大數據與商業分析

用 AI 及社群數據協助投資決策

—以新聞文章判斷 「0050 台灣 50」 股價之漲跌

大綱

01 將全部文章加上漲跌之標籤

02 依標籤挑選出看漲看跌文章集

03 將文章集切詞並建構出
看漲及看跌之關鍵字列表

04 建構不同模型並比較各自的
confusion matrix 與正確率

05 移動回測並計算正確率

大綱

01 將全部文章加上漲跌之標籤

02 依標籤挑選出看漲看跌文章集

03 將文章集切詞並建構出
看漲及看跌之關鍵字列表

04 建構不同模型並比較各自的
confusion matrix 與正確率

05 移動回測並計算正確率

判斷漲跌

方法

- ✓ 判斷第n天文章集為漲或跌
- ✓ 利用第n+1天到第n+3天股價分別去跟第n天股價做比較
- ✓ 再用三天中漲跌較多的數量去決定第n天之漲跌

股價

台灣50

舉例

將1/24之文章歸類為看漲文章集

2018/1/24	8,460.85

日期	股價	漲跌(與1/24比較)
2018/1/25	8,487.00	漲
2018/1/26	8,458.96	跌
2018/1/29	8,527.11	漲

結果

年月日	收盤價(元)	漲跌
2018/1/2	8,032.98	1
2018/1/3	8,107.28	1
2018/1/4	8,131.55	1
2018/1/5	8,148.79	1
2018/1/8	8,193.17	0
2018/1/9	8,199.95	0
2018/1/10	8,127.25	1
2018/1/11	8,099.62	1
2018/1/12	8,164.68	1
2018/1/15	8,246.26	1
2018/1/16	8,264.20	1

大綱

01 將全部文章加上漲跌之標籤

02 依標籤挑選出看漲看跌文章集

03 將文章集切詞並建構出
看漲及看跌之關鍵字列表

04 建構不同模型並比較各自的
confusion matrix 與正確率

05 移動回測並計算正確率

依據當天漲/跌標記將文章分為兩個檔案

結果：

跌

2018/01/08	焦點股：營收走弱＋外資砍目標價，大立光(3008)股價回測3800元
2018/01/08	AI、雲端帶動！DRAM供給上半年續吃緊、Q1報價走揚
2018/01/08	《SIMEX台股期貨》09:30，1月台股指數期貨跌0.2點，為401.6點
2018/01/08	《TX台股期貨》09:30，01月台指為10854，跌4；成交36729口
2018/01/08	UNIQLO日本同店銷售額連4揚、創1年5個月來最大增幅
2018/01/08	《人幣匯率》中間價8日報6.4832元 升值83基點
2018/01/08	焦點股：積極搶攻國艦國造與離岸風電，今年營運現轉機，台船盤中亮燈漲停
2018/01/08	《SIMEX台股期貨》09:45，1月台股指數期貨漲1.1點，為402.9點
2018/01/08	《TX台股期貨》09:45，01月台指為10879，漲21；成交45229口
2018/01/08	焦點股：新舊代理商交接導致12月業績重挫，類比科(3438)股價大跌
2018/01/08	《新聞分析》排除4大障礙，聯廣擬重啟上市承銷程序

漲

2018/01/02	《匯市》台幣強升，續創逾4年新高
2018/01/02	《社會》今年首颱最快今天生成，週四提前變天
2018/01/02	《SIMEX台股期貨》10:45，1月台股指數期貨漲1.9點，為394.9點
2018/01/02	《TX台股期貨》10:45，01月台指為10673，漲40；成交63273口
2018/01/02	《電子零件》今年EPS估逾4元，日電貿強漲
2018/01/02	陸推銀行卡境外取現新規 每人年度上限10萬人民幣
2018/01/02	【台股盤中】開紅盤台積電領軍 越過季線
2018/01/02	焦點股：OLED材料商機大，達運(6120)盤中奔漲停創7年新高
2018/01/02	《SIMEX台股期貨》11:00，1月台股指數期貨漲2點，為395點

大綱

01 將全部文章加上漲跌之標籤

02 依標籤挑選出看漲看跌文章集

03 將文章集切詞並建構出
看漲及看跌之關鍵字列表

04 建構不同模型並比較各自的
confusion matrix 與正確率

05 移動回測並計算正確率

利用jieba切詞並篩選關鍵字

先去除數字、英文及stop word

用jieba切詞 並取tf-idf較大之詞

刪去看漲關鍵字及看跌關鍵字中相同且tfidf相近之關鍵字

```python
import csv
import re
import jieba.analyse

down_words = ''
with open("/Users/wupeiyu/Desktop/大數據/bda2020_midterm_data/down2三天.csv", newline='') as csvfile:
    rows = csv.reader(csvfile)
    for row in rows:
        down_words += row[1] * 5
        down_words += row[2]
down_words = re.sub(r'[^\w]', '', down_words)
down_words = re.sub(r'[A-za-z0-9]', '', down_words)

up_words = ''
with open("/Users/wupeiyu/Desktop/大數據/bda2020_midterm_data/up2三天.csv", newline='') as csvfile:
    rows = csv.reader(csvfile)
    for row in rows:
        up_words += row[1] * 5
        up_words += row[2]
up_words = re.sub(r'[^\w]', '', up_words)
up_words = re.sub(r'[A-za-z0-9]', '', up_words)

jieba.analyse.set_stop_words('stopword.txt')
up_tags = jieba.analyse.extract_tags(up_words, topK=1500, withWeight=True)
down_tags = jieba.analyse.extract_tags(down_words, topK=1500, withWeight=True)

both = []
up_remove = []
down_remove = []
for up in up_tags:
    for down in down_tags:
        if up[0] == down[0]:
            if max(up[1], down[1]) < 2 * min(up[1], down[1]):
                up_remove.append(up)
                down_remove.append(down)
            else:
                if up[1] < down[1]:
                    up_remove.append(up)
                else:
                    down_remove.append(down)

for i in both:
    up_tags.remove(i)
```

Key Words

- 上漲關鍵字

up.csv

市場,中國,美國,指數,億元,美元,成長,公司,基金,營收,債券,投資,金額,台灣,年月日,企業,經濟,營業,台幣,資訊,董事,適用,報導,相關,影響,價格,單位,發行,預期,全球,法人,交易,發展,表示,持續,事項,第季,產品,未來,貿易,新聞,收益,以來,股價,累積,顯示,今年,銀行,億萬元,記者,觀測,公告,公開,新興,台北,財務,大陸,認為,電子,認購,台股,每股,預計,來源,日期,指出,外資,進行,訊息,組合,減少,客戶,超過,損益,產業,收盤,蘋果,預估,時間,證券,獲利,技術,名稱,資產,決定,萬元,金融,累計,貨幣,投信,報告,目前,國際,集團,整體,表現,投資人,處分,決議,股份,應敘明,機會,增加,資料,增長,明會,時報,銷售,機構,汽車,本次,建議,綜合,可望,現金,服務,營運,相對,透過,問題,出現,股票,目標,召開,數據,帶動,科技,淨值,漲幅,新台幣,下跌,關稅,需求,相較,實發,工業,反彈,中央社,有限公司,因應,手機,媒體,應用,資金,新高,合計,國家,設備,之關,亞洲,趨勢,去年,漲點,去年同期,辦理,期間,今日,地區,連續,計畫,為止,美股,利率,一個,風險,終場,股數,分別,成交,數量,參考,族群,智慧,股東,調整,重大,生日,幣元,項目,股利,已經,走勢,成為,特別,歐洲,合作,近期,子公司,生產,穩定,進入,普通股,期貨,獨立,財報,領域,規定,日本,第二季,策略,調查,台積,開發,內容,上半年,進口,醫療,上市,規模,當日,工商,總經理,平衡,持股,取得,包括,條件,壓力,股市,晶片,美式,價值,年度,編者,推動,配息,統計,分析,自營商,主要,價元,處理,種類,系統,積極,指標,歷史,費用,情況,繼續,比例,公布,成交量,維持,基準,第三季,網路,建設,消費者,權益,今天,總額,申請,貿易戰,受惠,政府,資金貸,保證,明年,淨額,提供,國內,損失,執行,下半年,收入,早盤,支撐,跌幅,平均,一覽表,百分,中心,說明,基礎,升息,被動,狀況,餘額,增資,半導體,政策,主管,環境,類型,多重,類股,消費,相當,比率,情形,創新,最近,解決,利益,美中,部分,股息,實際,總金額,產能,元件,創下,到期,輸入,編輯,商品,提升,業績,供應,增減,品牌,發生,成本,行動,多元,研發,推出,舉辦,措施,活動,一步,明顯,展望,過戶,發布,旗下,債權,水準,運用,事業,強調,出貨,稅前,條款,上櫃,元年,短線,結構,毛利率,股權,強勢,選擇,邀請,報價,新任,移轉,標準,變動,昨日,以上,匯率,三大,震盪,實現,獲得,指期,管理,強勁,委員會,歐盟,財經,方面,商業,宣布,背書,原油,授權,最新,經營,會議,申報,競爭,評估,資本,自行,香港,雙方,整理,發生緣,同期,推薦,採用,平台,呈現,過去,收購,訂單,旺季,優勢,月份,面板,最大,設計,中國大陸,張數,依據,協議,作收,專業,上揚,第四季,母公司,關注,轉換,開盤,重要,價證券,人者,本益比,關鍵,衝擊,同月,業務,下滑,受到,實施,激勵,部位,台北市,英國,空間,超億元,選舉,最高,自結,買賣,準備,方式,預測,除息,監察人,富邦,掛牌,舉行,加權,新科,修正,動能,姓名,導致,總統,數位,第一,認列,一度,傳統,交易日,談判,海外,規劃,紀錄,外匯,意見,前次,六個,環球,盈餘,上漲,報酬,僅供,機關,是否,大盤,變化,組織,帶來,年增率,機器,辦法,資源,代號,交易所,減資,無法,擴大,近年,大漲,契約,停止,擁有,完成,聯準,社會,看好,挑戰,原因,貶值,對象,陸續,印度,挹注,決策,無公開,幣中國,大廠,業者,波動,大幅,訂立

- 下跌關鍵字

down.csv

市場,營收,億元,指數,中國,美國,公司,營業,金額,成長,美元,基金,債券,投資,單位,年月日,損益,台灣,董事,企業,經濟,億萬元,資訊,適用,台幣,相關,第季,發行,影響,價格,報導,交易,預期,法人,事項,發展,累計,全球,財務,表示,持續,觀測,公開,減少,綜合,未來,萬元,以來,公告,新聞,銀行,產品,累積,收益,顯示,股價,每股,今年,貿易,認購,新興,記者,日期,台股,台北,訊息,合計,相較,大陸,電子,預計,來源,資料,認為,增加,名稱,超過,指出,進行,股份,資產,外資,證券,收盤,客戶,國際,預估,決議,報告,時間,金融,組合,產業,決定,應敘明,技術,獲利,增減,處分,明會,蘋果,貨幣,投信,集團,本次,整體,去年同期,實發,目前,表現,投資人,淨額,項目,銷售,費用,相對,透過,汽車,召開,建議,時報,營運,機構,有限公司,服務,增長,之關,出現,機會,數據,現金,目標,下跌,可望,收入,問題,生日,股票,資金貸,損失,資金,漲幅,保證,總額,科技,因應,需求,帶動,為止,地區,工業,淨值,中央社,子公司,利益,應用,權益,新台幣,期間,手機,繼續,利率,媒體,重大,新高,反彈,設備,連續,數量,國家,風險,亞洲,趨勢,調整,背書,今日,辦理,計畫,股數,普通股,參考,明年,美股,終場,股東,漲點,分別,智慧,規定,一個,同月,成本,獨立,近期,美式,穩定,成交,去年,第三季,餘額,條件,走勢,取得,歐洲,稅前,特別,族群,幣元,上市,總經理,壓力,本期,生產,已經,進入,持股,同期,調查,成為,合作,淨利淨損,日本,開發,台積,執行,處理,主管,股權,升息,領域,價值,公布,股市,期貨,策略,關稅,內容,規模,工商,比例,系統,包括,主要,情況,積極,維持,當日,百分,編者,分析,推動,種類,自結,新任,統計,跌幅,比率,財報,情形,晶片,自營商,平衡,歷史,輸入,基準,股利,提供,發生,網路,進口,醫療,說明,狀況,採用,指標,變動,實際,今天,一覽表,實現,消費者,價元,申請,政府,配息,年度,受惠,創新,當月,建設,國內,基礎,增資,消費,平均,成交量,母公司,總金額,支撐,半導體,多重,品牌,早盤,政策,千元,中心,貿易戰,活動,條款,事業,環境,最近,產能,百分比,類股,研發,到期,上半年,上櫃,移轉,認列,被動,元年,水準,毛利,計師,淨利,美中,舉辦,類型,編輯,相當,第二季,部分,創下,提升,管理,展望,委員會,債權,運用,本年,股息,商業,推出,業績,旗下,重分類,會議,出貨,供應,解決,結構,毛利率,明顯,一步,元件,發布,第四季,原油,收購,姓名,盈餘,措施,以上,標準,多元,評估,獲得,強調,短線,過戶,下半年,發生緣,授權,行動,匯率,昨日,邀請,依據,資本,財經,辦法,歸屬,人者,指期,月份,申報,選擇,簡歷,震盪,香港,所得,設計,三大,方面,最新,監察人,重要,選舉,報價,宣布,異動,呈現,支出,訂單,下滑,整理,分類,專業,認股,自行,協議,原因,經營,競爭,作收,商品,業主,轉換,意見,推薦,年增率,強勁,方式,平台,優勢,強勢,契約,前次,交易所,機關,關鍵,預測,價證券,開盤,最大,歐盟,決策,台北市,兌換,張數,受到,轉讓,旺季,預定,本益比,對象,英國,導致,加權,稅費用,大盤,第一,準備,財務業務,海外,掛牌,過去,無公開,衝擊,面板,買賣,融資,民國,擴大,激勵,聯準,一度,規劃,營業外,上揚,新科,印度,中國大陸,動能,是否,實施,關注,報酬,毛損,機器,合資,舉行,超億元,紀錄,減資,權利,迄事,資源,最高,雙方,完成,陸續,新低,空間,為關,數位,變化,高點

大綱

01 將全部文章加上漲跌之標籤

02 依標籤挑選出看漲看跌文章集

03 將文章集切詞並建構出
看漲及看跌之關鍵字列表

04 建構不同模型並比較各自的
confusion matrix 與正確率

05 移動回測並計算正確率

Data pre-process

- 將上漲與下跌各1500個關鍵字降維
- 將大量的0去除，濃縮成tf-idf的矩陣
- 相較於傳統的tf-idf算法利用「Okapi BM25」演算法來計算詞彙與文章間關聯度
- 公式：

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

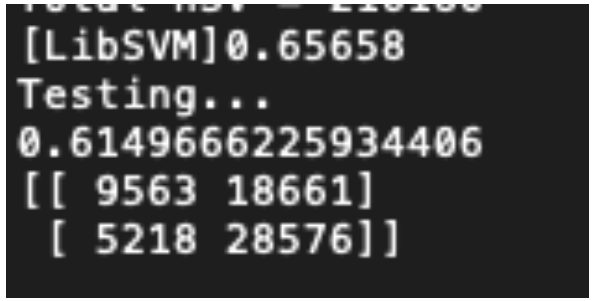
k_1,b：調節參數
q_i：關鍵詞
D：文章
avgdl：所有文章平均長度
|D|：此文章長度

- 概念與TF-IDF很像，但的是不同把文章長度加入為考量因素，同樣的關鍵字，詞頻較長文章重要度會低於較短文章，另外還加入兩個調節參數

Training-SVM

- 將前面預處理完成的資料用SVM模型訓練
- 設定上漲文章標記為「0」，下跌文章標記為「1」
- Training data以及Testing data分別為80%與20%
- 結果如下：

資料總數量：310,088筆
Training data準確率：65.6%
Testing data準確率：61.5%



Training-SVM

- 資料總數量：310,088筆

  Training data準確率：65.6% / Testing data準確率：61.5%

- Confusion Matrix :

	真實 為漲	真實 為跌
預測 為漲	9563	18661
預測 為跌	5218	28576

Training-Random Forest

```
0.8644909542390918
[[23219  5005]
 [ 3399 30395]]
```

- 資料預處理方式同SVM（Training data以及Testing data分別為80%與20%）
- 資料總數量：310,088筆　Training data準確率：91.4%　/　Testing data準確率：86.4%
- Confusion Matrix：

	真實為漲	真實為跌
預測為漲	23219	5005
預測為跌	3399	30395

SVM V.S. Random Forest

- 兩者皆可以處理高維度空間的資料
- Feature selection 若做不完善，會導致SVM結果不佳
- Random Forest 結合多個模型的優點，產出一個更好的結果
  並且也解決單個decision tree overfitting的問題
- SVM不適合訓練大量資料，速度會非常慢
- Random Forest 適合處理高維度空間以及大量資料

其他方法：KNN

- 隨機挑選100000筆資料（80%訓練、20%測試）
- 準確率：0.5833
- Confusion Matrix :

其他方法：Keras

- 隨機挑選100000筆資料（80%訓練、20%測試）
- 準確率：多數預測結果介於0.4～0.6，預測力不佳

```
In [17]: from keras import models
         from keras import layers

         model = models.Sequential()
         model.add(layers.Dense(16, activation='relu'))
         model.add(layers.Dense(1, activation='sigmoid'))

In [18]: from keras import optimizers
         model.compile(optimizer='rmsprop', loss='binary_crossentropy', metrics=['accuracy'])

In [19]: x_val = x_train[:20000]
         partial_x_train = x_train[20000:80000]
         y_val = y_train[:20000]
         partial_y_train = y_train[20000:80000]

In [21]: history = model.fit(partial_x_train,
                             partial_y_train,
                             epochs=10,
                             batch_size=512,
                             validation_data=(x_val, y_val))

         Train on 60000 samples, validate on 20000 samples
         Epoch 1/10
         60000/60000 [==============================] - 1s 13us/step - loss: 0.6868 - accuracy: 0.5487 - val_loss: 0.6872 - va
         l_accuracy: 0.5469
         Epoch 2/10
         60000/60000 [==============================] - 1s 11us/step - loss: 0.6866 - accuracy: 0.5483 - val_loss: 0.6873 - va
         l_accuracy: 0.5481
         Epoch 3/10
         60000/60000 [==============================] - 1s 11us/step - loss: 0.6864 - accuracy: 0.5486 - val_loss: 0.6872 - va
```

大綱

01 將全部文章加上漲跌之標籤

02 依標籤挑選出看漲看跌文章集

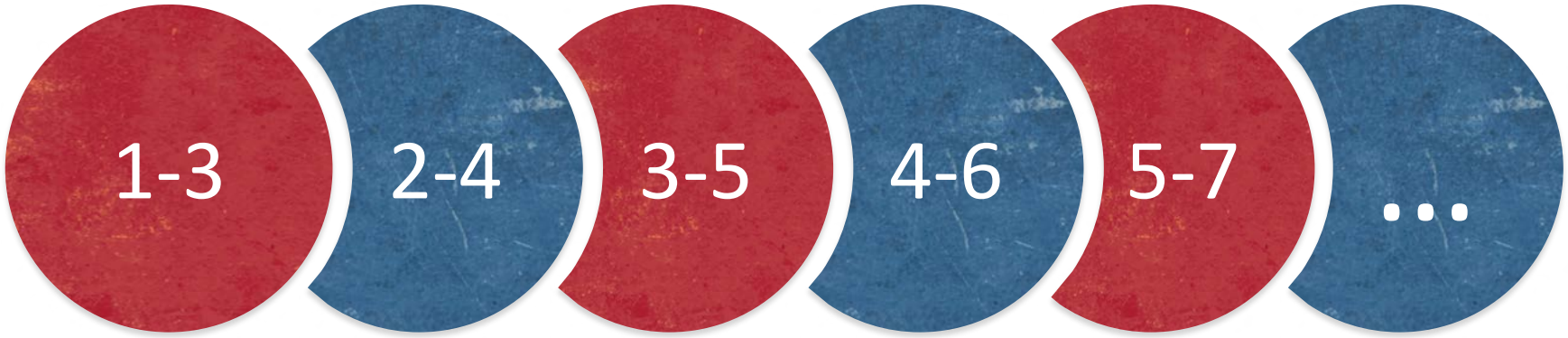
03 將文章集切詞並建構出
看漲及看跌之關鍵字列表

04 建構不同模型並比較各自的
confusion matrix 與正確率

05 移動回測並計算正確率

Back-test(利用random forest模型）

- 以2018年資料，每三個月資料預測下個月，共做了九次回測



Back-test(利用random forest模型)

- 以1~3月，預測4月

  資料數量：14456筆

  Training data準確率：95.4%
  Testing data準確率：53.1%

- 以2~4月，預測5月

  資料數量：10540筆

  Training data準確率：96.1%
  Testing data準確率：53.2%

```
training: 1 to 3, testing: 4
(14456, 64) (3566, 64) (14456,) (3566,)
Training...
0.954344216934145
Testing...
0.5316881660123387
[[1556  418]
 [1252  340]]
training: 2 to 4, testing: 5
(10540, 64) (6386, 64) (10540,) (6386,)
Training...
0.9612903225806452
Testing...
0.5327278421547135
[[2312 1392]
 [1592 1090]]
```

Back-test(利用random forest模型)

- 以3～5月，預測6月

  資料數量：14908筆

  Training data準確率：96.8%
  Testing data準確率：47.7%

- 以4～6月，預測7月

  資料數量：25886筆

  Training data準確率：97.3%
  Testing data準確率：52.0%

- 以5～7月，預測8月

  資料數量：74424筆

  Training data準確率：95.2%
  Testing data準確率：58.0%

```
training: 3 to 5, testing: 6
(18, 64) (15934, 64) (14908,) (15934,)
Training...
0.968204990609069
Testing...
0.47759507970377807
[[5116 1924]
 [6400 2494]]
training: 4 to 6, testing: 7
(25886, 64) (52104, 64) (25886,) (52104,)
Training...
0.9734991887506761
Testing...
0.5202671579917089
[[16122 14932]
 [10064 10986]]
training: 5 to 7, testing: 8
(74424, 64) (42204, 64) (74424,) (42204,)
Training...
0.9524078254326561
Testing...
0.5800398066533978
[[20638  5890]
 [11834  3842]]
```

Back-test(利用random forest模型)

- **以6~8月，預測9月**
  資料數量：110242筆
  Training data準確率：94.7%
  Testing data準確率：52.1%

- **以7~9月，預測10月**
  資料數量：136016筆
  Training data準確率：94.4%
  Testing data準確率：41.2%

- **以8~10月，預測11月**
  資料數量：131392筆
  Training data準確率：94.7%
  Testing data準確率：54.2%

- **以9~11月，預測12月**
  資料數量：44186筆
  Training data準確率：95.5%
  Testing data準確率：61.6%

```
training: 6 to 8, testing: 9
(110242, 64) (41708, 64) (110242,) (41708,)
Training...
0.9477694526586963
Testing...
0.5214347367411528
[[14752  2780]
 [17180  6996]]
training: 7 to 9, testing: 10
(136016, 64) (47480, 64) (136016,) (47480,)
Training...
0.9446241618633102
Testing...
0.4124262847514743
[[ 9544  4064]
 [23834 10038]]
training: 8 to 10, testing: 11
(131392, 64) (54998, 64) (131392,) (54998,)
Training...
0.9476071602532878
Testing...
0.5420560747663551
[[ 8146 13966]
 [11220 21666]]
training: 9 to 11, testing: 12
(144186, 64) (31252, 64) (144186,) (31252,)
Training...
0.95500256612986
Testing...
0.6168565211826443
[[ 1366  7246]
 [ 4728 17912]]
```

Back-test結果不盡理想原因

報告影片

https://www.youtube.com/watch?v=3bFmn2Riacw&feature=youtu.be

Thank You.