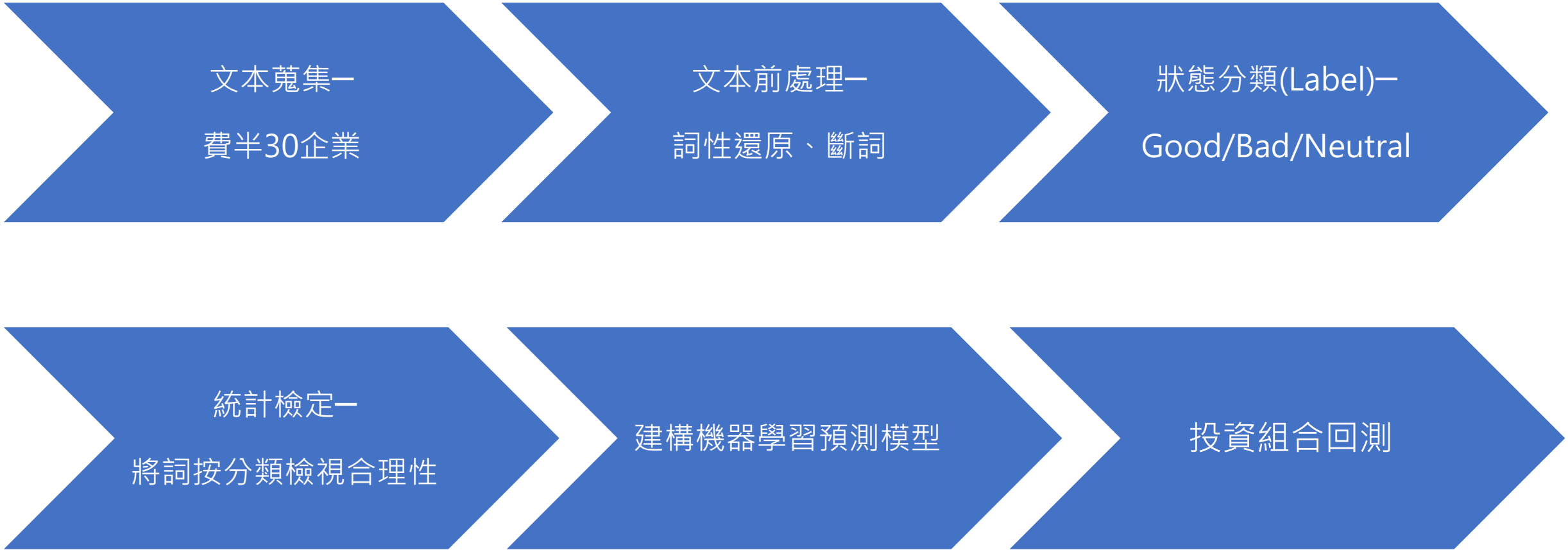
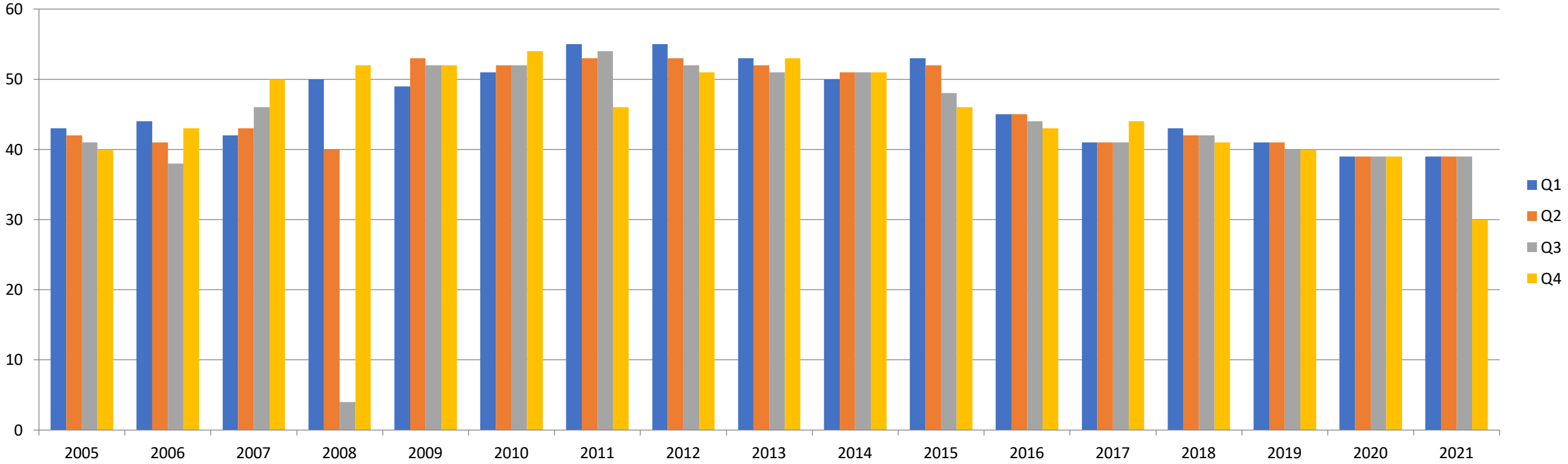
以半導體法說會逐字稿建立投資組合

建置流程


文本蒐集ㄧ
費半30企業
文本前處理ㄧ
詞性還原、斷詞
狀態分類(Label)ㄧ
Good/Bad/Neutral
統計檢定ㄧ
將詞按分類檢視合理性
建構機器學習預測模型
投資組合回測

資料 – 歷史費半企業

- 期間：2005/Q3 – 2021/Q4
- 總公司數：66
- 總文本數：3081



資料 – 歷史費半企業

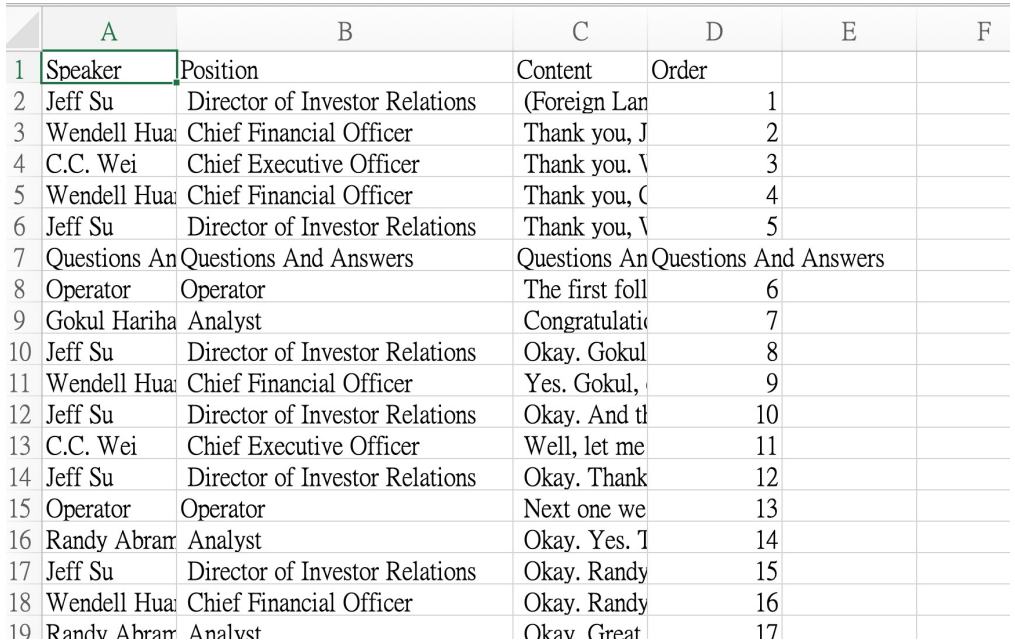
- 上游公司：**INFN、INTC、MU、NSM、NXPI、ON、QRVO、STM、SWKS、TXN、ADI、ALTR、AMD、ARMH、ATHR、ATML、AVGO、BRCM、CAVM、CRUS、CY、FSL、HITT、IDCC、IDTI、IFNNY、IPHI、LLTC、LSCC、MCHP、MLNX、MPWR、MRVL、MSCC、MSI、MXIM、NETL、NVDA、NVLS、POWI、QCOM、RFMD、SIMO、SLAB、SMTC、SNDK、SPRD、XLNX**

- 中游公司：**AMAT、ASML、BRKS、CCMP、CREE、ENTG、KLAC、LRCX、LSI、MKSI、RBCN、SUNEQ、TSM、VECO、IPGP、AZTA、IIVI、WOLF**

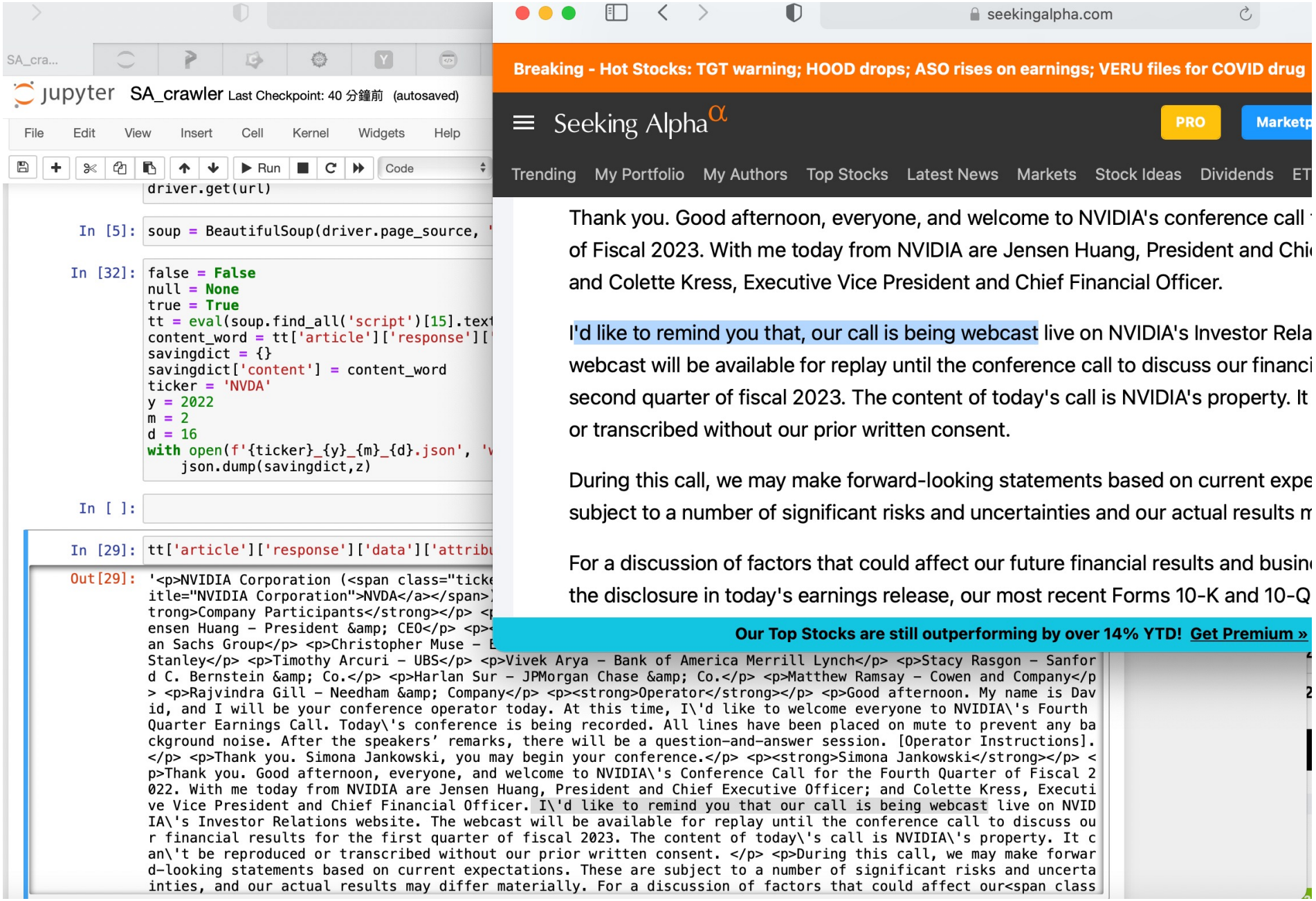
- 下游公司：**TER、AMKR**

文本蒐集—爬蟲(Web Crawler)

- 資料來源：SeekingAlpha、Bloomberg資料庫
- 使用套件：python Selenium + Requests
- 方法：因若需查閱一年以前之法說會文本需使用付費會員，故不能直接透過request方式分析前端網頁代碼，而需要使用Selenium開啟虛擬瀏覽器後，登入會員，分析網頁格式才能　取得需要之文本內容。取得內容後需進行文本處理，分辨發言者與發言內容，存入設定之格式。



	A	B	C	D	E	F
1	Speaker	Position	Content	Order		
2	Jeff Su	Director of Investor Relations	(Foreign Lan	1		
3	Wendell Hua	Chief Financial Officer	Thank you, J	2		
4	C.C. Wei	Chief Executive Officer	Thank you. V	3		
5	Wendell Hua	Chief Financial Officer	Thank you, (	4		
6	Jeff Su	Director of Investor Relations	Thank you, V	5		
7	Questions An	Questions And Answers	Questions An	Questions And Answers		
8	Operator	Operator	The first foll	6		
9	Gokul Hariha	Analyst	Congratulati	7		
10	Jeff Su	Director of Investor Relations	Okay. Gokul	8		
11	Wendell Hua	Chief Financial Officer	Yes. Gokul,	9		
12	Jeff Su	Director of Investor Relations	Okay. And th	10		
13	C.C. Wei	Chief Executive Officer	Well, let me	11		
14	Jeff Su	Director of Investor Relations	Okay. Thank	12		
15	Operator	Operator	Next one we	13		
16	Randy Abram	Analyst	Okay. Yes. T	14		
17	Jeff Su	Director of Investor Relations	Okay. Randy	15		
18	Wendell Hua	Chief Financial Officer	Okay. Randy	16		
19	Randy Abram	Analyst	Okay. Great	17		


jupyter SA_crawler Last Checkpoint: 40 分鐘前 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run Code
driver.get(url)
In [5]: soup = BeautifulSoup(driver.page_source,
In [32]: false = False
null = None
true = True
tt = eval(soup.find_all('script')[15].text
content_word = tt['article']['response']['
savingdict = {}
savingdict['content'] = content_word
ticker = 'NVDA'
y = 2022
m = 2
d = 16
with open(f'{ticker}_{y}_{m}_{d}.json', 'w
json.dump(savingdict,z)
In [ ]:
In [29]: tt['article']['response']['data']['attribu
Out[29]: ' NVIDIA Corporation (<span class="ticke
itle="NVIDIA Corporation">NVDA)
trong>Company Participants <p
ensen Huang - President & CEO <
an Sachs Group Christopher Muse - B
Stanley Timothy Arcuri - UBS Vivek Arya - Bank of America Merrill Lynch Stacy Rasgon - Sanfor
d C. Bernstein & Co. Harlan Sur - JPMorgan Chase & Co. Matthew Ramsay - Cowen and Company</p
> Rajvindra Gill - Needham & Company Operator Good afternoon. My name is Dav
id, and I will be your conference operator today. At this time, I\'d like to welcome everyone to NVIDIA\'s Fourth
Quarter Earnings Call. Today\'s conference is being recorded. All lines have been placed on mute to prevent any ba
ckground noise. After the speakers' remarks, there will be a question-and-answer session. [Operator Instructions].
 Thank you. Simona Jankowski, you may begin your conference. Simona Jankowski <
p>Thank you. Good afternoon, everyone, and welcome to NVIDIA\'s Conference Call for the Fourth Quarter of Fiscal 2
022. With me today from NVIDIA are Jensen Huang, President and Chief Executive Officer; and Colette Kress, Executi
ve Vice President and Chief Financial Officer. I\'d like to remind you that our call is being webcast live on NVID
IA\'s Investor Relations website. The webcast will be available for replay until the conference call to discuss ou
r financial results for the first quarter of fiscal 2023. The content of today\'s call is NVIDIA\'s property. It c
an\'t be reproduced or transcribed without our prior written consent. During this call, we may make forwar
d-looking statements based on current expectations. These are subject to a number of significant risks and uncerta
inties, and our actual results may differ materially. For a discussion of factors that could affect our<span class
seekingalpha.com
Breaking - Hot Stocks: TGT warning; HOOD drops; ASO rises on earnings; VERU files for COVID drug
Seeking Alpha
PRO
Trending My Portfolio My Authors Top Stocks Latest News Markets Stock Ideas Dividends
Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call
of Fiscal 2023. With me today from NVIDIA are Jensen Huang, President and Chi
and Colette Kress, Executive Vice President and Chief Financial Officer.
I'd like to remind you that, our call is being webcast live on NVIDIA's Investor Rela
webcast will be available for replay until the conference call to discuss our financi
second quarter of fiscal 2023. The content of today's call is NVIDIA's property. It
or transcribed without our prior written consent.
During this call, we may make forward-looking statements based on current expe
subject to a number of significant risks and uncertainties and our actual results m
For a discussion of factors that could affect our future financial results and busin
the disclosure in today's earnings release, our most recent Forms 10-K and 10-Q
Our Top Stocks are still outperforming by over 14% YTD! Get Premium »

文本前處理─詞性還原(Lemmatization)

英文單字會因時態、單複數不同而變化，若不處理會造成文字探勘研究的偏誤，例如 the performance looks good 和 the performance is better than last year 兩句話的 good 和 better 是比較級關係，卻會被當成兩個不同的單字

- 使用套件：NLTK + Stanza(美國 Stanford大學開發之語言處理套件)
  - 以 it’ s better than before 為例
  - NLTK: it ‘s good than before
  - Stanza: it be better than before
- Stanza 無法處理形容詞之詞性還原、NLTK不夠細緻，縮寫(ex. 無法處理 It ‘s)
- 目標：

  went/ goes → go
  cars → car
  better → good

文本前處理–斷詞(Segmentation)

先進行各種文本預處理，例如透過人工標記的方式保留完整片語、去除符號及stop words，使研究更精確

- 使用套件：NLTK
- 斷詞預處理：去除符號及stopwords 後，在保留片語的前提下將句子斷成單詞
- 以 However, there are a lot of companies doing this!為例

處理順序

  - 詞性還原後的句子: however, there be a lot of company do this!
  - 去除符號及 stopwords：however there a lot of company do this
  - 保留片語進行斷詞： however, there, a lot of, company, do, this
  - 若不保留片語語意會不精準：however, there, a, lot, of, company, do, this → 得到一串詞的 list 以進行後續分析

- Stop words 定義(Stanford)：some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called *stop words* .

狀態分類(Label)─Good/Neutral/Bad

隨機亂數抽取不同法說會之段落，並以個人主觀之方式給予Good/Neutral/Bad之標記，之後針對各字詞檢視是否在Good/Bad狀態差異顯著下具合理性，若為不具合理性之字詞則反視原段落之該詞前文是否有否定詞，若有則改變該詞為non-_____。

Good 範例字

```
t1,p1 ,t2,p2= check_words_t_test_result('enhancing')
print(f'good bad t-value :{t1}')
print(f'good bad p-value :{p1}')
print(f'good neu t-value :{t2}')
print(f'good neu p-value :{p2}')
```

```
good bad t-value :5.240565007
good bad p-value :0.000968667
good neu t-value :2.922203127
good neu p-value :0.021574189
```

Bad 範例字

```
t1,p1 ,t2,p2= check_words_t_test_result('distortion')
print(f'good bad t-value :{t1}')
print(f'good bad p-value :{p1}')
print(f'good neu t-value :{t2}')
print(f'good neu p-value :{p2}')
```

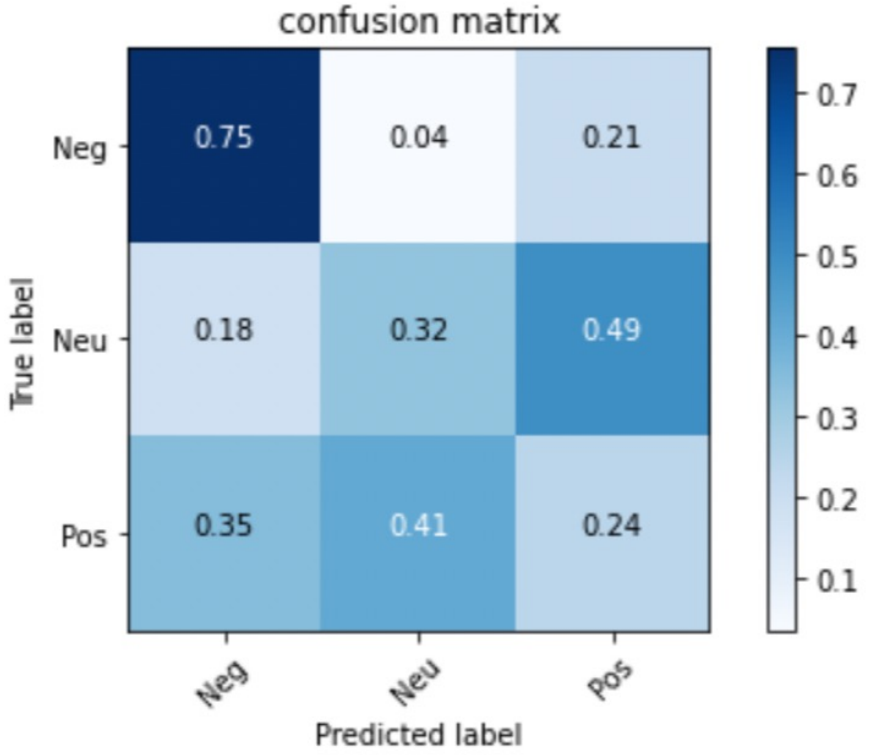
```
good bad t-value :-5.749282571
good bad p-value :0.005226515
good neu t-value :-4.082482905
good neu p-value :0.076464126
```

BERT model (base-cased)

- 訓練集: 1250段(pos:539, neu:438, neg:273)
- Test: pos: 20/83, neu: 44/136, neg: 61/81

	precision	recall	f1-score	support
Neg	0.53	0.75	0.62	81
Neu	0.54	0.32	0.41	136
Pos	0.19	0.24	0.21	83
accuracy			0.42	300
macro avg	0.42	0.44	0.41	300
weighted avg	0.44	0.42	0.41	300

****



投資組合回測

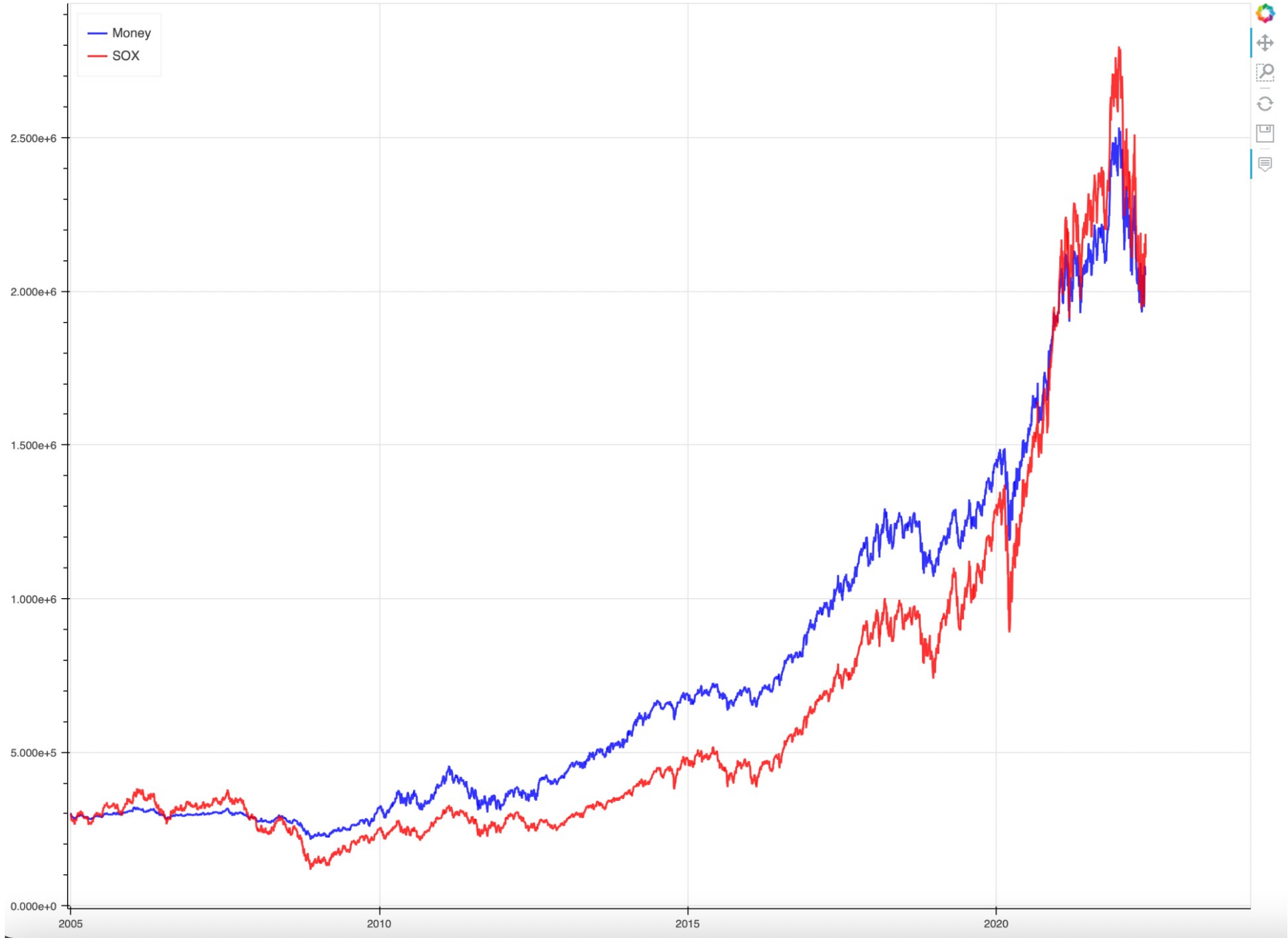
- 初始資金為30萬，平均持有當時費半30所有成分股
- 若模型判別文本分數情緒總和低於-0.25則該公司剔除於該季投資組合，直至下次法說會
- 若該公司已被排除在費半30指數之外，則剔除該公司於投資組合中
- 總文本數: 3081 (non-neg:2515, neg:566)，2004/12/31 – 2022/06/01

	Type	Content
2	Total_Return	584.5297545
3	Internal_Rate_Return	11.66782096
4	Volatility	0.985392058
5	Sharpe_Ratio	0.565496933
6	Max_Drawdown	-0.332013864
7	Max_Drawdown_Start	2011/2/17 00:00
8	Max_Drawdown_End	2011/10/3 00:00
9	Max_Drawdown_Continuos_Days	228 days 00:00:00
10	Profit_Risk_Ratio	17.6055827

- 投組各項數值回測

	Type	Content
2	Total_Return	603.660207
3	Internal_Rate_Return	11.84454862
4	Volatility	1.539003314
5	Sharpe_Ratio	0.438842248
6	Max_Drawdown	-0.689023588
7	Max_Drawdown_Start	2006/1/27 00:00
8	Max_Drawdown_End	2008/11/20 00:00
9	Max_Drawdown_Continuos_Days	1028 days 00:00:00
10	Profit_Risk_Ratio	8.761096397

- 費半30各項數值回測


Money
SOX
2.500e+6
2.000e+6
1.500e+6
1.000e+6
5.000e+5
0.000e+0
2005
2010
2015
2020

結論

- 該投組交易方式能降低波動度，提升夏普值
- 能反應產業實體經濟，然而不能完全反應總體經濟狀況
- 法說會屬於落後資訊，並不能完全避免跌勢
- 可以搭配其他量化資訊做參考，改進投組表現

改善目標

- 嘗試更多種類模型(XLNet、ELECTRA……)，並加入文本內對於產業數值的情緒判別
- 加入更多其他量化數據一同參考，進行樣本外測試
- 擴增標註文本數量
- 嘗試不同投資交易及調整不同參數方式

最新投組

	F_name	result	Date
2	TXN	-0.33152174	2022/4/26
3	TSM	-0.31132075	2022/4/14
4	SLAB	-0.30872483	2022/4/27
5	WOLF	-0.26875	2022/5/4
6	AMKR	-0.25675676	2022/5/2
7	POWI	-0.24324324	2022/4/28
8	QRVO	-0.24242424	2022/5/4
9	TER	-0.24186047	2022/4/27
10	MCHP	-0.23715415	2022/5/9
11	LSCC	-0.23404255	2022/5/3
12	ADI	-0.22839506	2022/5/18
13	AZTA	-0.22429907	2022/5/9
14	ASML	-0.21649485	2022/4/20
15	SWKS	-0.21481481	2022/5/3
16	ON	-0.20647773	2022/5/2
17	LRCX	-0.19655172	2022/4/20
18	IPGP	-0.19230769	2022/5/3
19	NXPI	-0.19130435	2022/5/3
20	AVGO	-0.19117647	2022/5/26
21	IIVI	-0.1862069	2022/5/10
22	KLAC	-0.17818182	2022/4/28
23	ENTG	-0.15625	2022/4/26
24	INTC	-0.15151515	2022/4/28
25	MRVL	-0.15062762	2022/5/26
26	AMAT	-0.14893617	2022/5/19
27	MU	-0.14285714	2022/3/29
28	QCOM	-0.13989637	2022/4/27
29	NVDA	-0.12903226	2022/5/25
30	MPWR	-0.12385321	2022/5/2
31	AMD	-0.10507246	2022/5/3

灰色區塊為本季未納入投組之個股

Thank You