壽險經營管理實務研討期末報告

- 追蹤 Trump 在 Twitter 的發文對市場影響性 -

指導教授: 石百達 教授

指導業師: 張明淇 專案經理

第 11-3 組 電機一 徐楷程

會計三 賴彥良

財金碩一 羅佳敏

財金碩一 楊雅婷

壹、 研究動機

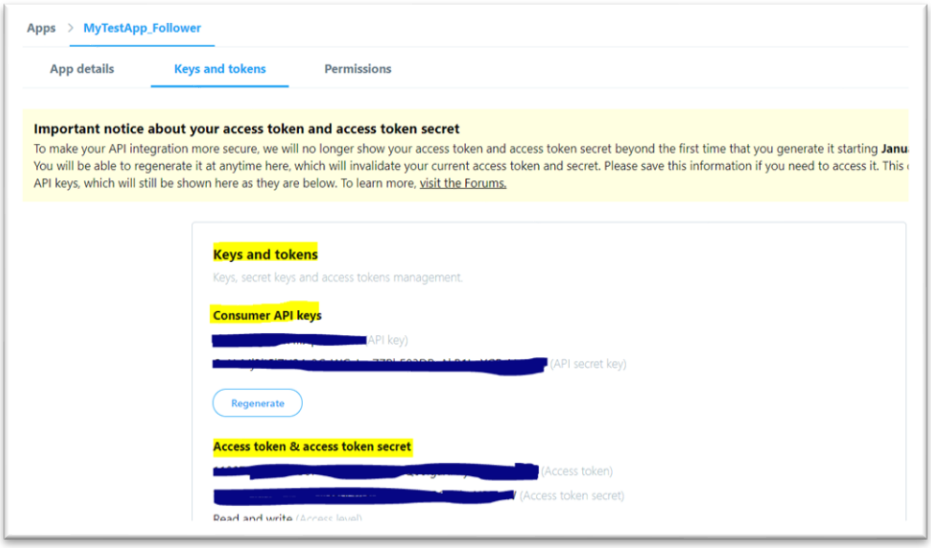
隨著大數據的重要性越來越受關注，運用現有數據來預測未來已經成為趨勢。但數據從何而來?我們又該如何運用數據呢？在 mentor 的引導下，我們這組決定運用彼此所學，將財經相關知識與程式設計做結合，運用 python 蒐集美國總統川普上任以來在 twitter 上的發言，並將蒐集來的資料與股市的波動連結，藉由篩選數據與分析數據建立出一套模型。如此一來，一旦川普在 twitter 上發言，程式就能評估推文對股市的影響，甚至預測未來股市的走向，輔助使用者做出更精明的投資決策。

貳、 研究流程及技術運用

研究流程	技術運用
1. 抓取 Trump 的 Twitter 發文	Twitter API, Python Tweepy, GetOldTweets3
2. 清洗資料並建立資料庫	Mongo DB
3. 自然語言分析	word2vec
4. Label 與資料分類模型	KNN

1. 抓取 Trump 的 Twitter 發文

首先，需註冊 Twitter 和 Twitter Developer 的帳號，註冊成功後，在 Develop 的帳號中創建一個新的 App 以獲取 API 的 Keys and tokens，如下圖:



獲得 Twitter API 所需的 Keys and tokens 後，即可編寫 Python 進行推文抓取，常見內建套件有 Tweepy (須用到 Keys and tokens)和 GetOldTweets3，但由於 Tweepy 有資料期間的限制，因此我們使用 GetOldTweets3，來抓取川普自上任以來(2017/1/20-最新)的推文並匯出成 csv 檔以利後續資料處理，程式碼及解釋請見下圖:

```
import GetOldTweets3 as got
import csv
import datetime
import re
```

```
#定義各參數 (startDate/endDate)
startDate="2017-01-20"
endDate=(datetime.datetime.now()+datetime.timedelta(days=1)).strftime("%Y-%m-%d")

#定義移除圖片的函數
def remove_picture(sample):
    return re.sub(r'pic.twitter.com\S+', "", sample)
```

```
#創建csv檔 (檔名為資料期間)
csvFile = open(startDate+ '_' +datetime.datetime.now().strftime( "%Y-%m %d")
           +'.csv','w',newline='',encoding='utf_8_sig')
csvWriter = csv.writer(csvFile)
```

```
#設定Twitter API抓取的User ID和想抓取的資料期間
tweetCriteria = got.manager.TweetCriteria().setUsername("realDonaldTrump")\
                                .setSince(startDate)\
                                .setUntil(endDate)
tweets = got.manager.TweetManager.getTweets(tweetCriteria);
```


```
#移除推文內容中的圖片以及超連結，並匯出csv檔
for tweet in tweets:
    tweet.text=remove_picture(tweet.text)
    tweet.text=tweet.text.partition("https")[0]
    csvWriter.writerow([tweet.date,tweet.text])

csvFile.close()
```

抓取結果如下圖:

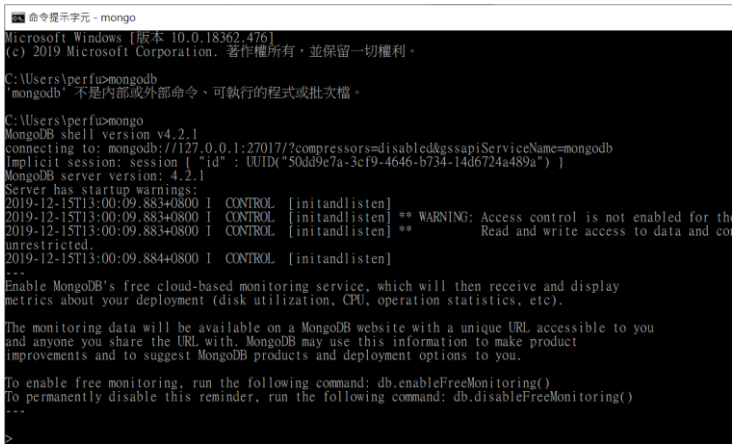
	A	B
1	2019-11-17 13:51:39+00:00	A new Republican Star is born. Great going @EliseStefanik !
2	2019-11-17 13:43:09+00:00	Thank you Pete. Our great warfighters must be allowed to fight. I would not have done this for Sgt. Bergdahl or Chelsea Manning!
3	2019-11-17 12:59:07+00:00	I agree Katrina, Pam Bondi is a great womem!
4	2019-11-17 05:09:17+00:00	Visited a great family of a young man under major surgery at the amazing Walter Reed Medical Center. Those are truly some of the best doctors anywhere in the world. Also
5	2019-11-17 03:11:35+00:00	
6	2019-11-17 02:23:27+00:00	
7	2019-11-17 02:22:20+00:00	
8	2019-11-16 18:48:46+00:00	True!
9	2019-11-16 18:44:16+00:00	Congratulations Kimberley. Great book!
10	2019-11-16 18:38:10+00:00	"I mean, come on. The Democrats are doing focus groups to try and figure out what words to use to move the needle. The Democrats know this is political and they're jus
11	2019-11-16 17:57:51+00:00	"Triggered," a great book by my son, Don. Now number one on @NYTIMES LIST. Keep it there for a while!
12	2019-11-16 15:38:36+00:00	....and Taylor, dismissing everybody involved from the Obama holdover days trying to undermine Trump, getting rid of those people, dismissing them, this is what it looks lil
13	2019-11-16 15:38:34+00:00	"My support for Donald Trump has never been greater than it is right now. It is paramountly obvious watching this, these people have to go. You elected Donald Trump to c
14	2019-11-16 15:18:27+00:00	Dow hits 28,000 - FIRST TIME EVER, HIGHEST EVER! Gee, Pelosi & Schitt have a good idea, "lets Impeach the President." If something like that ever happened, it w
15	2019-11-16 13:51:34+00:00	LOUISIANA, VOTE @EddieRispone TODAY! He will be a great governor!
16	2019-11-16 13:08:14+00:00	Good morning Louisiana! Polls are open at 7AM. Get out and VOTE for @EddieRispone to be your next Gov! He will get your taxes and auto insurance (highest in Country!
17	2019-11-16 02:39:50+00:00	#NewHoaxSameSwamp
18	2019-11-16 00:33:20+00:00	
19	2019-11-16 00:16:42+00:00	THANK YOU! #MAGA #KAG

(資料抓取日:2019/11/17；資料期間:2017/1/20-2019/11/17；空白內容為圖片或影片連結)

2. 清洗資料並建立資料庫

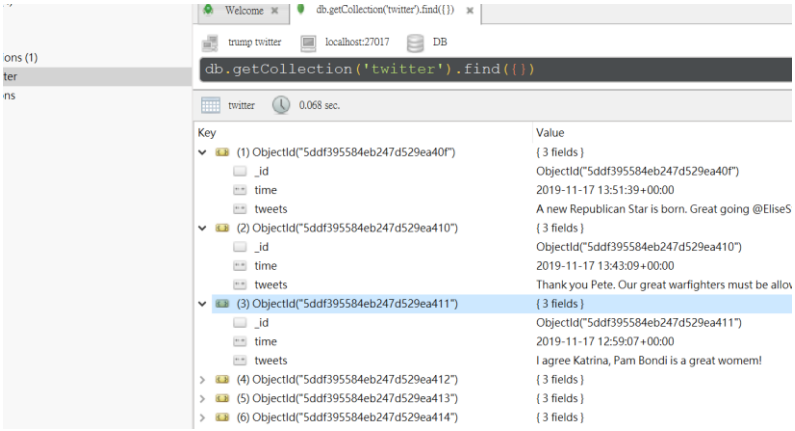
建立一個 MongoDB 資料庫，將蒐集到的大量推文去除圖片和影片後匯入 MongoDB。MongoDB 類似一個資料彙整的中心，組內的大家可以從中抓取資料，並把更新好的資料上傳。如此一來，就不需要把大量的資料存在每個人的電腦上，讓資料的存取與使用更加有效率。

i) 下載並啟動 Mongo



ii) 匯入 csv 檔

iii) 匯入的檔案用 robo3t 呈現就變成這樣啦!



3. 自然語言分析

利用 word2vec 技術分析川普發文中的單詞關聯，幫助之後的分類進行。另外找出漲/跌幅較大的貼文用詞中較常出現的單詞，有助於未來的分析。

4. Label 與資料分類模型

將川普推文的情緒分為正面(1)、負面(2)、中性(0)三種進行標記，並將情緒與 Word2Vec 計算出來的向量「分數」與 S&P 500 指數的漲跌幅合併成一個表作為訓練 KNN 模型的 input，輸出結果分為「大幅正面(1)」或「大幅負面(2)」或「無大幅變動(0)」。之所以不稱為「大幅上漲」或「大幅下跌」是因為推文效果可能使原本積極上漲的股市變為微幅上漲，雖未下跌，但仍有明顯的負面作用。

- 「大幅」的定義為漲跌幅超過資料期間一個標準差，S&P500 未開市的資料以前後最接近開市日之漲跌幅平均代替缺值
- World2Vec 計算文字向量時會扣掉代名詞、介係詞等本身不帶意思表示的單字
- KNN 之訓練及測試採用 train_test_split 方法 隨機將 dataset 中的資料分為 70%訓練、30%測試，避免 panel data 訓練集/測試集出現集中在特定情緒、日期區間或特定文字向量分數的問題

參、 研究結果

本次研究以 2018/3/2 中美貿易戰開始之後的川普推特發文為母體資料，從中選取當日 S&P 500 漲跌幅超過資料區間平均值 1 個標準差的推文做為文 word2vec 訓練集、再以近期推文(2019/12/10~2019/12/16)做為測試集計算向量分數，並標記推文情緒後做為 KNN 分類模型的 input，隨機切割 30%做為測試、70%做為訓練。我們做了兩種模型，皆是以川普當天發文內容預測當日收盤狀況，詳述如下:

一、 預測結果分為有大幅影響(1)及無大幅影響(0)兩類

在 random_stat = 121、k = 85 時整體準確率最佳──整體預測結果準確率為 72%，分類至 1(無大幅影響)的準確率較高，有 73%、分類至 0(無大幅影響)的準確率較差，只有 67%，混淆矩陣如下圖。其中，recall 為召回率，是在所有正樣本當中，能夠預測多少正樣本的比例，precision 為準確率為在所有預測為正樣本中，有多少為正樣本，F1-score 則是兩者的調和平均數。Accuracy 則是整體準確率。

	precision	recall	f1-score	support
0	0.67	0.31	0.43	32
1	0.73	0.92	0.82	66
accuracy			0.72	98
macro avg	0.70	0.62	0.62	98
weighted avg	0.71	0.72	0.69	98

二、 預測結果分為有大幅正面影響(1)、無大幅影響(0)、有大幅負面影響(2)

在 random_stat = 31、k = 94 時整體準確率最佳──整體預測結果準確率為 59%，分類至 2(大幅負面影響)的準確率為 100%準確 分類至 1(大幅正面影響)的準確率最差只有 56%，混淆矩陣如下。

	precision	recall	f1-score	support
0	0.60	0.74	0.66	46
1	0.56	0.69	0.62	32
2	1.00	0.10	0.18	20
accuracy			0.59	98
macro avg	0.72	0.51	0.49	98
weighted avg	0.67	0.59	0.55	98

肆、 研究過程遇到的問題及目前解決方法

Q1: Python 內建套件 Tweepy 有資料期間上的限制，最早僅能追溯回 2019 年 7 月，會造成資料分析上樣本數不足。

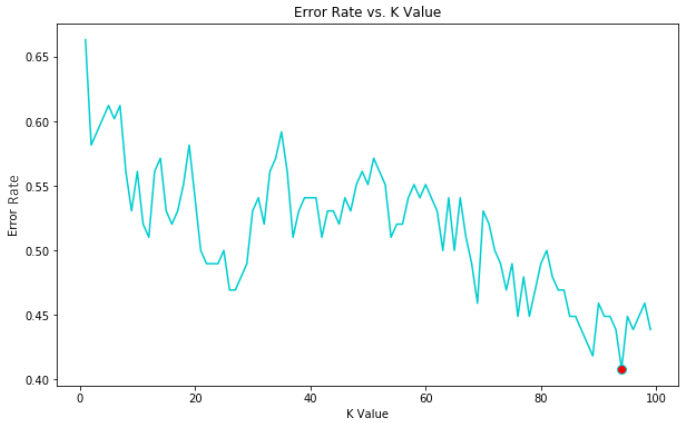
A1: 改採 GetOldTweets3 套件，優點在於無須 Twitter API 的 Keys and Tokens，更無資料期間限制，可自行選取想要的期間,惟須注意 GetOldTweets3 僅支援 Python 3。

Q2: 自然語言分析上資料量太少，分析出的模型太鬆散。

A2: 等川普有更多發文或可以考慮加入其他人的發文一起分析，增加預測準確度。

Q3: 機器學習模型常有各種複雜的參數，本次使用 KNN 模型遇到的問題即為不知該如何設定 K 值、使用 train_test_split 不知該如何設定 random_stat 值。

A3: 最後決定自己訂立特定整數區間一一嘗試，找出錯誤率最低的參數。例如下圖為 random_stat = 31 時，不同 k 值下的 error rate 走勢，因 k=94(紅色點)時 error rate 最低，故選擇該值。



伍、 未來延伸

有了初步的模型，再來就需要大量的數據來驗證，一旦模型達到足夠的準確性，這個計畫就不僅限於預測川普的推文，更可以是其他公開社群網路的任何資訊，亦可將情緒分析的部分也改以機器學習的方式標記，避免人工標記的失誤。我們的構想，就是從這些日常生活中觸手可及的數據中，做出最有價值的運用；只要是文字資料，都能經由統整與歸納，被賦予對使用者來說有使用價值的意義。因此，只要有足夠的資料，使用者可以根據自己的需求，運用相同的概念製作模型，也能達到歸納甚至預測未來的目的。在這個高度資訊化的社會，我們除了是數據的提供者，也可以是數據的使用者，掌握數據的同時，也掌握了未來!