

Fintech-Text Mining and Machine Learning

智能新聞評分系統 第二組

2020/6/18

指導業師:詹益安

指導老師:蔡芸琤

本組組員:

東吳巨資系大四 劉品妤台大國企所碩二 王昱達

台大經濟系大四 楊廣元 台大會計所碩二 呂明諺



CONTENTS

01 現

現有問題描述與本專案優勢

02

專案流程圖

03

資料及樣態說明

04

成果展現與介紹

01

現有問題描述 & 本專案優勢

新聞重要性篩選 評分延遲 專家看法不一致 準確的新聞評分 過濾不重要的訊息 提供即時資訊 可作為投資參考工具





- 訊息來源眾多
- 那些新聞重要?

海量新聞資訊 專家看法不一致 評分延遲 現有系統每日下午 • 專家經驗判斷水平 才發佈對最近一日 不一致 的新聞評價 • 沒有一致的評斷標 準 無法及時掌握新聞 之影響

本專案產品用途與競爭優勢





優勢一 準確的新聞評分:



透過機器學習模型,避免多位專家評分看法不一致的偏誤





專家經驗判斷水平不一 致

機器學習



利用機器學習參考各專 家評分結果,降低評分 變異程度

優勢二 過濾不重要的訊息: 建立篩選機制,讓使用者只看得到重要的新聞





ن ن ن



那些新聞重要?

重要新聞篩選機制



系統只展現重要的新聞 內容及評分(-3, -2, 2, 3)

優勢三 提供即時資訊: 爬取最新新聞,即時更新新聞評分



評分延遲



無法及時掌握新聞之影響

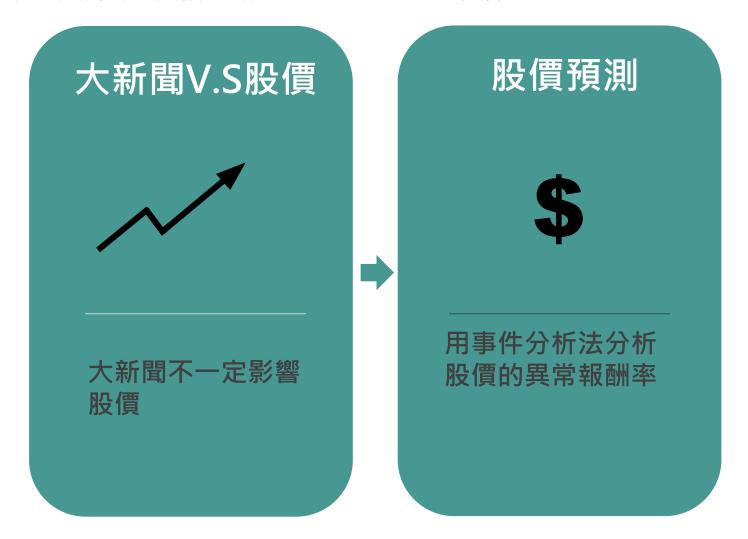
新聞爬蟲



針對公開資訊觀測 站,每十秒爬一次, 即時更新

優勢四 可作為投資參考工具: 透過預測預期漲跌幅指標,作為投資參考





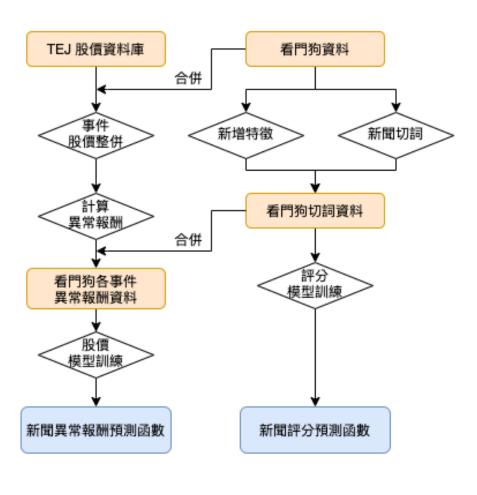
02

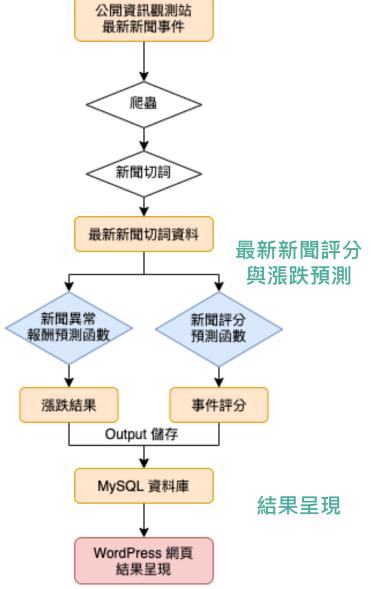
專案流程圖



專案流程圖

模型訓練過程





03

資料及樣態說明



資料集樣態說明

/	A	В	С	D	Е	F	G	Н	I	J	K
1	個股代號	公司簡稱	事件日	TCRI(年/月)	事件強度	大事件類別	小事件類別	事件內容			
2	1218	泰山	20190101	6(2018/09)	0	M_經營層	MT06_高管異動	發言人林俐婉內部調動	,由江巍峰	接任。。	
3	1503	士電	20190101	4(2018/09)	0	M_經營層	MT06_高管異動	内部稽核主管林志強內	部調動,由	莊文清接任	E · ·
4	1504	東元	20190101	4(2018/09)	0	M_經營層	MT06_高管異動	會計主管藍俊雄內部調	動,由林鴻	3名接任。。	
5	1709	和益	20190101	5(2018/09)	0	M_經營層	MT06_高管異動	内部稽核主管游本詮內	部調動,由	曾筱茜接任	E · ·
6	1721	三晃	20190101	7(2018/09)	0	M_經營層	MT06_高管異動	財務經理洪廷宜內部調	動,由王妈	渝接任。。	
7	1817	凱撒衛	20190101	6(2018/09)	0	M_經營層	MT06_高管異動	研發主管吳政峰內部調	動,由朱清	î 立接任。。	
8	2064	晉椿	20190101	7(2018/09)	-1	M_經營層	MT06_高管異動	總經理高進義離職,由	陳譽接任。	。發言人高	5進義離職,日
9	2207	和泰車	20190101	4(2018/09)	0	M_經營層	MT02_董監異動	改派1董。董事大野勝仁	二(豐田自動	車代表)卸任	E。董事長沼-
10	2330	台積電	20190101	1(2018/09)	-1	M_經營層	MT02_董監異動	辭任1董。獨立董事湯馬	斯?延吉布	斯卸任。	
11		華碩	20190101	2(2018/09)	0	M_經營層	MT06_高管異動	總經理沈振來內部調動			
12	2377	微星	20190101	3(2018/09)	0	M_經營層	MT06_高管異動	總經理徐祥內部調動,			
13	2442	新美齊	20190101	7(2018/09)	0	M_經營層	MO04_經營權轉讓疑慮	新美齊澄清報載提及本	公司停業及	、隠射屬大同	集團上市櫃2
14	2724	富驛-KY	20190101	D(2018/09)	0	M_經營層	MT02_董監異動	改派1董。董事周威良(F	urama Hotel	l International	Management In
15	2750	桃禧	20190101		0	M_經營層	MT06_高管異動	會計主管林慧茹內部調	動,由楊崑	岳接任。。	
16	2852	第一保	20190101		0	M_經營層	MT06_高管異動	財務經理李易致內部調	動,由施冬	森接任。。	
17	2888	新光金	20190101		0	M_經營層	MT06_高管異動	會計主管施貽昶內部調	動,由呂雅	- 遊接任。。	
18	3004	豐達科	20190101	5(2018/09)	0	M_經營層	MT06_高管異動	總經理邱智科內部調動	,由林威村	接任。。	
19	3167	大量	20190101	5(2018/09)	0	M_經營層	MT06_高管異動	研發主管宋漢釧內部調	動,由商國	強接任。。	
20	4104	佳醫	20190101	5(2018/09)	0	M_經營層	MT06_高管異動	總經理高省內部調動,	由張明正接	任。。	
21	4138	曜亞	20190101	6(2018/09)	0	M_經營層	MT06_高管異動	總經理傅若軒內部調動	,由吳國龍	接任。。	
22	4168	醣聯	20190101	7(2018/09)	0	M_經營層	MT06_高管異動	總經理張東玄內部調動	,由楊玫君	接任。。	
23	4402	福大	20190101	D(2018/09)	0	M_經營層	MT06_高管異動	會計主管莊清揚內部調	動,由何子	龍接任。。	

• 資料重要內容

A. 事件強度

B. 大事件類別

C. 小事件類別

D. 事件內容

• 資料樣本數

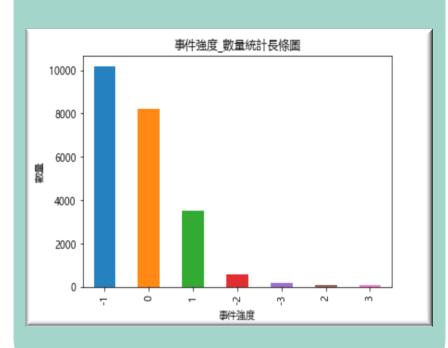
A. 2019(1-12): 23,703

B. 2020(1-3): 13,104

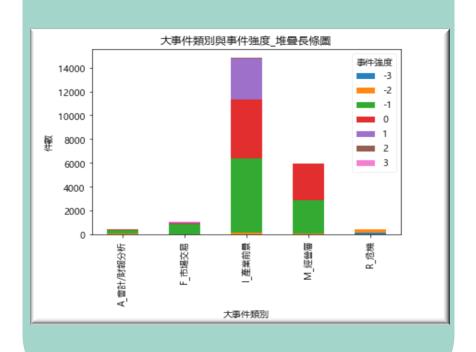


資料集樣態說明

事件強度本身樣本分布 不均



大事件類別事件強度分布 有所差異



04

成果展現與介紹

斷詞、爬蟲、新聞評分模型

股價預測模型、資料庫建立、網頁呈現



新聞資料爬蟲

- 本組採用 pandas 套件中的 read_html 功能,對公開資 訊觀測站的重大訊息主旨做 爬蟲
- 設定成每 10 秒爬一次最新 消息,將新增的新聞訊息加 入資料庫
- 後續將新聞訊息做切詞處理 後,放入模型跑重要性評分 與股價漲跌預測



E.SUN

資料預處理(斷詞)



- 詞是最小有意義且可以自由使用的語言單位
- 任何語言處理的系統都必須先能分辨文本中的詞才能進 行進一步的處理
- 將一段中文切分出有「意義」的小單位(詞)



• 本次報告須做處理的部分為新聞的內容



• 分別使用2種不同的斷詞系統做新聞斷詞



Jieba vs Ckiptagger



Jieba



Ckiptagger

- Jieba 這個中文斷詞程式 是由中國開發者所開發
- 可同時支援簡體與繁體的 斷詞
- 中研院CKIP Lab中文詞 知識庫小組開發之中文斷 詞工具

TOOL	(WS) PREC	(WS) REC	(WS) F1
Ckiptagger	97.49%	97.17%	97.33%
Jieba	90.51%	89.10%	89.80%



Ckiptagger 🚡



- Ckiptagger 斷詞結果較準確
- 使用 Ckiptagger 斷詞後模型預測結果較佳
- 在切詞工具方面我們選擇 Ckiptagger



模型建立

利用長短期記憶模型(LSTM)建立:

大事件類別分類器

小事件類別分類器

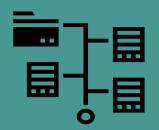
事件強度分類器

股價異常報酬分類器

大事件分類器:資料分割與不平衡資料處理



大事件分類器



將新聞分類為以下五個 大事件類別:

- 'A 會計/財報分析'
- 'F 市場交易'
- 'I_產業前景'
- 'M 經營層'
- 'R_危機'

資料分割



- 所有資料的64%作為訓練資料
- 所有資料的16%作為 驗證集
- 所有資料的20%作為 測試集

不平衡資料處理



由於大事件類別的分 布相當不平衡,所以 我們使用了以下兩種 方法來處理資料不平 衡的問題

- 1. 使用Oversampling
- 2. 調整損失函數 (loss function)的 權重

大事件分類器:模型架構



Layer (type)	Output	Shape	Param #
embedding_2 (Embedding)	 (None,	None, 128)	1280000
lstm_2 (LSTM)	(None,	16)	9280
dense_2 (Dense)	(None,	5)	85
Total params: 1,289,365 Trainable params: 1,289,365 Non-trainable params: 0			

Embedding 用來進行詞嵌入 layer

LSTM layer 長短期記憶模型

Dense layer 作為此模型的output layer

大事件分類器:模型表現(在驗證集上)



Accuracy: 0.971

	A_會計/財 報分析	F_市場交易	I_產業前景	M_經營層	R_危機
Precision	0.672	0.995	0.979	0.959	0.888
Recall	0.741	0.974	0.982	0.960	0.879
F1 score	0.705	0.984	0.980	0.960	0.883

小事件分類器



小事件分類器



Created by Eucalyp

利用新聞中的文字資料,將新聞分類為以下15小事件類別:

- MT02_董 監異動
- MT06_高 管異動
- 經營層_其他

- FS02_股 價暴跌或 異常
- FS03_其 他市場交 易議題
- 市場交易_ 其他

- AF05_財 務警示
- AI01_延 遲公告
- 會計/財報分析 其他

- RB01_TC RI負向觀 察
- RB02_TC RI降等
- 危機_其 他

- IP01_成本 /產能變動 或資本支 出
- IS01_營收 變動或客 戶/商品/ 通路策略
- 產業前景_ 其他



資料分割



- 所有資料的64%作為訓 練資料
- · 所有資料的16%作為驗 證集
- 所有資料的20%作為測 試集

不平衡資料處理



由於小事件類別的分 布相當不平衡,所以 我們使用了以下兩種 方法來處理資料不平 衡的問題

- 1. 使用Oversampling
- 2. 調整損失函數 (loss function)的 權重

小事件分類器:模型架構



Layer (type) ====================================		•	Param # ======== 1280000	Embedding layer	用來進行詞嵌入
lstm_6 (LSTM)	(None,	64)	49408		
dense_15 (Dense)	(None,	32)	2080	LSTM layer	長短期記憶模型
dense_16 (Dense)	(None,	32)	1056	23 Tivi layer	
dense_17 (Dense)	(None,	15)	495		
Total params: 1,333,039 Trainable params: 1,333,039 Non-trainable params: 0				Dense layer (3 dense layers):	進行小事件類別的分 類

小事件類別分類器:模型表現(在驗證集上)



Accuracy: 0.905

	AF05_財務 警示	AI01_延遲 公告	FS02_股價 暴跌或異常	FS03_其他 市場交易議 題	IP01_成本/ 產能變動或 資本支出	IS01_營收 變動或客戶 /商品/通路 策略	MT02_董 監異動
Precision	0.128	1	1	0.8	0.797	0.943	1
Recall	0.451	0.375	0.999	0.333	0.701	0.945	0.995
F1 score	0.191	0.545	0.999	0.47	0.746	0.944	0.997

	MT06_高 管異動	RB01_TC RI負向觀 察	RB02_TC RI降等	危機_其 他	市場交易 _其他	會計/財 報分析_ 其他	產業前景 _其他	經營層_ 其他
Precision	1	0.806	0.888	0.941	0.8	0.7	0.759	0.907
Recall	0.995	0.781	0.8	0.592	0.47	0.491	0.839	0.87
F1 score	0.997	0.793	0.842	0.727	0.592	0.577	0.797	0.888

事件強度分類器:資料分割與不平衡資料處理



事件強度分類器



Created by Eucalyp

利用新聞中的文字資料, 將新聞分類為以下七個事 件強度:

- 3
- 2
- 1
- (
- -1
- -2
- -3

資料分割



- 所有資料的64%作為訓練 資料
- 所有資料的16%作為驗證 集
- 所有資料的20%作為測試 集

不平衡資料處理



由於事件強度的分布 相當不平衡(極端事件:-3,-2,+2,+3 出 現的頻率相對較少) 所以我們使用了以下 兩種方法來處理資料 不平衡的問題

- 1. 使用Oversampling
- 2. 調整損失函數 (loss function)的 權重

事件強度分類器:模型架構



Layer (type)	Output	•	Param #	- I II:	
embedding_3 (Embedding)		None, 128)	1280000	Embedding layer	用來進行詞嵌入
lstm_3 (LSTM)	(None,	16)	9280	I CTN 4 I	
dense_3 (Dense)	(None,	7)	119	LSTM layer	長短期記憶模型
Total params: 1,289,399 Trainable params: 1,289,399 Non-trainable params: 0				Dense layer	作為此模型的 output layer

事件強度分類器:模型表現(在驗證集上)

✓ E.SUN

accuracy: 0.878

	-3	-2	-1	0	1	2	3
precision	0.333	0.629	0.957	0.829	0.789	0.857	1.
recall:	0.710	0.542	0.941	0.843	0.770	0.5	1.
F1 score:	0.454	0.582	0.949	0.836	0.779	0.632	1.



股價預測核心方法論——事件研究法

何謂事件研究?

事件研究法(Even Study) 為研究結果之驗證方法,其起源於1960年代 Ball and Brown,及Fama, Fisher, Jensen and Roll(沈中華、李建然, 2000),為近代會計及財務領域實證研究所廣泛運用之研究設計之一

事件研究的目的?

事件研究法(Even Study) 主要目在於利用統計方法檢定異常報酬狀況,藉以明瞭特定事件是否對公司股價造成影響,並可以了解股價的波動與該事件是否相關

事件研究流程

- 1. 決定事件與事件日
- 2. 估計異常報酬率



事件日、事件期、估計期之定義



事件日

如何設定事件期、估計期長度

- 事件影響的區間應包括在事件期之內,如新聞發布之日。通常事件期間比發生日期 (事件日)更寬廣一些,包括事件發生前後的一段時間。因為事件發生後一段時間的 資訊能顯示應變數(如盈利、股價)變化的情況;而考察事件發生後一段時間的股價 則有利於捕捉事件前徵兆與事前洩漏資訊所造成的影響。
- 估計期間或稱估計窗口(estimation window)的目的,是利用該期間的數據去估算在事件未出現情況下應變數之值,即預期報酬率。將預期報酬率與事件期間應變數變異後(即實際報酬率)相比較,變得出事件所帶來的異常報酬率。
- 一般而言,估計期選取要比事件期間長,本組採用年(250個交易日)、季(60個交易日)、月(20個交易日)三個區間去估計異常報酬,而事件期則是事件日前後一天(明日收盤價-昨日收盤價)



異常報酬計算結果

- 本組採用TEJ股價資料,將所有看門狗中所有事件對應的股價資料合併起來,進 而依序計算三種估計期的正常報酬與異常報酬
- 由於事件期長度定為兩日,因此所有的報酬率皆以兩日報酬率的形式去做計算
- 根據初步的模型預測結果發現,事件期定為一個月的準確率較高,因此後續的 股價預測模型皆採用「一個月事件期」的報酬率資料

個股代號	公司簡稱	事件日	TCRI(年/ 月)	事件強度	收盤價	明日收 盤價/ 昨日收 盤價	明日收 盤價 / 昨日收 盤價 - 1	兩日報 酬率(%)	前5~245日平 均兩日報酬率 (%),年平均正 常報酬(%)	前5~65日平均 兩日報酬率(%) ,季平均正常 報酬(%)	前5~25日平均 兩日報酬率(%) ,月平均正常 報酬(%)	前5~245日(年 平均)兩日異常 報酬(%)	前5~65日(季平 均)兩日異常報 酬(%)	前5~25日(月平 均)兩日異常報 酮(%)	大事件 類別	小事件類別	事件内容	切詞詞 數	切詞結果
1218	泰山	20190101	6(2018/09)	0	19	0.972973	-0.02703	-2.7027	0.072008802	-0.126391428	0.91151213	-2.774711504	-2.576311275	-3.614214833	M_經營層	MT06_高	發言人材	19	['發盲人', '7
1503	士電	20190101	4(2018/09)	0	40.95	0.99754	-0.00246	-0.246	-0.000782104	-0.459070858	-0.042079001	-0.245220356	0.213068398	-0.203923459	M_經營層	MT06_高	内部稽核	22	[内部], 稽
1504	東元	20190101	4(2018/09)	0	17.45	0.994302	-0.0057	-0.5698	-0.373847067	-0.699126047	0.436006766	-0.195953503	0.129325477	-1.005807335	M_經營原	MT06_高	會計主管		[會計,'主
1709	和益	20190101	5(2018/09)	0	14.8	1.003367	0.003367	0.3367	-0.097104514	-0.364537588	-0.160869687	0.433804851	0.701237924	0.497570024	M_經營原	MT06_高	内部稽核	22	[内部], 稽
1721	三晃	20190101	7(2018/09)	0	10.05	1	0	0	-0.275157057	-0.026064579	0.557660856	0.275157057	0.026064579	-0.557660856	M_經營層	MT06_高	財務經理	20	[財務], '經
1817	凱撒衛	20190101	6(2018/09)	0	36.95	0.989276	-0.01072	-1.07239	-0.15792161	-0.081807145	0.3658829	-0.914464449	-0.990578914	-1.438268959	M_經營原	MT06_高	研發主管	20	[研發,'主
2064	晉楷	20190101	7(2018/09)	-1	15.55	1.01634	0.01634	1.633987	-0.069460762	-0.219389167	-0.009464759	1.70344769	1.853376095	1.643451687	M_經營原	MT06_高	總經理高		[總經理', ']
2207	和泰車	20190101	4(2018/09)	0	255.5	0.988	-0.012	-1.2	-0.262075048	-0.048626627	1.179844775	-0.937924952	-1.151373373	-2.379844775	M_經營原	MT02_董	改派1董	41	[改派', '1',
2330	台積電	20190101	1(2018/09)	-1	225.5	0.984305	-0.0157	-1.56951	-0.024201137	-0.551439256	0.027877375	-1.545305589	-1.01806747	-1.597384101	M_經營原	MT02_董	辭任1董	20	[辭任', '1',
2357	華碩	20190101	2(2018/09)	0	201.5	1.0175	0.0175	1.75	-0.240915578	-0.822803805	-0.693052721	1.990915578	2.572803805	2.443052721	M_經營原	MT06_高	總經理沈	19	[總經理], "
2377	微星	20190101	3(2018/09)	0	76.4	1.005222	0.005222	0.522193	-0.026185798	-0.34696121	0.924262083	0.548379009	0.869154422	-0.402068872	M_經營原	MT06_高	總經理徐	18	[總經理], "
2442	新美齊	20190101	7(2018/09)	0	12	1.0125	0.0125	1.25	0.460998352	-0.435840484	-0.448641991	0.789001648	1.685840484	1.698641991	M_經營原	MO04_經	新美齊澄	536	[新美齊', '〉

E.SUN

模型建立—股價異常報酬分類器

- 若利用機器學習做股價 迴歸預測,效果其實並 不理想,因此改用二分 法(亦即只預測漲跌)
- 預測股價的異常報酬為正值或負值
- 將「預測股價異常報酬」 的問題視為二分類問題 (正值 or 負值)

股價異常報酬分類器:模型架構

Layer (type)	Output Sh	nape	Param #
embedding_4 (Embedding)	(None, No	one, 128)	1280000
lstm_4 (LSTM)	(None, 64	1)	49408
dense_13 (Dense)	(None, 32	2)	2080
dense_14 (Dense)	(None, 32	2)	1056
dense_15 (Dense)	(None, 32	2)	1056
dense_16 (Dense)	(None, 1))	33
Total params: 1,333,633 Trainable params: 1,333,633 Non-trainable params: 0			

	資料分割狀況	
訓練資料	驗證集	測試集
Training set	Validation set	Testing set
64%	16%	20%

模型架構與表現



模型架構

- ➤ Embedding layer:用來進行詞嵌入(word embedding)
- ➤ LSTM layer:長短期記憶模型
- ➤ Dense layers (4 dense layers): 進行股價異常報酬的分類

模型表現

> Accuracy: 0.596

事件強度類別	負	正
Precision	0.582	0.619
Recall	0.716	0.474
F1 score	0.642	0.537

(為了維持易讀性,將所有數字取四捨五入到小數點後三位數。)

模型架構與表現



模型結果

- 若將評分與異常報酬做迴歸分析,R平方只有0.1,代表平均來說,新聞事件的變動只能解釋10%的股價變動,解釋力不足
- 若只預測上漲與下跌(也就是只分兩個類別),Accuracy Rate 只有 0.596, 還有進步空間

結果解釋

- 新聞事件對股價影響的反應時間極短,可能事件前後幾小時內股價就已經反應完畢,而本組採用的兩日區間過長,造成解釋力不足的結果(但受限於沒有 Intraday 的股價資料,只能這麼做)
- 不重要的新聞事件數占資料絕大部分,而這些新聞對股價影響力不大,卻又大量使用這些資料做機器學習,將評分絕對值高的事件的影響力給稀釋掉

E.SUN

建立資料庫,存取最新新聞斷詞與評分預測結果

- 在將資料輸入上傳到網 頁前,我們會將:
 - 1.股票代碼
 - 2. 公司名稱
 - 3. 發生時間
 - 4. 預測漲跌狀態
 - 5. 事件強度評分
 - 6. 新聞內容
 - 7. 新聞切詞結果

等資料存放在資料庫中,若未來公司需要 使用即可直接存取

比較過後,本組採用 MySQL 作為本組的資 料庫工具



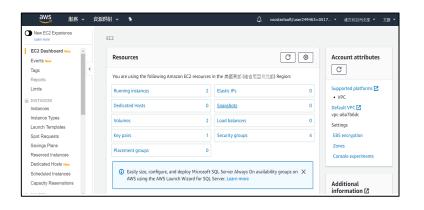
	Google Sheets	MySQL
與Python 連接方式	gspread 套件	pymysql 套件
容量	有容量限制,不適 合放大量數據	無容量限制,適合 放大量數據
存取量	每 100 秒有資料存取量限制,不適合作為存取頻繁的資料庫	無資料存取量限制, 適合作為存取頻繁 的資料庫
費用	免費試用期一年, 之後要收費	開源軟體,免費使 用

✓ E.SUN

透過 WordPress 網頁來呈現我們的預測結果

- ➤ 本組採用 AWS 與 WordPress 架構個股新聞評分 系統網站,並利用 wordpress-xmlrpc 套件完成 自動發文的功能
- ▶ 我們預期將使用爬蟲爬取最新新聞內容後,經過 我們的評分與股價預測系統模型預測,將重要的 新聞(評分絕對值大於1的新聞)呈現在我們的網 站中

AWS EC2控制台



發文更新畫面



手機版首面



小組分工



劉品妤: AWS 與 WordPress 網頁製作、爬蟲最新新聞資料

王昱達:股價資料整理、資料庫建立、網頁串接、簡報整理

楊廣元:新聞評分與股價預測模型的建立與調整參數

呂明諺:切詞與進一步的優化、簡報整理

Trello 連結: Github 連結:

https://reurl.cc/Mvozzv https://reurl.cc/E7vQQk

成果網站連結:

http://ec2-52-87-157-

<u> 212.compute-</u>

1.amazonaws.com/

P.S. 目前最後的資料庫與網頁串接部分尚未完成,因此目前網站僅有架構,整個產品還未正式上線,完成後會在下週以影片來呈現

感謝聆聽 敬請指教