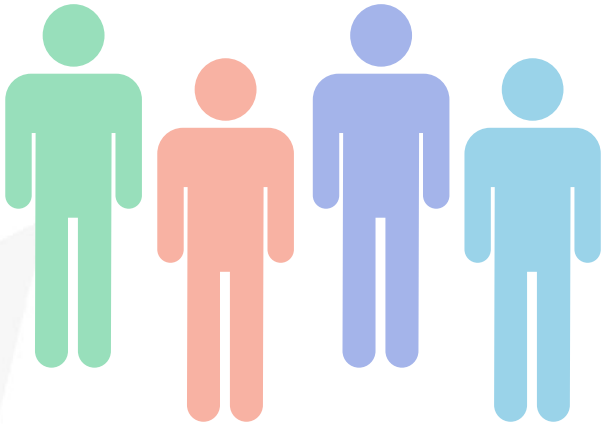
# 文字探勘
# 野村投信題目二 第1組

融入質化因子的基金加減碼策略

# 小組成員



- 巨資大三 陳偉傑
- 財金碩一 楊雅婷
- 財金大四 吳浩夫
- 工科碩二 王晉偉

# Outline

1. 資料與爬蟲
2. 量化 & 質化因子
3. Final model - SVM
4. Telegram Chatbot

# 專案流程規劃

01

## 確認財務量化因子

想做出好的投資決策，無可避免需要關注金融市場及經濟動態，因此想出幾種財務因子，希望可以和質化因子相輔相成，預測基金淨值走向。目前有 Sortino ratio、美林時鐘、政策利率

02

## 確認質化因子

我們有兩個想法，第一是用論壇討論文章判斷市場是否過於樂觀或過於悲觀，第二是用新聞分類方式預測市場走向

03

## 爬蟲

爬10-k、Reddit、WSJ、Forbes

04

## 最終模型

使用SVM，將未來基金淨值分為上漲(1)和下跌(0)兩類進行預測，並將回測績效以 Telegram Chatbot呈現

# Reddit 與 10-K 資料爬蟲

# Reddit 資料爬蟲


- API: **Pushshift**
- Subreddit**: investing, fund**
- Total Data: 15092
- From: 2017

| Date | Name | Score | Title | Content |
|---|---|---|---|---|
| **2019-09-01** | MasterCookSwag | 271 | Formal posting guidelines for political topics... | Alright everyone, it looks like we had pretty ... |
| **2020-06-03** | AutoModerator | 11 | Daily Advice Thread - All basic help or advice... | If your question is "I have $10,000, what do I... |
| **2020-06-03** | jayjayy123 | 446 | Choosing a stock is like choosing your gf, if ... | What are some of your favourite picks? |
| **2020-06-03** | SD987 | 335 | List of stocks still down more than 40% | I'm trying to compile a list of stocks that st... |
| **2020-06-03** | achicomp | 1066 | May ADP employment report: -2.76 million vs ex... | Absolutely monster job report beat. Stock futu... |

# 10-K 資料爬蟲

- API: **sec-edgar-downloader**
- Total Company : 98
- Total Data: 471


UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

**Form 10-K**

(Mark One)

☒ ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the fiscal year ended September 28, 2013

Or

☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the transition period from _____ to _____

Commission file number: 000-10030

**APPLE INC.**

(Exact name of registrant as specified in its charter)

| California | 94-2404110 |
|---|---|
| (State or other jurisdiction of incorporation or organization) | (I.R.S. Employer Identification No.) |
| 1 Infinite Loop<br>Cupertino, California | 95014 |
| (Address of principal executive offices) | (Zip Code) |

Registrant's telephone number, including area code: (408) 996-1010

Securities registered pursuant to Section 12(b) of the Act:

| Common Stock, no par value | The NASDAQ Stock Market LLC |
|---|---|
| (Title of class) | (Name of exchange on which registered) |

Securities registered pursuant to Section 12(g) of the Act: None

# 10-K 資料文字預處理

擷取公司 10-K 財報中 Item 1，Item 1A， Item 5， Item 7 的內容，並以公司名以及年份作區別，當作文章內容。

利用自然語言處理 NLTK 套件將文章斷詞，並標註詞性，以及移除 stopwords

| | Year | Company | Content |
|---|---|---|---|
| **0** | 2014 | MU | [corporate, information, micron, delaware, cor... |
| **1** | 2015 | MU | [overview, micron, technology, inc., including... |
| **2** | 2016 | MU | [following, discussion, contains, trend, infor... |
| **3** | 2017 | MU | [overview, manufacture, product, worldwide, wh... |
| **4** | 2018 | MU | [overview, manufacture, product, wholly-owned,... |

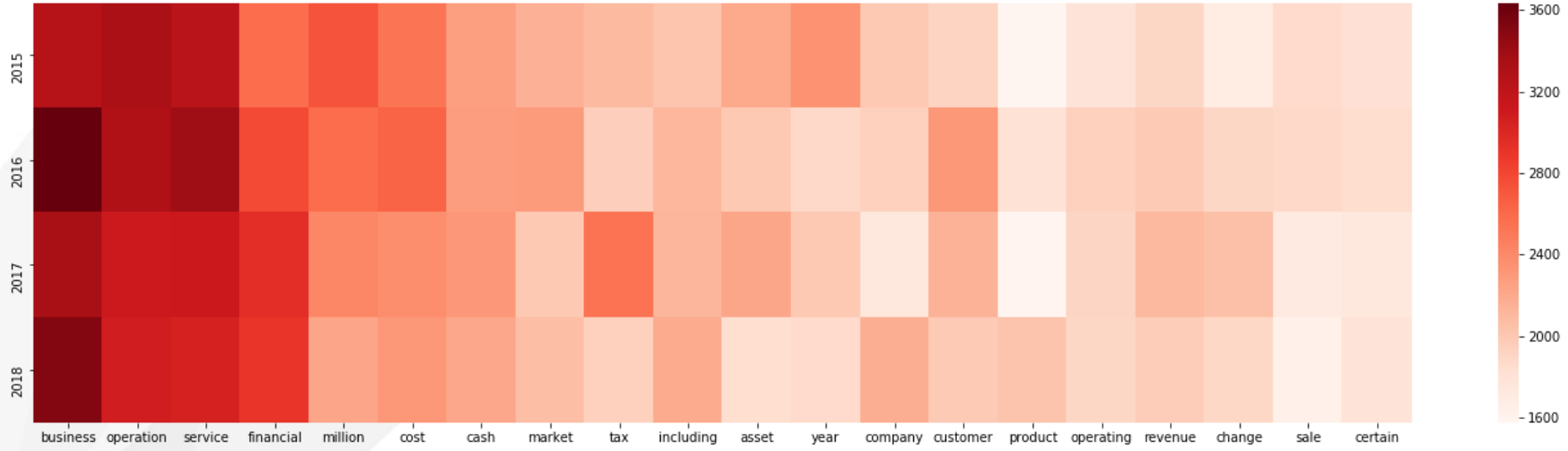
# 10-K 資料文字分析結果

針對預處理完的財報資料，進行 TF-IDF 詞頻分析，並算出共線性矩陣以及熱點圖。

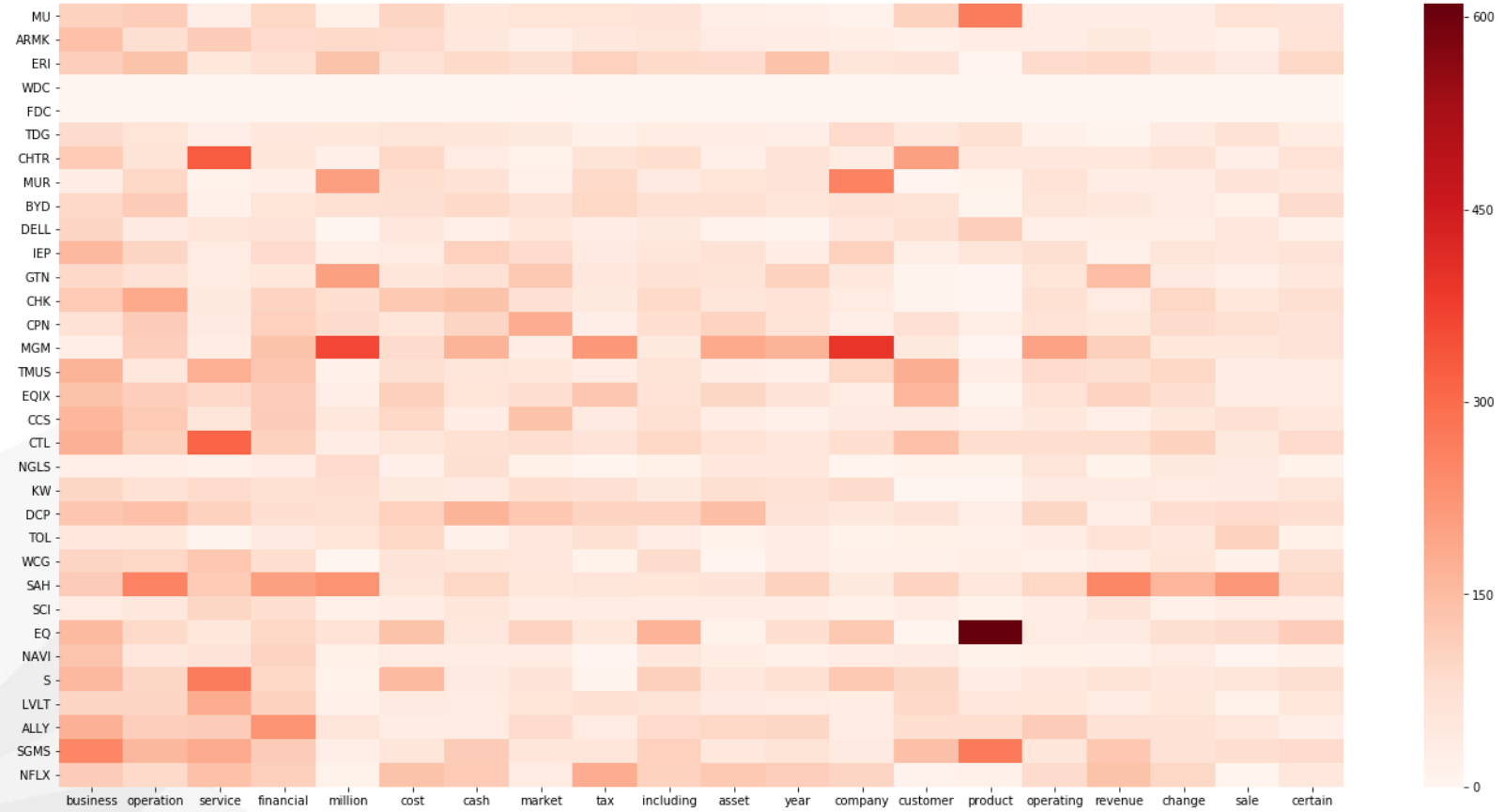
從結果發現財報各家公司用字類似，針對詞頻較高的詞彙並沒有年份有顯著的區別。

```
business : 21 Degree
operation : 21 Degree
service : 21 Degree
financial : 21 Degree
million : 21 Degree
cost : 21 Degree
cash : 21 Degree
market : 21 Degree
tax : 21 Degree
including : 21 Degree
asset : 21 Degree
year : 21 Degree
company : 21 Degree
customer : 21 Degree
product : 21 Degree
operating : 21 Degree
revenue : 21 Degree
change : 21 Degree
sale : 21 Degree
certain : 21 Degree
```

# 10-K 資料隨不同年份詞頻的熱點圖



# 10-K 資料隨不同公司詞頻的熱點圖



# 華爾街日報爬蟲

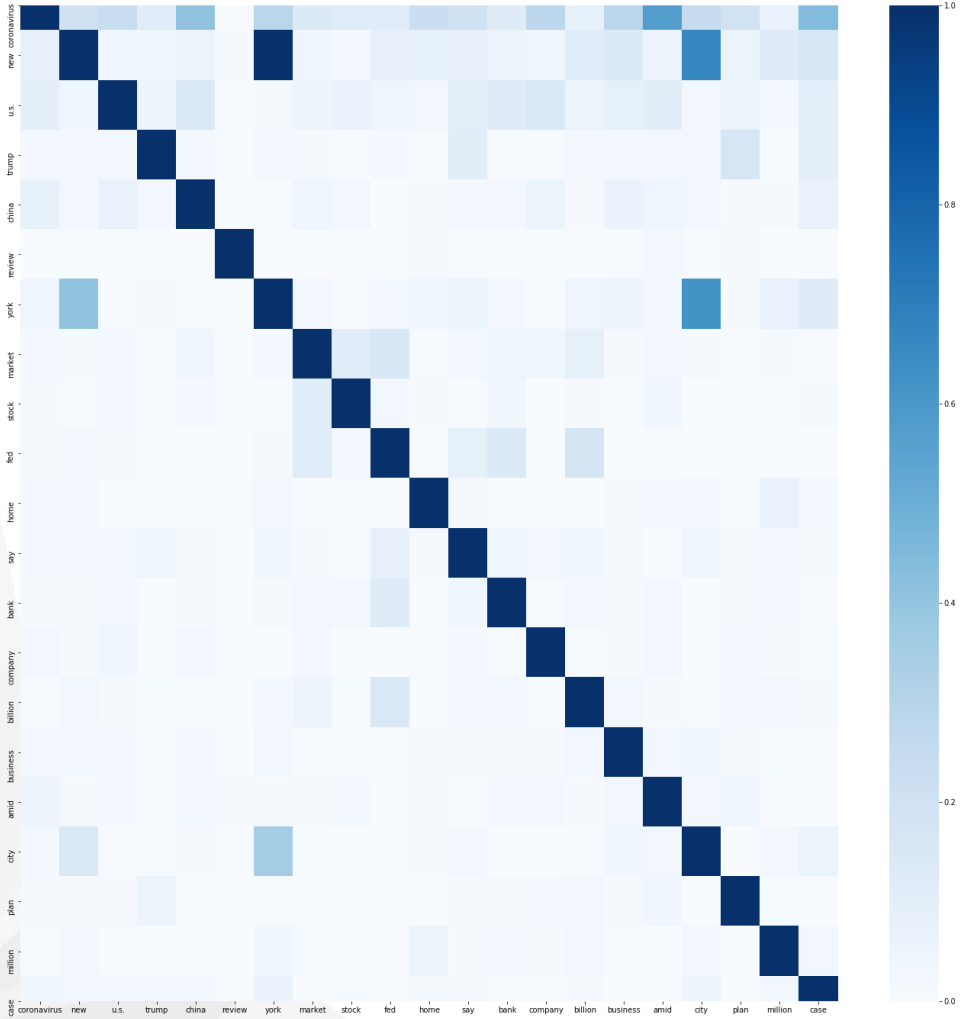
- WSJ Articles
- Total Data:
- Date: 2003 / 2004 / 2011 / 2020

| | date | title | sub_title | content |
|---|---|---|---|---|
| 0 | 1-4-2020\n | Affordable Care Act Sign-Ups Total 11.4 Mi... | ACA sign-ups steady for third straight year de... | About 11.4 million consumers signed up for hea... |
| 1 | 1-4-2020 | Florida Issues Statewide Limits as U.S. Co... | Tally of confirmed cases world-wide passes 900... | In the span of a month, the coronavirus pandem... |
| 2 | 1-4-2020 | Coronavirus Lockdowns Prompt Smaller Resta... | Eateries examine different ways to get food to... | Restaurants' increasing dependence on companie... |
| 3 | 1-4-2020 | The New York Neighborhoods With the Most C... | Working-class and Orthodox Jewish neighborhood... | The new coronavirus has struck hardest in work... |
| 4 | 1-4-2020 | New York State Lawmakers Reach Tentative B... | Gov. Cuomo says falling revenue complicated th... | New York state lawmakers reached a tentative s... |

- 基金净值 ： TEJ
- 野村基金介紹 ： 野村官網
- 專案標的： 野村環球債 /野村優質 /野村全球品牌證券

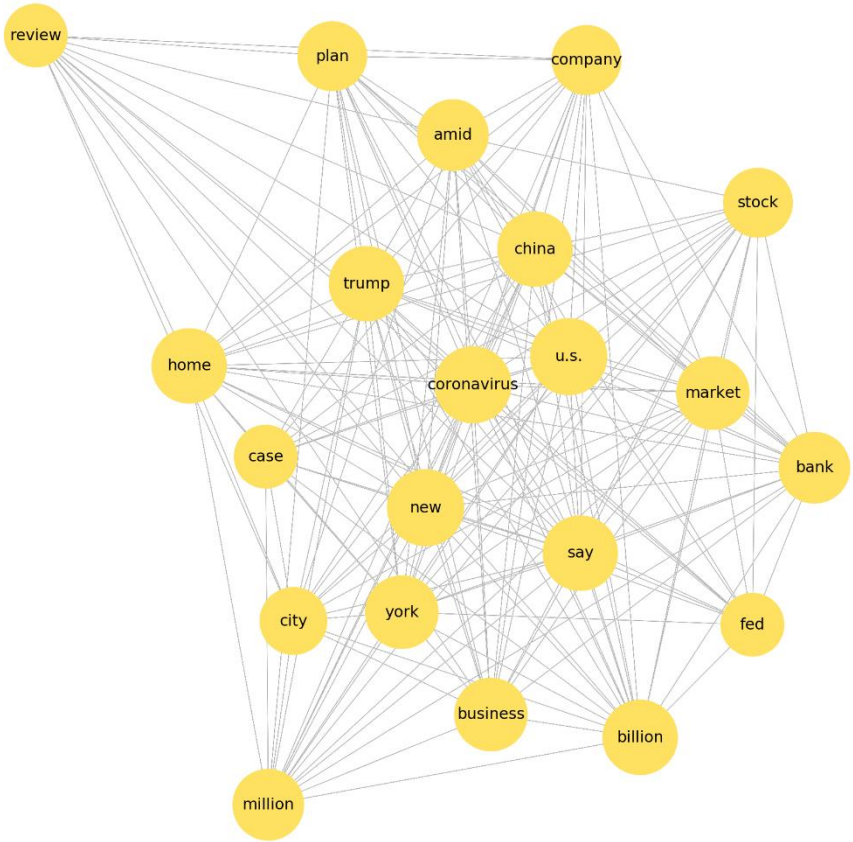
NOMURA
Nomura Asset Management Taiwan

A JOINT VENTURE WITH
ALLSHORES




## 華爾街日報標題

藉由熱點圖驗證從標題中關鍵字之分佈情況，并把關鍵字時間的關係以網路圖呈現。




## 小結

在瞭解各個關鍵字之間的關係後，列出出現關鍵字報導的日期，接著對應到基金趨勢情況，最終確認華爾街日報與市場反應存在一定的關係。於是針對華爾街日報 與 Reddit 進行模型訓練。

| | Date with keyword |
|---|---|
| 0 | 1-4-2020 |
| 1 | 31-3-2020 |
| 2 | 29-3-2020 |
| 3 | 28-3-2020 |
| 4 | 27-3-2020 |

# 量化因子

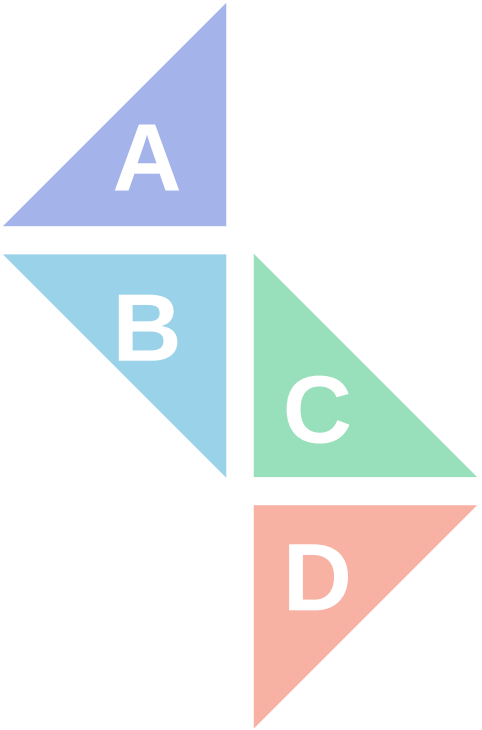
# 財務量化因子

**Sortino ratio**

用於計算投資組合超額報酬績效，僅考慮下方風險而不考慮上方風險

**美林時鐘**

最初由美銀美林提出，藉由非食品 CPI、PMI、M1、M2、PPI-CPI 等經濟數據預測未來經濟走勢，將經濟情況分為四象限




**政策利率**

由於目前我們將重心放在債型基金，利率與債券價格息息相關，因此我們將美國政策利率加入因子之中

**其他**

考慮將美林時鐘使用的數據直接作為因子，而不先轉換成時鐘象限

# Sortino Ratio

```
s_list = []
for i,row in df_merge.iterrows():

        r = sell_return(row['淨值(元)'], nv0, row['年月日'], date0) # 計算每日年化報酬率
        if r < (row['1M']/100):
           s_list.append( (r-row['1M']/100)**2 ) #挑出報酬率 < 必要報酬率的return差值平方，因為s
        else:
           s_list.append(0)
        #print(i," ",row['年月日']," ",r," ",row['1M'])
        if  len(s_list) > 1 and downDeviation(s_list)> 0:
           sortino = calculate_sortino(r,row['1M']/100,downDeviation(s_list))
           df_showSortino = df_showSortino.append({"年月日": row['年月日'],
                                                   "Sortino": sortino}, ignore_index = True)

pd.set_option('display.max_rows', None)
df_merge = pd.merge(df_merge,df_showSortino,on='年月日',how='left')
df_merge
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:8: RuntimeWarning: invalid val
```

| | 年月日 | 幣別 | 淨值(元) | 基金名稱 | 1M | Sortino |
|---|---|---|---|---|---|---|
| 0 | 2015-11-09 | USD | 10.0007 | 野村美利高收債B美金 | 0.065667 | NaN |
| 1 | 2015-11-10 | USD | 9.9956 | NaN | 0.065167 | -1.000000 |
| 2 | 2015-11-11 | USD | 9.9981 | NaN | 0.065267 | -0.352665 |
| 3 | 2015-11-12 | USD | 9.9809 | NaN | 0.065667 | -1.353460 |
| 4 | 2015-11-13 | USD | 9.9734 | NaN | 0.065767 | -1.256955 |

基金資料來源：TEJ

基本利率： USD 1M Libor

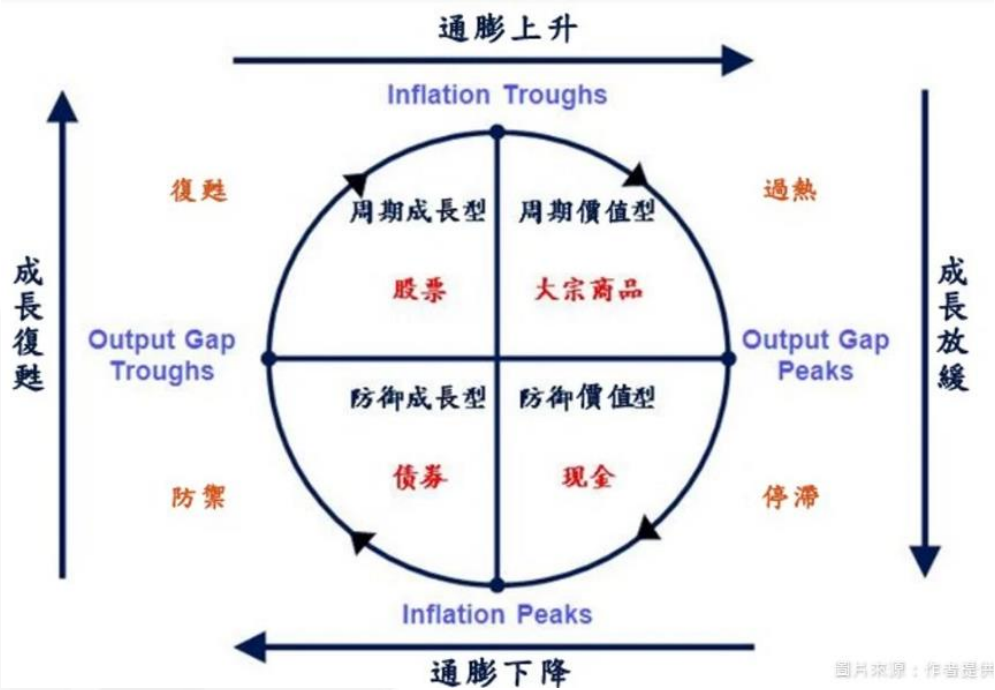
$$\text{Sortino Ratio} = \frac{R_p - R_f}{DD}$$

$R_p$ = 投資組合 $p$ 的報酬率

$R_f$ = 無風險報酬率

$DD$ = 下行標準差

# 美林時鐘




```
#======= PPI - CPI =======
def PPI_CPI( n, n_3): # 第 n 期 ppi - cpi 與第 n-3期 ppi - cpi
    if n<0 and n_3 <0 and (n - n_3) < 0: return 3
    if n - n_3 >0: return 2 # ppi - cpi 擴張 => 復甦
    elif n - n_3 <0:  # ppi - cpi 縮小 => 擴張 or 滯脹
     if n < 0: return 4  # 最新的 ppi 為負數 => 滯脹
     else: return 1

#======= PMI =======
def PMI(pmi, pmi_3):
  if pmi - pmi_3 >0 and pmi > 50 : return 1
  elif pmi - pmi_3 <0 and pmi<= 52: return 4
  elif pmi - pmi_3 >0 and pmi_3 < 50: return 2
  else: return 3

#======= M1 & M2 =======
def M1_M2(m1, m1_3, m2, m2_3):  # M1 是貨幣寬鬆狀況，M2 是信用寬鬆狀況
  if m1 - m1_3 > 0:  # 寬鬆貨幣 => 復甦 or 衰退
    if m2 - m2_3 >0: return 2
    else: return 3
  else:  # 擴張 or 滯脹
    if m2 - m2_3 >0: return 1  # 緊縮貨幣 + 仍寬信用
    else: return 4

#======= CPI_nofood =======
def CPI_nofood(cf, cf_3, ppi, ppi_3):
  if cf < 0 and cf_3 >= 0: return 2    # 原本下跌後停止下跌
  if ppi > 0 and ppi_3 >0 and cf > 0: return 1
  if ppi<0 and ppi_3 <0 and (cf - cf_3)>=0: return 4
```

# 質化因子

# Fasttext 與 BERT

## Fast text
## 預測市場反轉

「別人貪婪時我恐懼，別人恐懼時我貪婪」是巴菲特的名言，而我們即是希望用論壇發文，搭配由 Facebook 團隊設計出的 Fast text 詞向量模型，來計算每一天 Reddit 上投資相關發文的詞向量，並使用 PCA 降維取得單日發文詞特徵，來判斷市場是否過於恐懼或過於貪婪。

## BERT
## 預測漲跌

除了Fast text 外，詞向量模型還有 Google 團隊設計的 BERT，BERT比 Fast text更複雜且龐大，我們希望用BERT進行新聞漲跌分類，但BERT最初是由廣泛而無特定類型的文本訓練成，目前我們用來做財經文本分類效果並不好，可能需要用更多商業及財經類型的文章繼續訓練。

# 詞向量

## 詞向量原理

詞向量是把文字「量化」的一種方式，用一個或多個數字的及何來描述一個詞，好的詞向量應該要讓意思相近的詞在向量空間中的位置相近。詞向量最初用one hot encoding的方式配置，但這會有Sparse matrix 浪費記憶體空間及運算效能的問題，後來出現word2vector，用淺層神經網路建立詞嵌入模型，並計算詞向量。

## 從W2V 到 Fast text 及BERT

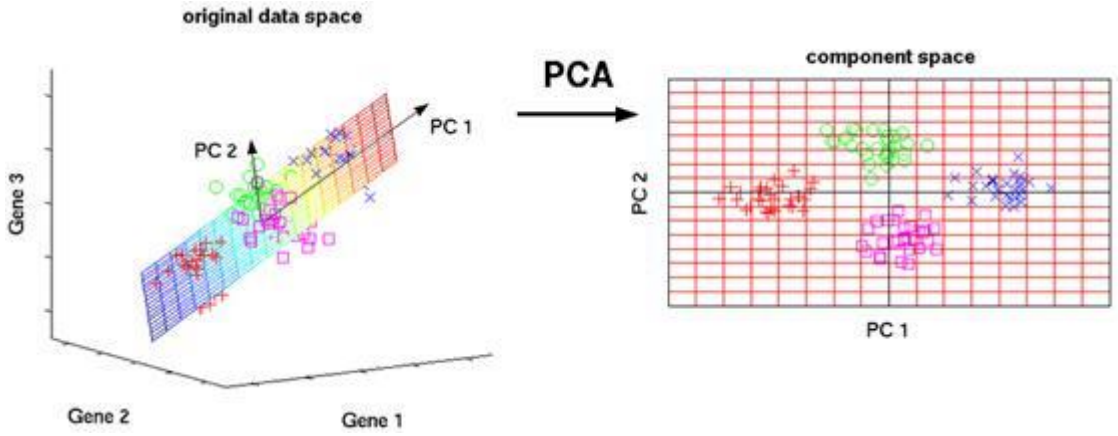
Word2vector 的效果不錯，但缺點是無法處理沒看過的詞(例如名詞單複數)，且無法考慮詞彙在句子中出現的順序，例如 Here you are 跟 You are here 對 w2v 來說一模一樣，因此後來 Facebook 團隊用 n-gram 及 n-char 方式、Google 團隊用 Attention、subword 方式來改良w2v的缺點。

# Fasttext

Reddit – StockMarket & Stocks

| Date | Title | Content | title_edited | content_edited |
|---|---|---|---|---|
| 20190401 | U.K. Watchdog to Review Conduct, Governance at... | 701 pm ET The U.K. watchdog for accounting an... | [u.k., watchdog, review, conduct, governance, ... | [701, u.k., watchdog, accounting, audit, tuesd... |
| 20190401 | U.K. Business Sentiment Declines as Brexit Unc... | 701 pm ET Corporate decision makers in the U.... | [u.k., business, sentiment, decline, brexit, u... | [701, corporate, decision, maker, u.k., strugg... |
| 20190401 | Slack Chooses NYSE for Direct Listing | 619 pm ET Slack Technologies Inc. has selecte... | [slack, chooses, nyse, direct, listing] | [619, slack, technology, inc., selected, new, ... |
| 20190401 | U.S. Dollar Slips as Investors Seek Risk After... | 614 pm ET The dollar fell against emerging-ma... | [u.s., dollar, slip, investor, seek, risk, str... | [614, dollar, fell, emerging-market, currency,... |
| 20190401 | U.S. Government Bond Prices Fall On Global Gro... | 429 pm ET U.S. government bond prices fell sh... | [u.s., government, bond, price, fall, global, ... | [429, u.s., government, bond, price, fell, sha... |
| 20190401 | Minneapolis Fed Chief Kashkari Says MMT Isn?□ ... | 406 pm ET Minneapolis Fed leader Neel Kashkar... | [minneapolis, fed, chief, kashkari, say, mmt, ... | [406, minneapolis, fed, leader, neel, kashkari... |
| 20190401 | Apollo-Backed Adhesives Maker Hexion Files for... | 402 pm ET Hexion Inc., an adhesives and coati... | [apollo-backed, adhesive, maker, hexion, file,... | [402, hexion, inc., adhesive, coating, busines... |
| 20190401 | Apollo Global-Controlled Hexion Files for Chap... | 358 pm ET Hexion Inc., a thermoset-resins con... | [apollo, global-controlled, hexion, file, chap... | [358, hexion, inc., thermoset-resins, controll... |
| 20190401 | Copper Prices Slip on Signs of Industrial Down... | 322 pm ET Copper prices retreated on Monday, ... | [copper, price, slip, sign, industrial, downturn] | [322, copper, price, retreated, monday, haltin... |
| 20190401 | Gasoline Prices Creep Toward $3 a Gallon | 151 pm ET Gasoline prices typically move high... | [gasoline, price, creep, gallon] | [151, gasoline, price, typically, higher, time... |

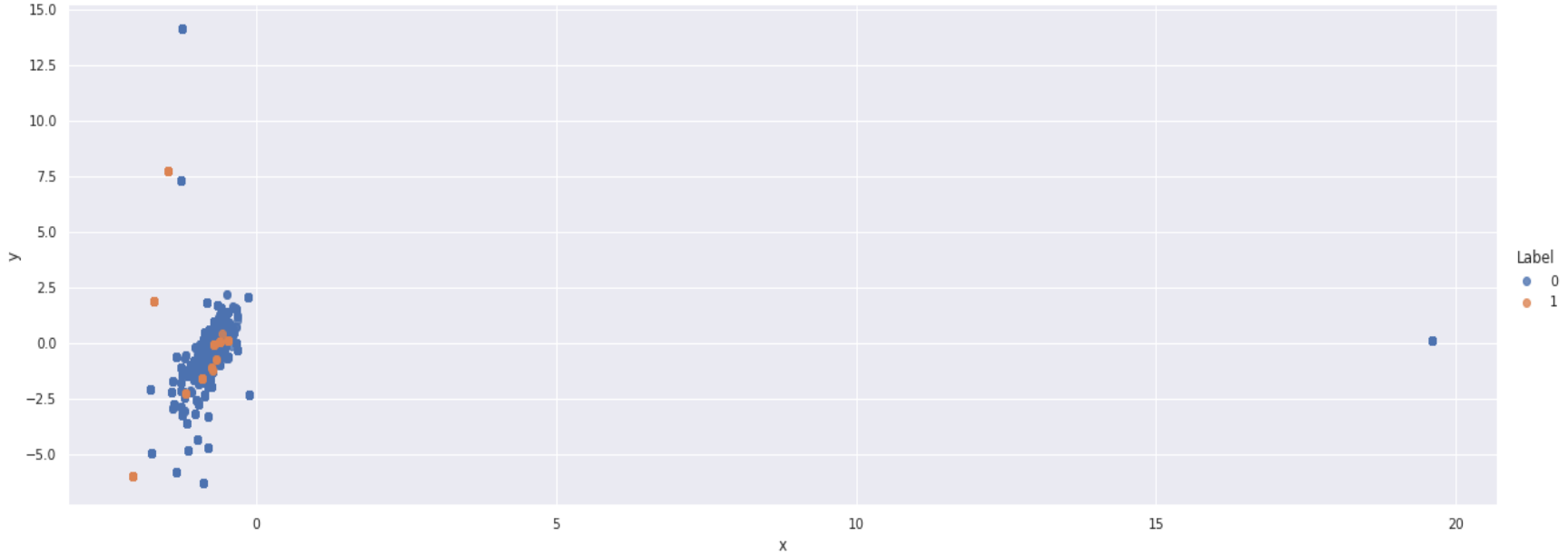
# PCA 主成分分析




利用線性轉換，找出能高維度空間中不同「點」的投射，且這些投射彼此間的差異性最大。以3D→2D為例，就是找出能讓三維空間中每個點投射後差異最大的平面(2D)，也就是保存能區分不同點的主要特徵

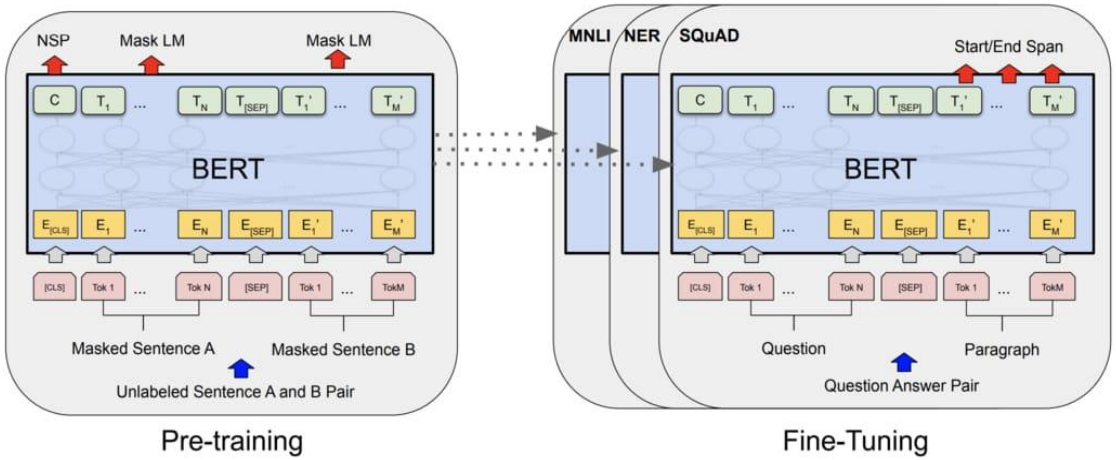
# PCA後標記

10天內跌10%，則10天都標1或隔日單日跌10%標1



# BERT

BERT 是 Google 以無監督的方式，利用大量無 Label 文本所開發的語言模型，本文將使用 Fine-Tuning 的模型對 WSJ 報導進行分類




**Ref: Jacob Devlin (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

# 文字資料預處理

將 WSJ 報導內非文字的符號去除，並且將文字轉為 token 與 segments 以符合模型的輸入格式

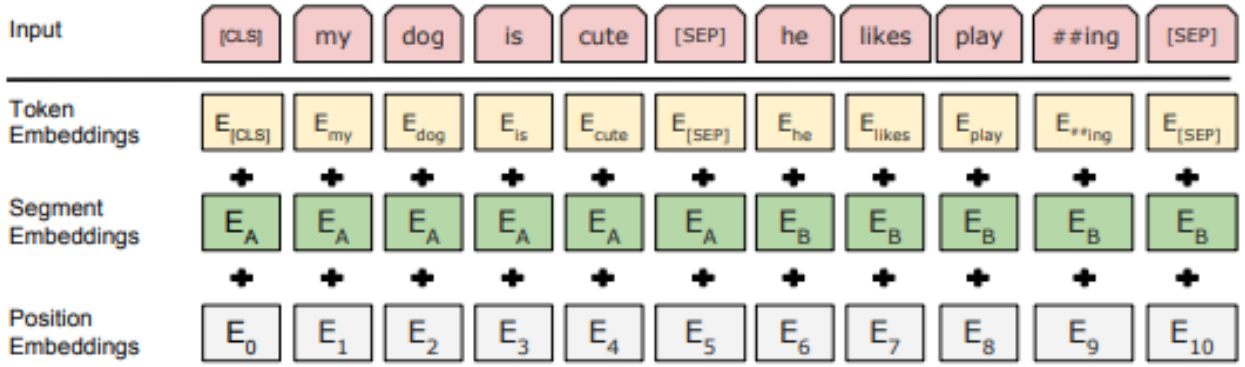
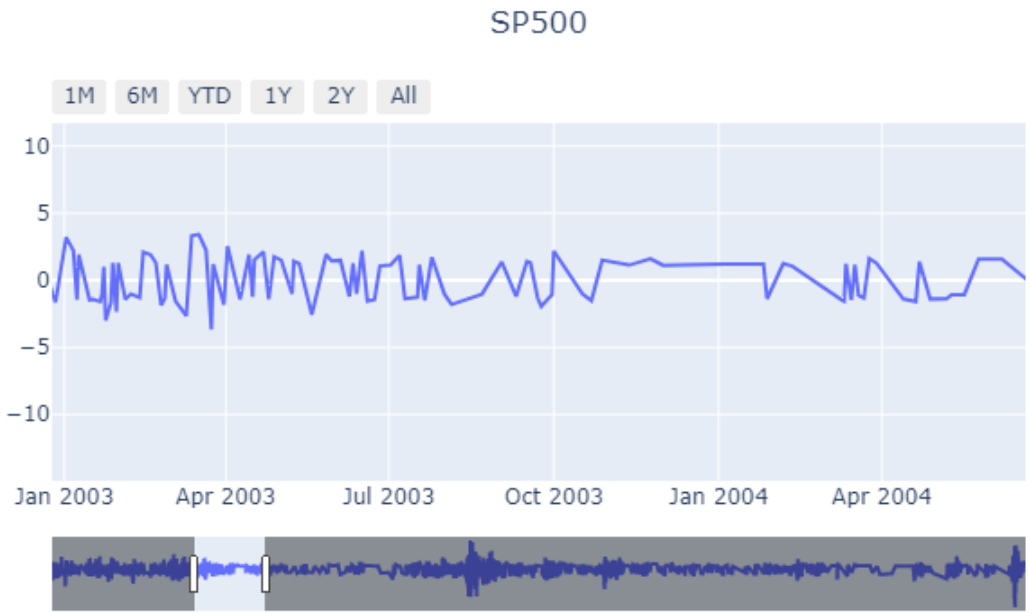



Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

# S&P500 資料預處理

由於 WSJ 資料為 2003年， S&P 500 的漲跌幅如右圖，而分類的標準如下

| 漲跌幅 (%) | Label |
|---|---|
| >1% | 1 |
| <1% | 0 |



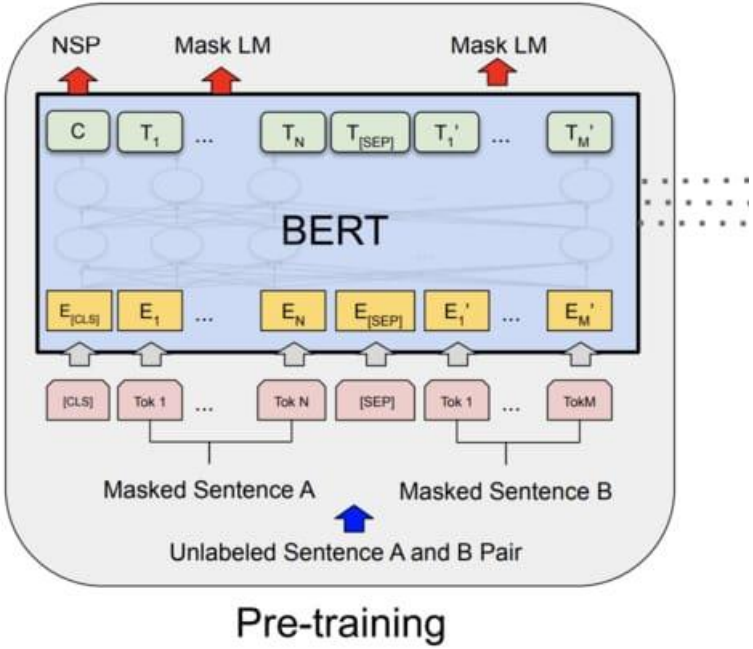
# 合併資料

考慮報導為延遲指標，針對漲跌幅對時間往後 Shift 天數，再以 Shift 欄的數值做分類

| | Date | Text | Close | 漲跌幅 | Shift | Label |
|---|---|---|---|---|---|---|
| **0** | 2003-06-04 | dj market talk/sg wall st ends up on greenspan... | 986.23999 | 1.488481 | 1.488481 | 1.0 |
| **1** | 2003-06-04 | wsj brokerage firm seeks to early bird in bagh... | 986.23999 | 1.488481 | 1.488481 | 1.0 |
| **2** | 2003-06-04 | dj market talk/kl wall st ends up on greenspan... | 986.23999 | 1.488481 | 1.488481 | 1.0 |
| **3** | 2003-06-04 | wsj update indictment of martha stewart is clo... | 986.23999 | 1.488481 | 1.488481 | 1.0 |
| **4** | 2003-06-04 | dj this is dow jones in new yorkdow jones fina... | 986.23999 | 1.488481 | 1.488481 | 1.0 |

# Pre-trained 模型測試

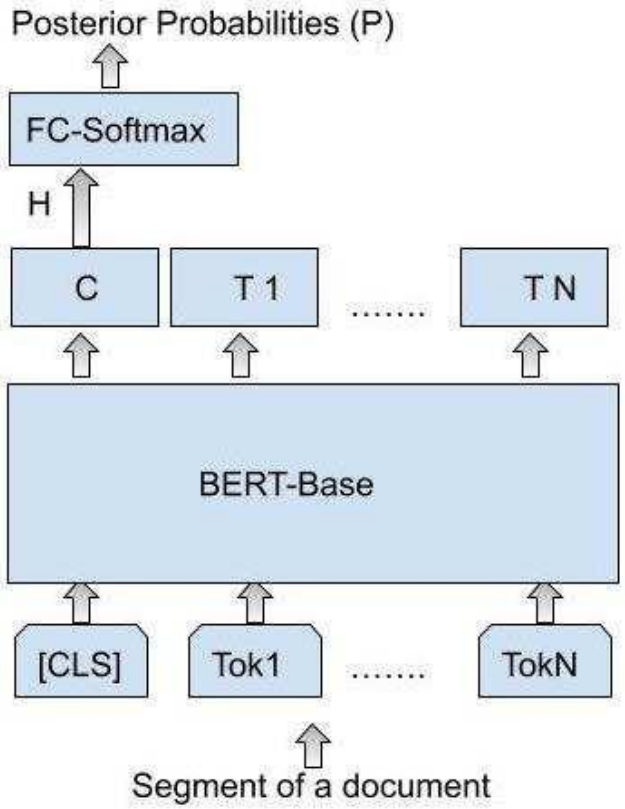
Google 官方提供 Pre-trained 的模型權重，因此，針對 WSJ 文章，本文先使用 Mask LM 的方式，隨機將文章中的字遮罩，並讓 BERT 模型從字典中預測此字，其預測效果不錯，但在某些金融用語上預測就相對較差



# Fine-Tuning 最終模型

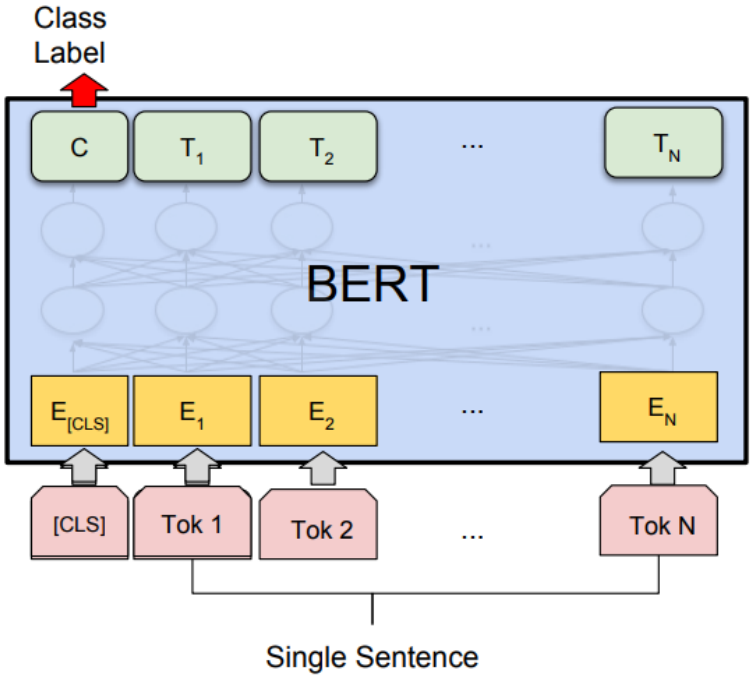
我們的目標是對單一文章做漲跌的分類，因此，在 BERT 模型後加入簡單的分類層*。由於 Pre-trained 模型測試對於金融用語有待改善，勢必要對 BERT 模型再訓練

| | |
|---|---|
| Training Data | 51053 |
| Testing Data | 12764 |
| Positive Data | 27122 |
| Negative Data | 23931 |




***Ref: Pappagari (2019) Hierarchical Transformers for Long Document Classification**

# Fine-Tuning 最終模型

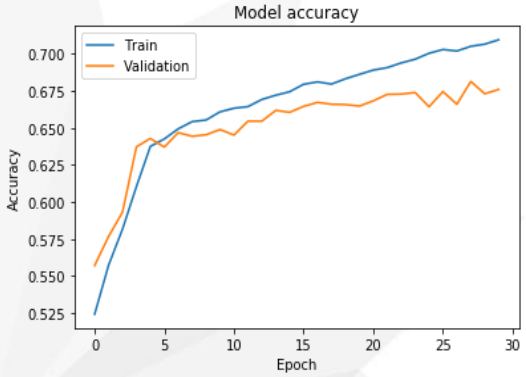


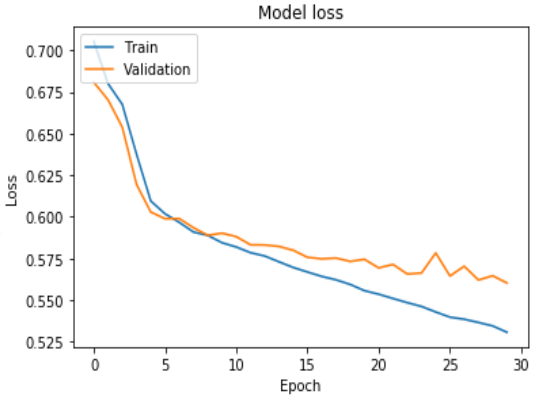

| Layer (type) | Output Shape | Param # | Connected to |
|---|---|---|---|
| input_1 (InputLayer) | (None, None) | 0 | |
| input_2 (InputLayer) | (None, None) | 0 | |
| model_2 (Model) | multiple | 108873384 | input_1[0][0]<br>input_2[0][0] |
| lambda_1 (Lambda) | (None, 768) | 0 | model_2[1][0] |
| dense_1 (Dense) | (None, 64) | 49216 | lambda_1[0][0] |
| dense_2 (Dense) | (None, 2) | 130 | dense_1[0][0] |

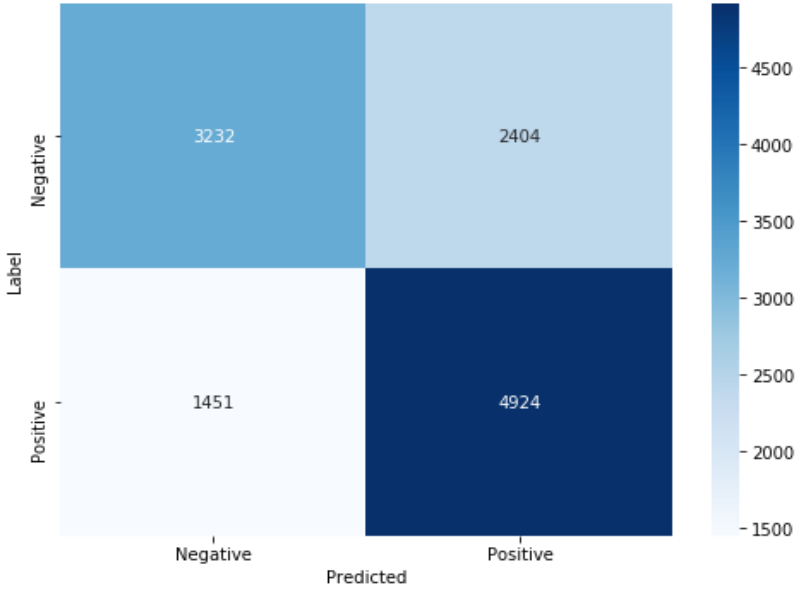
Total params: 108,922,730
Trainable params: 362,858
Non-trainable params: 108,559,872

# 預測結果

| Data Shift | None |
|---|---|
| SEQ LEN | 128 |
| Learning Rate | 3e-5 |
| Classifier Neuron | 64 |
| Batch Size | 32 |
| Accuracy | 67.90% |

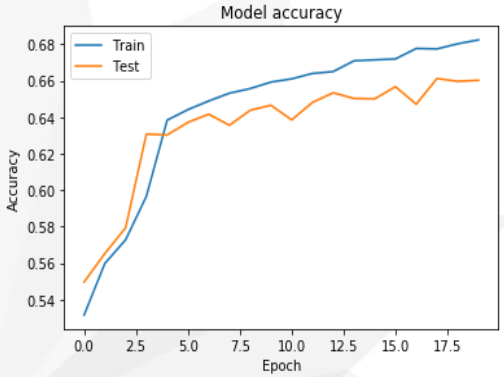


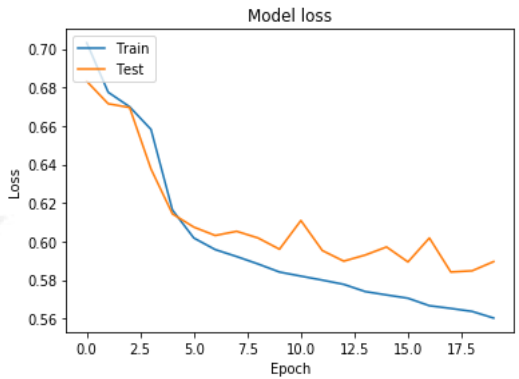



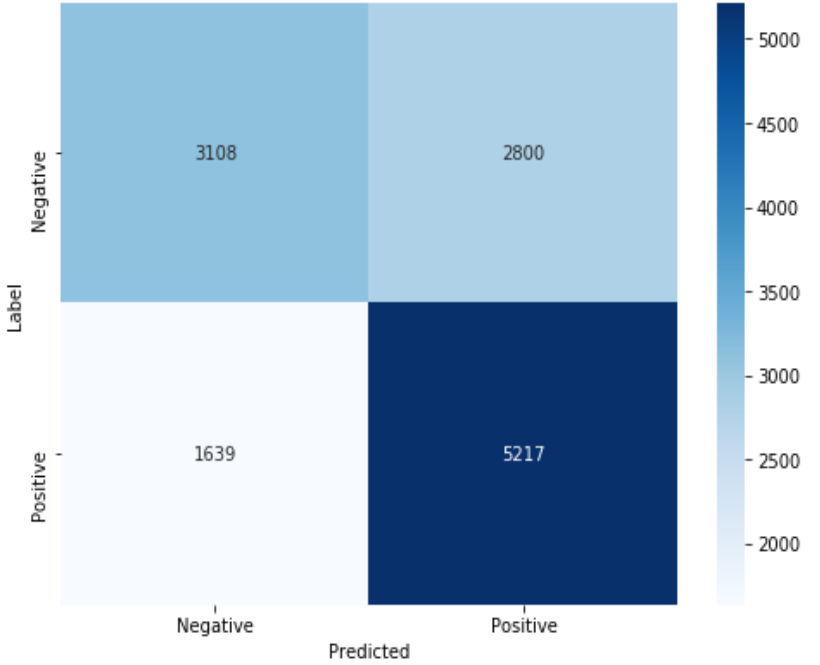



# 預測結果

| Data Shift | 1 Day Delay |
|---|---|
| SEQ LEN | 128 |
| Learning Rate | 3e-5 |
| Classifier Neuron | 64 |
| Batch Size | 32 |
| Accuracy | 65.22% |



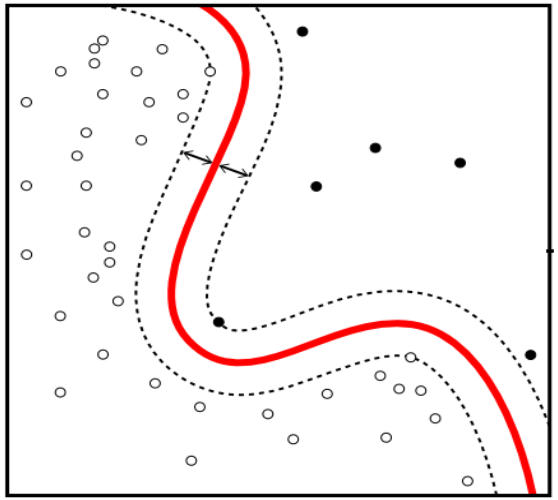



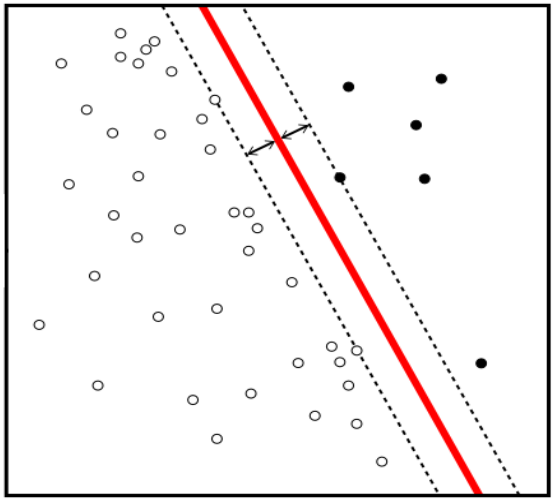

SVM

# 支持向量機 SVM

Support Vector Machine

SVM是一種監督式的學習方法，用統計風險最小化的原則來估計一個分類的超平面(hyperplane)，其基礎的概念非常簡單，就是找到一個決策邊界讓兩類之間的邊界(margins)最大化，使其可以完美區隔開來。





# Kernel 函數

當不同類別的資料在原始空間中無法被線性分類器區隔開來時，經由非線性投影後的資料能在更高維度的空間中可以更區隔開。

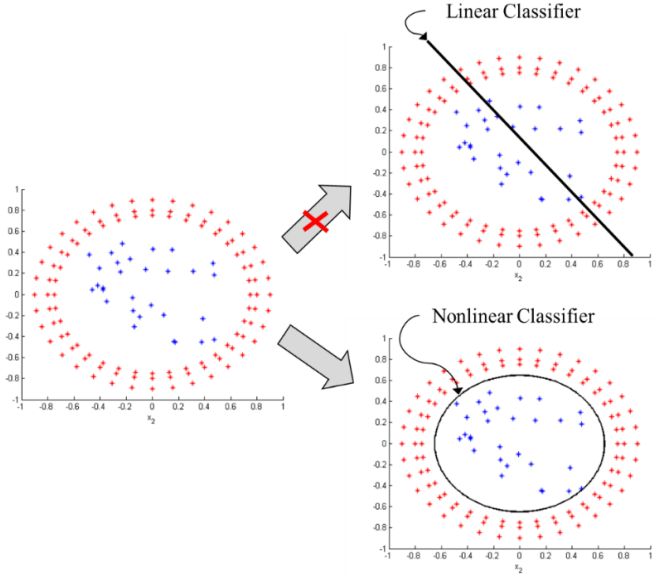
比較常用的**kernel**函數:

$$\text{Linear kernel: } k(x,y) = < x, y >$$

$$\text{Polynomial kernel: } k(x,y) = (< x, y > + c)^d$$

$$\text{Gaussian Radial Basis Function kernel (RBF): } k(x,y) = e^{-\frac{\|x-}{2}}$$

$$d \in Z^+, \sigma \in \mathcal{R} - \{0\}$$

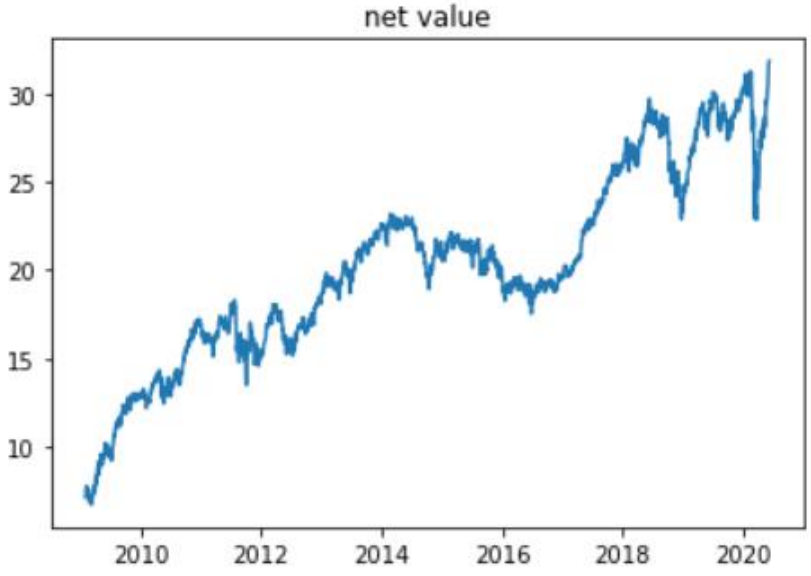


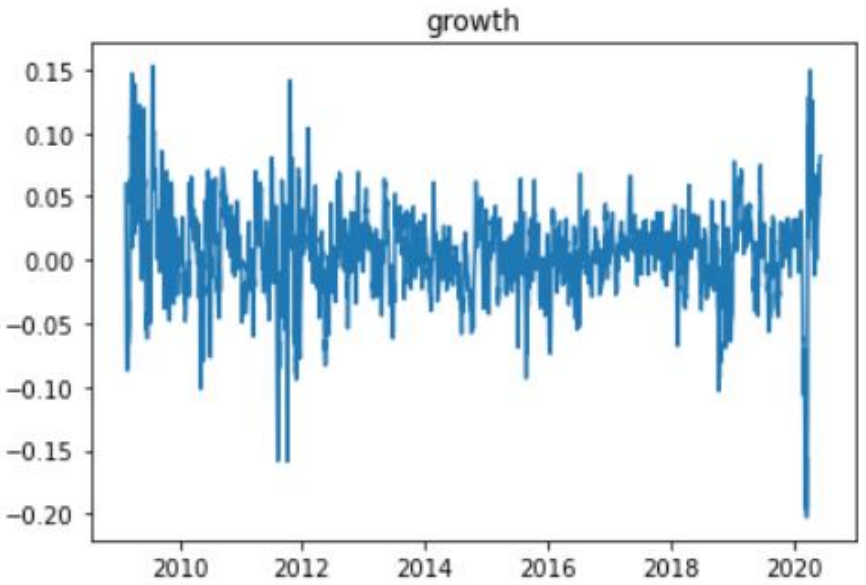
# 野村全球品牌證券投資信託基金

| | |
|---|---|
| 目前狀態 | Current |
| 基金統編 | 14697373 |
| 基金ISIN Code | TW000T3208Y3 |
| 基金全稱 | 野村全球品牌證券投資信託基金 |
| 基金英文全稱 | Nomura Global Luxury Brands Fund |
| 經理公司 | 野村證券投資信託 |
| 保管機構 | 彰銀 |
| 保證機構 | |
| 成立日 | 2002/6/21 |
| 進場日 | 2002/6/27 |
| 上市日 | |
| 清算/合併日 | |
| 清算/合併說明 | |
| 類型 | 開放 |
| 型態 | 國內募集,投資國內外 |
| 投資標的 | 股票 |
| TEJ分類 | 跨國--全球股票 |
| 公會分類 | 跨國--全球股票 |
| 風險收益等級 | RR4 |
| 主基金標示 | Y |
| 主基金代碼 | T3208Y |

| | 年月日 | 幣別 | 淨值(元) |
|---|---|---|---|
| 0 | 2020/06/08 | NTD | 31.84 |
| 1 | 2020/06/05 | NTD | 31.69 |
| 2 | 2020/06/04 | NTD | 31.12 |
| 3 | 2020/06/03 | NTD | 31.29 |
| 4 | 2020/06/02 | NTD | 30.85 |
| ... | ... | ... | ... |
| 2775 | 2009/02/06 | NTD | 7.73 |
| 2776 | 2009/02/05 | NTD | 7.39 |
| 2777 | 2009/02/04 | NTD | 7.36 |
| 2778 | 2009/02/03 | NTD | 7.23 |
| 2779 | 2009/02/02 | NTD | 7.16 |

# 野村全球品牌證券投資信託基金



# Label 方式

| | Date | 10day_growth | 1day_growth | BuyOrNot | 10day_label | 1day_label |
|---|---|---|---|---|---|---|
| **0** | 2020-06-02 | 0.057593 | 0.011807 | 0 | 0 | 0 |
| **1** | 2020-06-01 | 0.072082 | 0.007268 | 0 | 0 | 0 |
| **2** | 2020-05-29 | 0.074929 | 0.004313 | 0 | 0 | 0 |
| **3** | 2020-05-28 | 0.074127 | 0.007016 | 0 | 0 | 0 |
| **4** | 2020-05-27 | 0.058345 | 0.007066 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... |
| **2761** | 2009-02-20 | -0.060802 | -0.017591 | 1 | 0 | 1 |
| **2762** | 2009-02-19 | 0.000000 | 0.005442 | 0 | 0 | 0 |
| **2763** | 2009-02-18 | -0.001359 | -0.009434 | 0 | 0 | 0 |
| **2764** | 2009-02-17 | 0.026279 | -0.022398 | 1 | 0 | 1 |
| **2765** | 2009-02-16 | 0.060056 | -0.005242 | 0 | 0 | 0 |

```
BuyOrnot = 1
390
```

參考D檔投資法與10天前的淨值比較，若跌幅超過10%則標記1，反之為0

另外與前一天的淨值作比較，若跌幅超過1%則標記1，反之為0

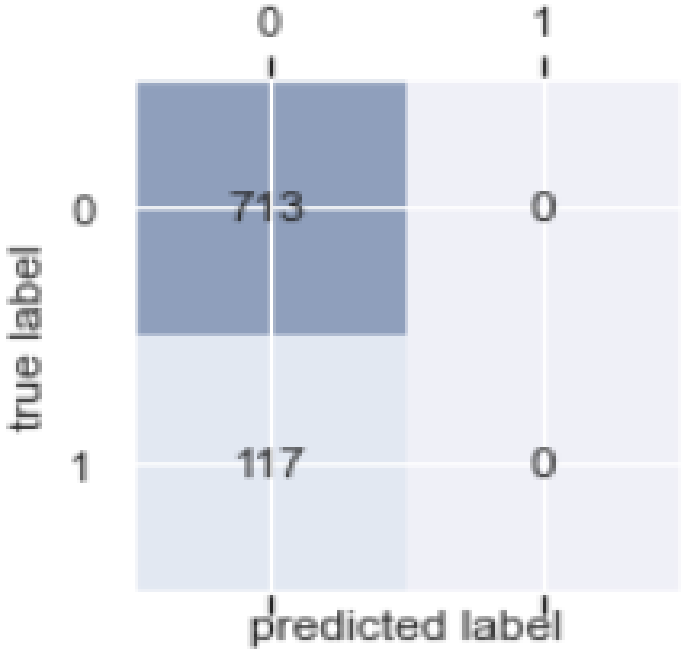
# PCA 因子

| | Date | 幣別 | 淨值(元) | x1 | x2 | x3 | x4 | x5 |
|---|---|---|---|---|---|---|---|---|
| **0** | 2020-06-02 | NTD | 30.85 | -4.698641 | -1.447645 | -4.712656 | 20.986408 | 5.384891 |
| **1** | 2020-06-01 | NTD | 30.49 | -4.698641 | -1.447645 | -4.712656 | 20.986408 | 5.384891 |
| **2** | 2020-05-29 | NTD | 30.27 | -4.698641 | -1.447645 | -4.712656 | 20.986408 | 5.384891 |
| **3** | 2020-05-28 | NTD | 30.14 | -4.698641 | -1.447645 | -4.712656 | 20.986408 | 5.384891 |
| **4** | 2020-05-27 | NTD | 29.93 | -4.698641 | -1.447645 | -4.712656 | 20.986408 | 5.384891 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **2771** | 2009-02-06 | NTD | 7.73 | -4.517556 | -1.063704 | -0.224528 | -0.640680 | 0.096782 |
| **2772** | 2009-02-05 | NTD | 7.39 | -4.517556 | -1.063704 | -0.224528 | -0.640680 | 0.096782 |
| **2773** | 2009-02-04 | NTD | 7.36 | -4.517556 | -1.063704 | -0.224528 | -0.640680 | 0.096782 |
| **2774** | 2009-02-03 | NTD | 7.23 | -4.517556 | -1.063704 | -0.224528 | -0.640680 | 0.096782 |
| **2775** | 2009-02-02 | NTD | 7.16 | -4.517556 | -1.063704 | -0.224528 | -0.640680 | 0.096782 |

使用PCA得到的因子來訓練SVM模型分割資料

# 訓練結果

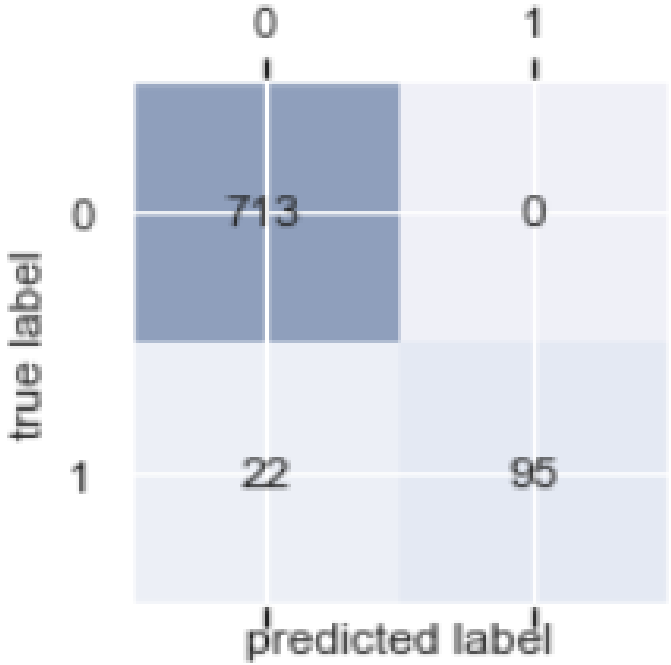
Only PCA




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 1.00 | 0.92 | 713 |
| 1 | 0.00 | 0.00 | 0.00 | 117 |
| accuracy | | | 0.86 | 830 |
| macro avg | 0.43 | 0.50 | 0.46 | 830 |
| weighted avg | 0.74 | 0.86 | 0.79 | 830 |

# 訓練結果

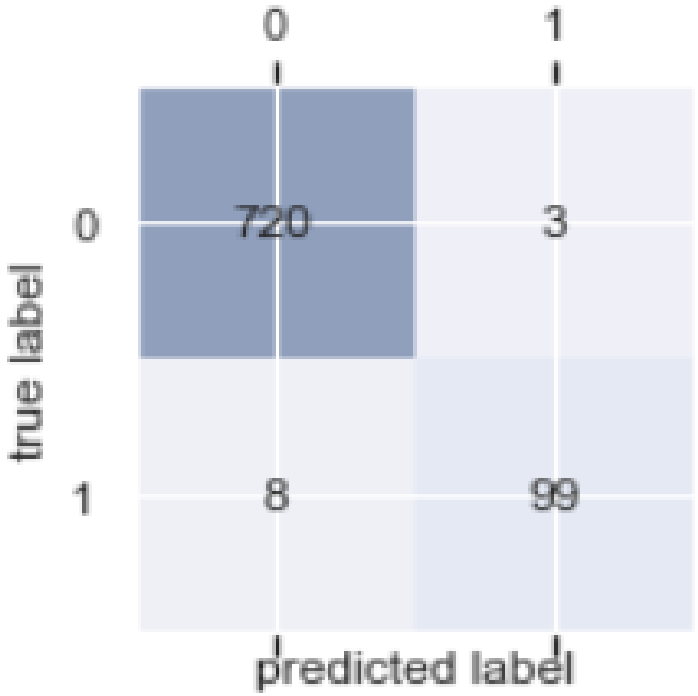
PCA + 1_day growth



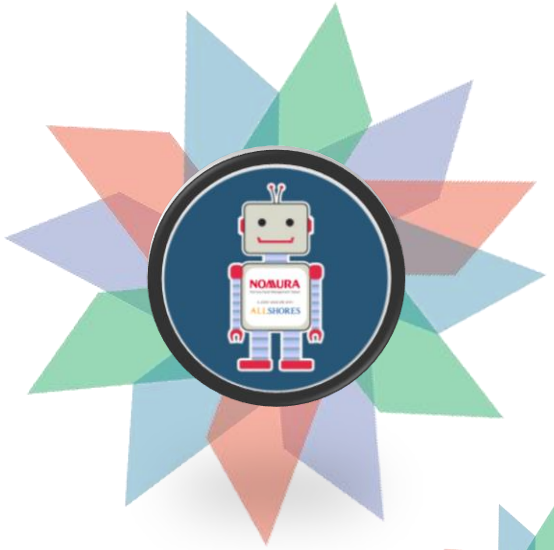

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 1.00 | 0.98 | 713 |
| 1 | 1.00 | 0.81 | 0.90 | 117 |
| accuracy | | | 0.97 | 830 |
| macro avg | 0.99 | 0.91 | 0.94 | 830 |
| weighted avg | 0.97 | 0.97 | 0.97 | 830 |

# 訓練結果

PCA + 1_day growth + 10_day growth




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 1.00 | 0.99 | 723 |
| 1 | 0.97 | 0.93 | 0.95 | 107 |
| accuracy | | | 0.99 | 830 |
| macro avg | 0.98 | 0.96 | 0.97 | 830 |
| weighted avg | 0.99 | 0.99 | 0.99 | 830 |




# Chatbot

Telegram Chatbot : http://t.me/nomura_beta_bot

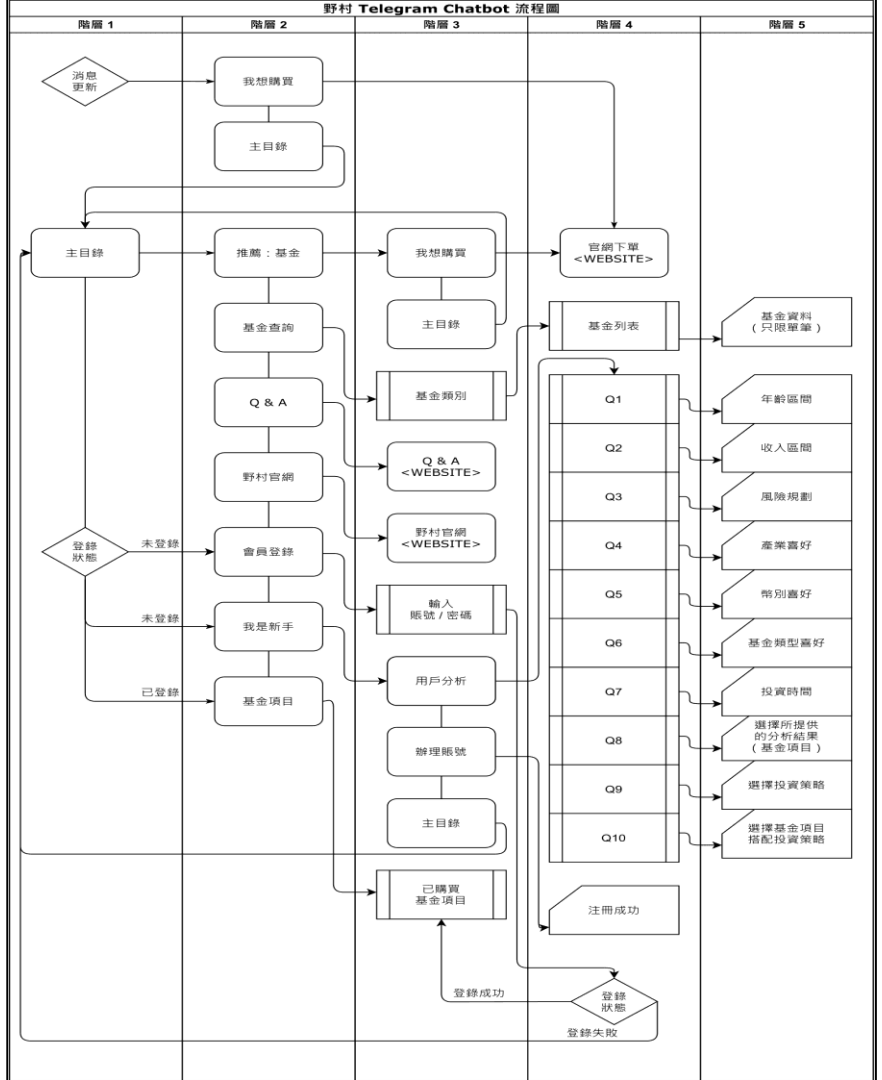
# 野村小夥伴

客戶服務是聊天機器人最有影響力的領域，因此，我們打造了一站式服務的任務機器人，搭配用戶使用場景，創造了以下功能：

- 用戶分析
- 推薦基金
- 辦理賬號
- 基金項目（基金經理人 / 用戶）
- 基金查詢
- 官方網站
- Q & A

# 流程圖

為提供使用者最佳的使用體驗，透過設計流程圖歸納各項按鈕之間的關係。



# Thank you