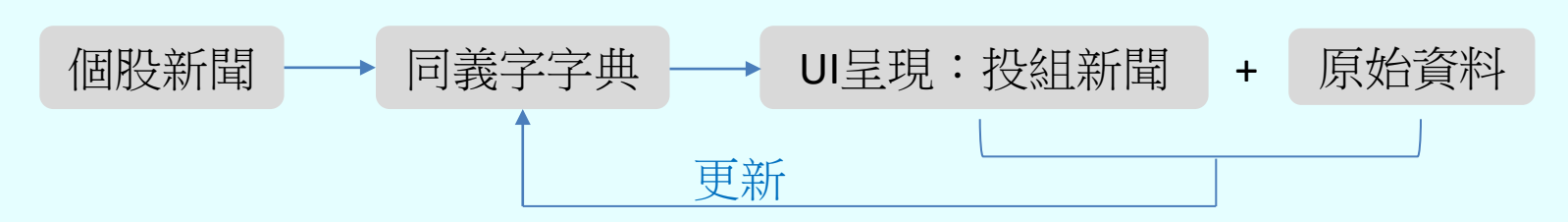
- 蒐集資料總是花太多時間？
  - 利用程式每日自動用API、RSS至各大網站抓取新聞，以及推文
  - 自2018年1月到2020年5月底，總共蒐集約10萬筆新聞資料


# 解決使用者痛點 – 投組新聞

- 手中的投組，如何尋找相關資訊(公司名稱同義字太多，難搜尋)？

➢透過投組新聞, 馬上能找出與投組中公司相關的新聞



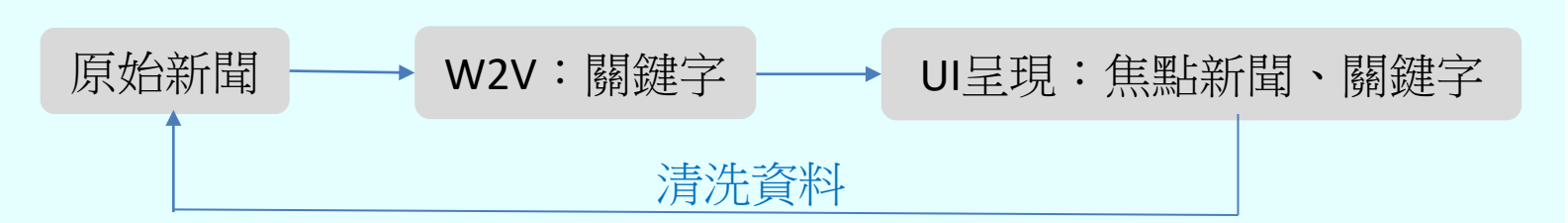

- 人工查詢個股新聞的關鍵字中，有出現那些該公司的其他名稱，並且加入同義字字典中
- 藉由UI呈現的資料中，跟原始資料比對，將一些遺漏的同義字更新進字典

# 解決使用者痛點 – 焦點新聞

- 每天那麼多新聞，到底有哪幾則是真的重要的？

➢透過焦點新聞, 馬上能找出最近焦點的主題是什麼, 及包含哪些新聞

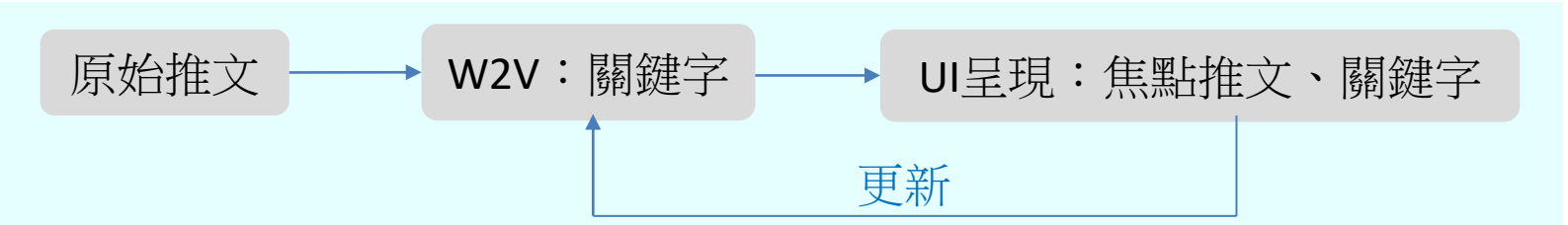
➢甚至能透過焦點新聞判斷投組表現與大盤相關性如何




- 將該日的前兩周所有新聞標題，利用W2V建立模型，量化每個字的重要性係數，取出前三個重要性最強的關鍵字挑選焦點新聞
- 根據UI呈現的關鍵字，判斷哪些字是不重要的關鍵字，重新清洗資料、訓練模型

# 解決使用者痛點 - Twitter

- 除了正式的新聞資料，若我們也想同時知道社群上的資訊？
  - 透過焦點推文、人氣推文，便可知道目前最受矚目的推文關鍵字，以及最近一些名人的推文




- 將該日的前兩周所有推文內容，利用W2V建立模型，量化每個字的重要性係數，取出前四個重要性最強的關鍵字挑選焦點推文
- 根據UI呈現的關鍵字，判斷哪些字是不重要的關鍵字，重新清洗資料，訓練模型

問題：更多口語化的字需篩選

# 結論

2

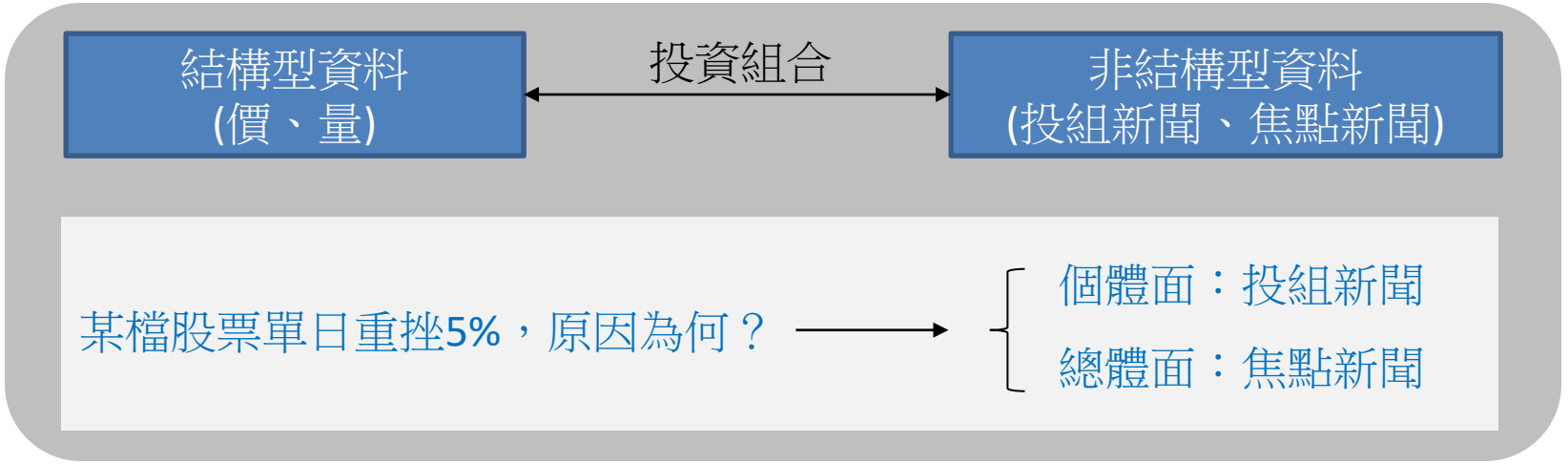
# 資料處理的重要性

- 資料處理才是決定演算法準確性的重要因素，因此我們在上個部分提到，在處理投組新聞、焦點新聞及Twitter時，我們花了很多時間在人工處理的部分


馬雲：通常是沒有數據的公司才會談論AI。

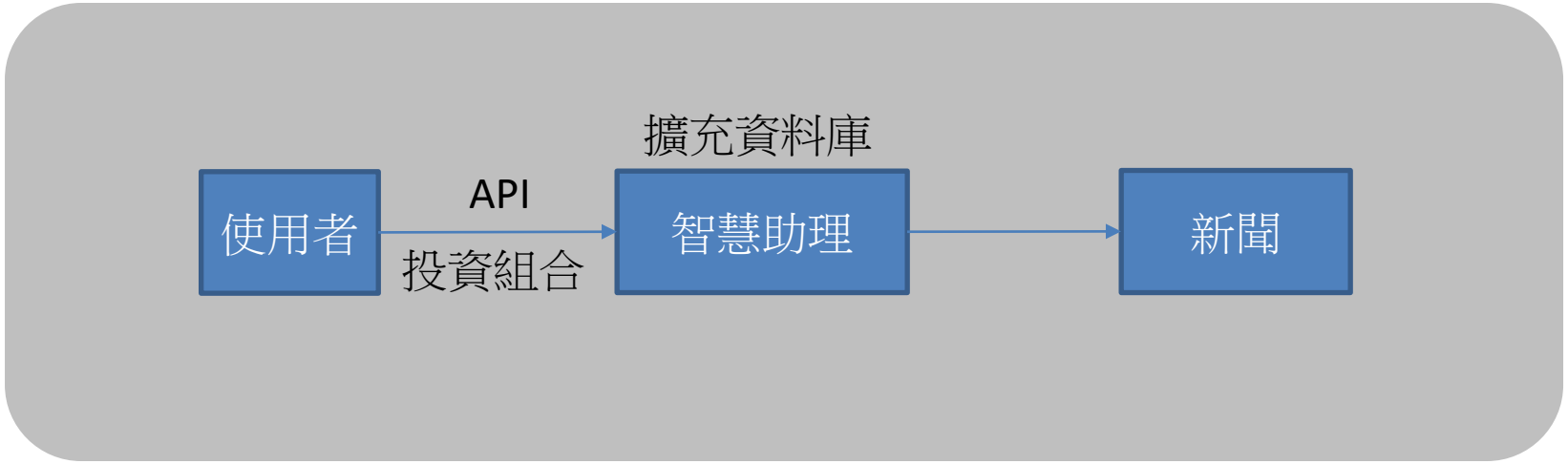
# 價值－結合結構型、非結構型資料

- 傳統的文字探勘，往往只鑽研純粹文字部分所產生的結果
- 智慧助理透過投資組合做橋梁，將價、量等結構型資料結合文字做呈現、研究



# 價值－非封閉式系統

- 只要能提供每日的投組清單(如API等方式串接)，即可利用該投組進行篩選新聞
- 未來若要擴充到台股，就將標的的資料庫擴充即可



# 價值 – 時間序列的關鍵字

- 先前在篩選每天焦點新聞的關鍵主題時，我們都有將每個關鍵字的重要性紀錄下來，未來若要研究關鍵字隨時間的變化時，將可以繼續延伸

- 如下圖，3/15~3/20間，雖然關鍵字前兩名都是”CORONAVIRUS”, “FED”, 但可以發現前者的重要性正在下降，且後者重要性不斷在上升。可能代表社會在關注新冠病毒的同時，也越來越關注FED的決策

```
keyword['20200315']

[('CORONAVIRUS', 0.9849397590361446),
 ('FED', 0.713855421686747),
 ('VIRUS', 0.608433734939759),
 ('BIDEN', 0.5843373493975904),
 ('SANDERS', 0.5481927710843374),
 ('CEO', 0.4307228915662651),
 ('HOUSE', 0.4246987951807229),
 ('OUTBREAK', 0.3885542168674699),
 ('TRUMP', 0.3704819277108434),
 ('BANK', 0.26506024096385544),
 ('OIL', 0.2469879518072289),
 ('BUY', 0.21987951807228914),
 ('BUSINESS', 0.20180722891566266),
 ('CUTS', 0.19578313253012047),
 ('TECH', 0.1686746987951807),
 ('COURT', 0.16265060240963855),
 ('AHEAD', 0.1566265060240964),
 ('ITALY', 0.14759036144578314),
 ('DEMOCRATIC', 0.12349397590361445),
```

```
keyword['20200320']

[('CORONAVIRUS', 0.9581151832460733),
 ('FED', 0.7931937172774869),
 ('TRUMP', 0.5209424083769634),
 ('ITALY', 0.450261780104712),
 ('CEO', 0.42670157068062825),
 ('HOUSE', 0.3795811518324607),
 ('VIRUS', 0.34554973821989526),
 ('BIDEN', 0.32722513089005234),
 ('DOWN', 0.31151832460732987),
 ('OIL', 0.3089005235602094),
 ('BUSINESS', 0.27486910994764396),
 ('BANKS', 0.2094240837696335),
 ('OUTBREAK', 0.20680628272251309),
 ('COURT', 0.20157068062827224),
 ('BANK', 0.18848167539267016),
 ('ECONOMIC', 0.18324607329842932),
 ('SPORTS', 0.18324607329842932),
 ('WORK', 0.17539267015706805),
 ('COMPANIES', 0.17015706806282724),
 ('PANDEMIC', 0.14136125654450263),
 ('CLOSE', 0.12827225130890052),
```

# 未來延伸 – 客製化推薦助理

- 在使用者操作推薦助理時，後台端會記錄使用者的點擊、搜尋紀錄，讓智慧助理可以進一步優化推薦的演算法，進而推薦更適合該使用者的新聞

- 若使用者為分析師，我們更可以從搜尋紀錄、點擊紀錄中試著尋找出為何該分析師會點擊該則新聞、為何搜尋該關鍵字，進而使我們的推薦助理逐漸訓練成為新聞的分析師

使用者紀錄


```
bruce:
▼ 0:
    date: "05/05/2020"
    pf: "pph_2"
    kw: "baseball"
▼ 1:
    date: ""
    pf: ""
    kw: ""
  ▼ clc:
      url: "https://time.com/5832060/korean-baseball-coronavirus/"
      title: "Korea’s Baseball Season Begins After Remarkable Turnaround in Coronavirus Cases"
      tab: "1"
      click_time: "20200627 15:15:11"
```


使用者名稱

搜尋紀錄(日期、投組、關鍵字)

點擊紀錄(新聞標題、點擊的時間)

# Thanks