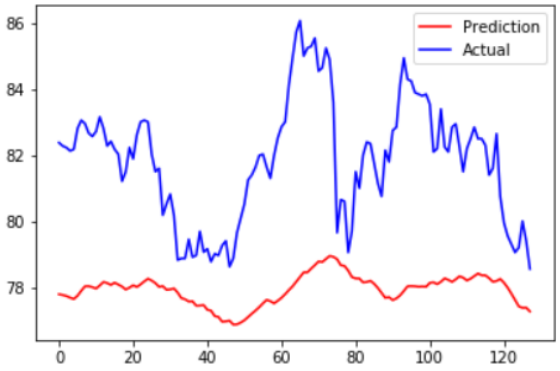
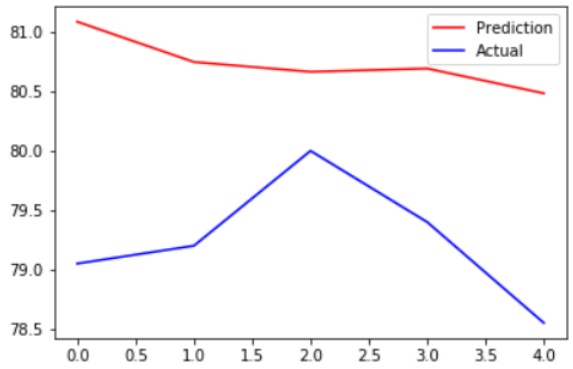
帳號:Q356200
密碼:nckueebigdata
課程網站連結:https://kid.ee.ncku.edu.tw/~course/107-big_data/homepage.html

ssh team5@140.116.158.49
jupyter notebook --ip 140.116.158.49
yarn application -kill application_1521777128978_0332 xdd

5/5(六)下午13:00

- 資料處理 :
  - 孟軒趨勢資料增加:把漲跌1,0, -1 多一欄updown(漲跌分類)
  - 帥哥整理的alldata用學長的資料前處理方式(一)
- 丟進分類漲跌model的資料
  - 趨勢給的資料 : open、high、low、close、volume、updown
- 分類classification model(漲跌)
  - 想法
    - input data為 多種變數(我們的很多feature x, y, z...) 對到 三個類別(1, 0, -1)
    - 老師和大多教學為　一種變數(x) 對到 兩種類別(貓或狗)
  - To-do
    - input處理(多對多分類)
      - https://zhuanlan.zhihu.com/p/34712246
    -
    - nn建立(老師的教學)
- 估值model
  - 預測一百二十幾天
  - 

  - 預測五天



5/1 19:00討論
每個人選參數，自己丟進去model看看（大家試試看能不能讀懂參數><以後也會需要修改）
孟軒、遇人不淑、帥哥：（三）爬蟲 ＊1
小精靈、儷瑛宅：csv ＊3(完成~)
學長想漲跌怎摸整理資料
5/5(六)下午約出來討論，大家這週要衝出來交出去給趨勢！

4/26進度:
csv爬蟲:基本上只要取得下載CSV的URL，知道日期格式就能用寫好的MATLAB程式抓資料
資料整合:用skiprows可讀進dataframe

4/24晚上進度討論：

- 爬蟲都搞定，謝謝精靈謝謝帥哥
- 等學長4/25中午前給我們input
    - 台灣證交所是抓前五年？
    - 彭博是只要抓下一個禮拜的資料 or 前五年？
    - 資料整合
- 開始測xgboost（找出input的重要性）
- 找分類、估值的model
    - 分類：BP
    - 估值（回歸）：
        - CNN
        - SVR
          http://blog.fukuball.com/lin-xuan-tian-jiao-shou-ji-qi-xue-xi-ji-fa-machine-learning-techniques-di-7-jiang-xue-xi-bi-ji/
        - **LSTM**
          http://bangqu.com/79qx8H.html
          https://brohrer.mcknote.com/zh-Hant/how_machine_learning_works/how_rnns_lstm_work.html

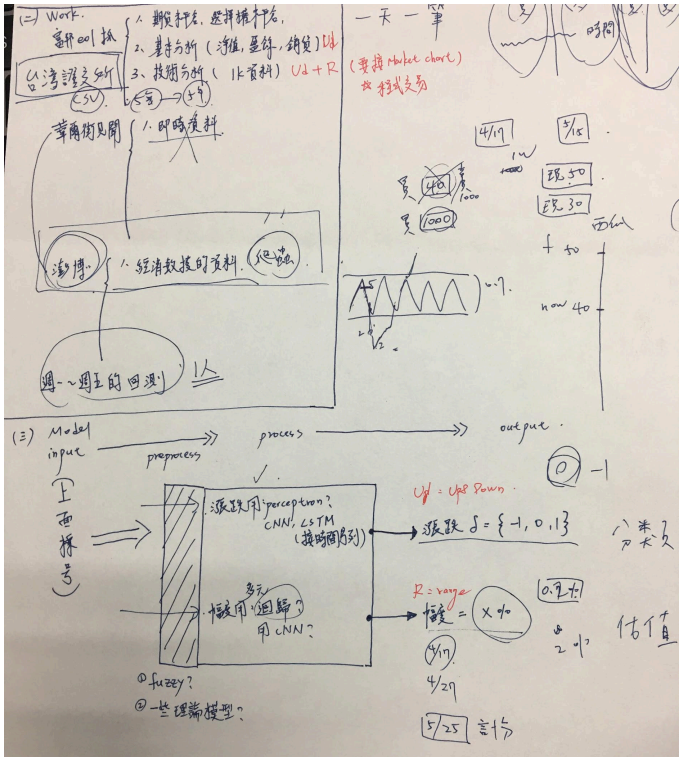
        - ？？？寄信問老師><
        - 參考資料：

(New York Stock Exchange)
https://www.kaggle.com/dgawlik/nyse
https://www.kaggle.com/pablocastilla/predict-stock-prices-with-lstm

2018/4/18 禮拜三16:00 @奇美樓95914 9f
向助教請教的問題

- 題目:趨勢比賽:台股預測
- 如何爬蟲？一天一筆資料，要爬5年，這樣要怎摸處理><
- 適合model?分類model比較了解，估值model助教還有什麼建議嗎><
- 助教您要不要也加入比賽><非常感謝您勹指點

2018/4/17 20:00@電機系館
和資管所交易員學長討論

- 金融的know-how
  - ETF選50支(市值前五十)的驚喜包
    - 比賽共有五種類股:
      - 權值股
      - 電子股
      - 金融股
      - 摩台
      - 高股息
  - input修正 ex. 波動率、選擇權
  - 操盤的流程: input->model->交易策略
  - 

- Input:
  - 機密input
  - 週一～週五回測(趨勢有給)

    - 各主力
    - 經濟數據的資料
- Model
    - 漲跌(分類):
        - CNN, BP, GA(數字轉圖片)
        - LSTM
        - Classification, clustering
    - 幅度(估值):
        - Linear regression
        - LSTM
- 時程:
    - 4/18-22 資料前處理(Data都爬下來)
    - 4/23-27試試model
    - 4/27比賽開始測試
    - 5/25正式計分
- 禮拜五Final Proposal slide:
    - Why:
        - 股票是否能被預測分為兩派, 而我們支持能被預測
        - 交易員的流程-->自動化
    - How: 我們怎摸跟別人做出不同
        - 資管所學長
        - 仁暐大弟子(放老師照片xdd)
        - 最強姿穎
    - What:
        - Input
        - Model

2018/4/16 21:30@電機系館

- 確認題目
    - 趨勢比賽:台股預測
- 隊名:
    - 讓你賺到飽
- 時程:
    - 4/27測試
    - 5/25正式開始
- Input:(會再去問資管所學長金融知識)
  影響台股的因素 (有沒有opendata)
    - 趨勢本身的台股資料
    - 美股**ETF**(家年)
    - 黃金(昱仁)
      每月黃金價格自從1950/2/1-2018/2/1
    - 房價(儷瑛)
      股市和房市走勢息息相關嗎
      房價dataset: https://data.nat.gov.tw/dataset/6213
    - 匯率
    - 陸股ETF(姿穎)

- 歐洲ETF database(孟軒)
- 新聞(重大事件)
- 全球金融趨勢
- 文字分析？(ptt Gossiping+Stock 推爆的新聞分析)
- 71個免費數據集
- 有個說臉書分享他就給數據`xdd

- Model(禮拜三中午前每一個人研究一個**model**打到共筆，才能問助教><)
  - CNN(🌰媛)
  - SVM(昱仁)
  - ARIMA模型(方)
    - **A**uto**r**egressive **I**ntegrated **M**oving **A**verage model
    - 時間序列預測分析方法之一
    - 基本思想：將預測對象隨時間推移而形成的數據序列視為一個隨機序列，用一定的數學模型來近似描述這個序列。這個模型一旦被識別後就可以從時間序列的過去值及現在值來預測未來值。現代統計方法、計量經濟模型在某種程度上已經能夠幫助企業對未來進行預測。
    - 參考資料(含模型預測流程)
    - 時間序列模型基本概念：AR, MA, ARMA, ARIMA
  - ARCH模型(方)
    - **A**uto**r**egressive **c**onditional **h**eteroskedasticity model
    - 模型的應用:準確地模擬時間序列變數的波動性的變化，它在金融工程學的實證研究中應用廣泛，使人們能更加準確地把握風險(波動性)
    - 參考資料
    - 如果變異數用ARMA模型來表示，則ARCH模型的變形為GARCH模型
    - GARCH模型:Generalized ARCH
  - LSTM (家年)
    - 時間序列
  - SVR

    套件:libsvm

    台大林智仁教授團隊所做，可調整參數，用來幫助分類回歸等

    林教授有一套建議:

    1. 將資料轉成libsvm看得懂的格式
    2. training data和testing data做**Scale** data
    3. 選用效能較佳的RBF kernel(預設值即是RBF)
    4. 使用cross validation選擇較好的參數(即是grid.py所做的事)
    5. 套用剛所找到的參數來**Train** model
    6. 將testing data做**Predict** data

    libsvm原始版(林智仁教授提供)

    笨蛋也可以用的 libsvm(中文版快速說明)
  - BP類神經網絡模型(方)
    - Back Propagation Neural Network 倒傳遞類神經網路
    - 監督式學習

- 學習階段 - 找到合適的權重讓輸入特徵可以計算出目標結果
  回想階段 - 利用學習完的特徵計算結果
- 依照最小誤差平方來進行權重的訓練
- 訓練的流程如下:(參考資料)
  1.初始化權重 (random)
  2.利用目前的權重計算輸出結果
  3.計算輸出結果和目標結果的誤差
  4.調整權重
  5.重複2~5直到收斂
- 優勢:很强的非線性映射能力和柔性的網絡结構
  劣勢:學習速度慢，易陷入局部極小值(參考資料)
- 與線性模式(AR、GARCH等)相比，非線性倒傳遞類神經網路模式預則能力較好。基因演算法配合類神經網路對於短期股價預測有很好的預測效果。(參考資料)
- SVM、PCA-SVM、BP、GA-BP四種方法比較MSE後，最佳:GA-BP(參考資料)(BP、GA-BP)

- XGBoost(feature importance)(子瑛)
  - 用於解決監督式問題(適用於變數較少的資料，更容易於調整參數)
  - 內容: XGBoost 所對應的模型為 Tree Ensemble，主要是由分類和迴歸樹 (Classification And Regression Trees，CART) 所組成，所謂 CART 會將輸入根據不同的屬性分配至各個葉子節點，每個葉子節點皆會對應一個分數(參考資料)
  - eXtreme Gradient Boosting(xgboost)
  - 優化版的GBDT(Gradient Boosting Decision Tree梯度提升決策樹)
  - 背後原理推導(數學式推導)
  - xgboost細節實作(含參數設定)
- paper蒐集找一個上面沒提過的model(孟軒)
  - Clustering, Classification
    Multifaceted Predictive Algorithms in Commodity Markets
  - Random forest
  - Reinforcement learning 馬可夫
- Case study
  - JP Morgan 2017 Value Strategies based on Machine Learning
    - Input:

ROE / Dupont Decomposition / Profitability / Efficiency		Valuation / Payout	
	Details		**Details**
ROE	Return on Equity	Earnings yield	Trailing EPS / Price
Net Margin	Net Profit Margin	Forward earnings yield	1 / (12-month forward PE)
Asset turnover	Asset Turnover	Dividend yield	Trailing DPS/ Price
Gearing	Total Assets / Equity	Forward dividend yield	12-month forward dividend yield
ROA	Return on Assets	Sales yield	Trailing Sales / Price
ROC	EBIT / Capital employed	Forward sales yield	12-month forward sales yield
ROE FY1	Return on Equity FY1	Shareholder yield	(Total Dividends + net Repurchases + net changes in Debt) / Market Cap
Gross-Profit-to-Assets	Gross profit / Total Assets	Cash flow yield	Trailing cash flow per share / price
Cash-Flow-to-Assets	Free Cash Flow / Total Assets	Sales to EV	Sales or Operating Revenue / Enterprise Value
CFROI	Cash Flow Return on Investment	EBITDA to EV	EBITDA to Enterprise Value
Positive Earnings	Dummy variable: 1 = positive last reported EPS, 0 otherwise	Payout ratio	Dividend per share / Earnings per share
		Dividend payer	Dummy variable: 1 = positive last reported DPS, 0 otherwise

Growth / Sentiment		Quality / Risk	
	Details		**Details**
Forecast EPS growth	Forecast Earnings growth from FY1 to FY2	Earnings Certainty	- [ABS(EPS FY1 Std Dev / EPS FY1) + ABS(EPS FY2 Std Dev / EPS FY2)]
Forecast DPS growth	Forecast 1-year growth in Dividends	ROE volatility	5 year standard deviation in ROE
EPS long-term growth	EPS long-term growth	Accruals	Change in Net Operating Assets / Average Assets
Recommendation change 1M	- (1-month change in Analyst recommendation) (1=Strong Buy, 5=Strong Sell)	Realized Volatility 90D	Historical volatility (past 90 days)
Earnings momentum	Average of 1M and 3M changes in FY1 and FY2 earnings estimates	Beta	MSCI Country Beta
Momentum		Market Cap	Investible Market Cap
	Details	ADV 1M	Average daily value traded (past 20 days)
Momentum 1M	1M price momentum		
Momentum 12M-1M	12M minus 1M price momentum		

Source: J.P. Morgan Quantitative and Derivatives Strategy

- Model:
    - 惩罚回归(LASSO)，梯度提升（XGBoost)和线性回归3个模型组成
    - 利用新闻情绪数据构造策略信号，对股票进行筛选
- Error設定：
  毛利润(Gross Profit)而不是净资产收益率（ROE）作为盈利能力衡量标准对股票进行筛选

- 原文沒公開但找到了節錄
- 幫大家找到了中文xdd

■ 國信證券
■ 長江證券
- 常見model在實務上應用

表 1：情景问题、具体的金融实例及其对应的机器学习方法

问题	金融实例	机器学习方法
给定输入变量，预测资产价格的方向	使用技术指标对于对应的指数进行择时	SVM、Logistic 回归、Lasso 回归等
一种资产的剧烈变动如何影响其他资产	美元指数的变动对于美国国债收益率及黄金走势的影响	格兰杰因果检验、脉冲响应函数
一种资产走势是否偏离其他相关资产	黑色系商品走势的分化	一对多分类
找出资产价格的驱动因素	行业中有效因子的筛选	PCA、ICA
目前市场状态判断	对于利率上行或下行周期判断	隐马尔科夫、Soft-max分类
一个事件发生的概率/是否会发生	高送转事件的预测	决策树、随机森林、Logistic回归
在噪音数据中寻找信号	资产周期的分析	低通滤波器、SVM
一篇文章或一段文字的感情色彩、主题	公司公告的舆情分析	词袋分析、词频-逆向文件频率（TF-IDF）
有哪些常见的市场压力指标		K-means聚类分析
计算图像中某物体数量		卷积神经网络
最优执行速度		基于部分可观察马尔科夫过程的强化学习
基于大量输入数据预测波动率		受限玻尔兹曼机、SVM

资料来源：JP Morgan Macro QDS，长江证券研究所

    - QDA, GDA, SVM長期預測
      Stanford SVM forecast
    - bayesian regression
      MIT Bayesian

- Output:
    - 台灣十八檔上市櫃成分證券ETF在下一週五天的漲跌及價格
- Reference:
  類神經網路應用於股票投資策略之研究
-
    -

2018/4/15絕對不要再看日出之提早開始
之前暫定題目：美西旅遊路徑推薦(Foursquare 打卡資料)

- Input:

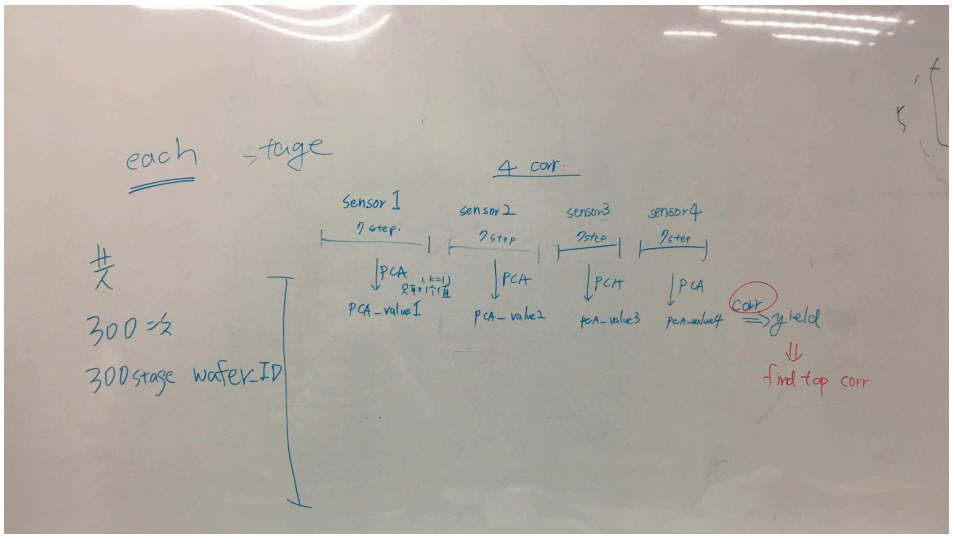
    - **NYC Restaurant Rich Dataset (Check-ins, Tips, Tags)**
        - 打卡（user ID and venue ID）
        - 餐廳（user ID and venue ID）
        - 標注（user ID and tag set）
    - **NYC and Tokyo Check-in Dataset**
      1. User ID (anonymized)
      2. Venue ID (Foursquare)
      3. Venue category ID (Foursquare)
      4. Venue category name (Fousquare)
      5. Latitude
      6. Longitude
      7. Timezone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC)
      8. UTC time
    - **Global-scale Check-in Dataset**
      1. City name
      2. Latitude (of City center)
      3. Longitude (of City center)
      4. Country code (ISO 3166-1 alpha-2 two-letter country codes)
      5. Country name
      6. City type
    - **User Profile Dataset**
      1. User ID
      2. Gender
      3. Twitter friend count
      4. Twitter follower count
- Output:
    - 旅遊路徑?
- Model:
    - Mining Frequent Patterns, Associations, and Correlations
    - Classification, clustering?
    - CNN?
    - 推薦系統（content-based filtering, collaborative based filtering）

更多新比賽:

- 台股預測 5/25（有dataset, 題目很明確, 這個好像也可以試試耶）
    - 股價影響情緒還是情緒影響股價?
- 政府氣象、交通、經濟 4/19（政府開放dataset, 題目還要自己想, 搞不好我們旅遊的可以延伸）
- 跟AI有關都可以 5/20（沒有dataset, 可以把我們自己題目包裝一下丟上去）
- 智慧工廠及智慧自動化領域相關(如工業4.0 物聯網,雲端計算,大數據) 含半導體光電設備、醫療設備、自動光學檢測設備、智慧化工具機、智動化資訊系統、數位化模具、智慧化生物機電系統進行專題研究 5/4（好像比較系統性的專題研究）

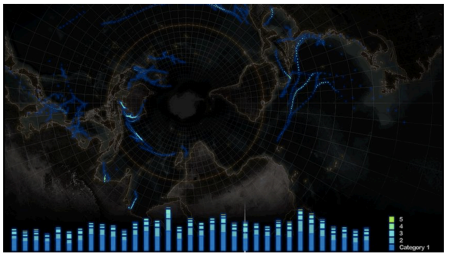
2018/04/07線上討論作業紀錄

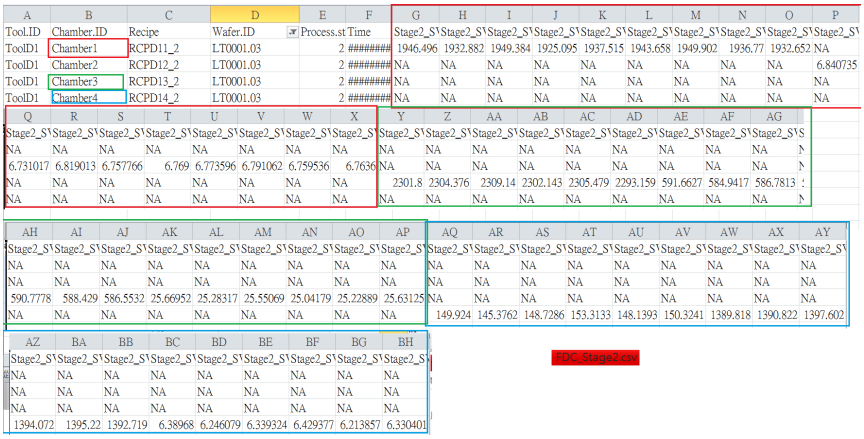
- 以一個stage來分, 同一個wafer同一個sensor算平均值(Xi), i由sensor數量決定, Y是yield值。

    - =>要合併很多份檔案冂(整合wafer資料)
- 以一個recipe來分，同一個wafer同一個sensor算平均值(Xi)，i由sensor數量決定，Y是該wafer的yield值。
    - =>什麼意思 同一個檔案 以recipe來group的話 每個recipe只會有一片wafer id(就是每一筆分開討論的意思?)
    - Recipe = RCP D1(tool) 1(chamber)_2(process stage)
    - 那NA呢 填0/捨棄/填平均?
- 每個檔案 SVID(sensor)數 不同 => 有檔案只有x1 x2 有的有x8? => 全部取統一的x數?
    - 共線性來取要用什麼feature?
- 有要考慮時間嗎(同時兩個chamber 兩片wafer做同樣的事) => 無法驗證
- 要不要用wafer log檔案(經過什麼stage之類的)=>決定不用力 跑太久力
- 資料前處理(3姿穎、儷媛、濰宥、昱仁)、PCA(1子瑛)、資料視覺化(2家年、孟軒)、報告(大家)
- 資料前處理：將檔案輸出成Recipe,Wafer_id,每個sensor的step平均
- PCA: 找前三名和yield最有關係的stage
    - 先假設300 stage 每個裡面有4個sensor 的平均值
    - 先把4個sensor 的平均值跑 pca降維，找出相關性高的並在一起（像是咳嗽和感冒正相關，就可以只做咳嗽）
    - 或4 sensor * 7 step ＝28跑一次pca 再當值和stage找yield相關係數
    - (4/10)將每個sensor的所有step做PCA(ex:stage1有4個sensors，則會出現4個值)，最後找哪個sensor和yield間最大的相關係數作為代表
    - 

    1. array
    2. pca
    3. vector 轉list
    4. correlation
    5. for loop
    6. 視覺化 scatter回歸線
    7. SVM 預測（不用作）
    8. CNN（不用作）

- Data Visulization

- Input:
  - 和yield關係最大的Stage裡面的Sensor
  - Correlation的值, sensor, stage, yield
- 語法：https://plot.ly/python/apache-spark/（但都是基本圓餅、曲線圖，可以看其他人做圖，看能不能自己湊出有yield意義的相關圖形）
- 靈感來源：
  - https://mashable.com/2013/03/05/data-visualization-projects/#U.8jD6.0dgqO
  - https://www.awwwards.com/websites/data-visualization/
- 討論：
  - 如果把step plot成像地圖一樣？用面積代表sensor數、顏色深淺yield代表高低？！
  - 相關的stage可以在一片大陸那種？！
  - 或者可以像這樣再來出一個長駐圖 想有什麼變量可能需要？！
  - 



2018/03/30上課筆記

- spark從client改成cluster比較快 (在設定環境那裡)
- 作業data會有空 本來rawdata就是如此
- Data Visulization 把結果(多變量)怎麼plot 或呈現，三個含量以上
- Data以 process stage為主

參考資源:(1)https://www.datacamp.com/community/tutorials/scikit-learn-python
(PCA和視覺化等tutorial)
(2)https://blog.csdn.net/u012162613/article/details/42192293
(PCA實作方法)
sklearn 沒辦法匯入ㄚㄚㄚ
==========

(3)https://stackoverflow.com/questions/31774311/pca-analysis-in-pyspark

2018/03/22連線討論紀錄(上課前投票)

1. 阿里媽媽搜索廣告預測(已有阿里巴巴資料)-->需要中國電話才可以註冊資料???
   服飾標籤(已有阿里巴巴資料)+

2. 新聞期貨相關的關鍵字，預測期貨的價錢(新聞網站的opendata)
   - New York Timeshttps://developer.nytimes.com
   - CNNhttps://developer.cnn.com
   - BBChttp://www.bbc.co.uk/developer/technology/apis.html
   - The Guardian Open Platformhttp://open-platform.theguardian.com/documentation/
   - NPRhttps://www.npr.org/api/index

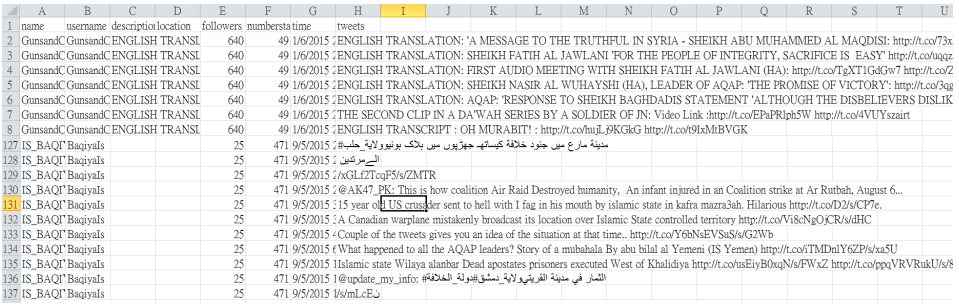
3. Twitter文字內容，預測會被轉推幾次(按讚數，文字內容類型)或使用者性格、政治態度、流行新元素(資料無法下載)
4. 政府公開datasets公開資料:天氣、電廠用量預測空汙 ++
5. 影像處理預測疾病
6. facebook 預測(有data可用)++++
   a. https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset
   b. https://archive.ics.uci.edu/ml/datasets/Facebook+metrics
   c. 臉書twitter分析方法(科技部報告)
   d. 預測打卡位置 （沒user id 經度跟時間不知道是啥 ）
      https://www.kaggle.com/c/facebook-v-predicting-check-ins

row_id	x	y	accuracy	time	place_id
0	0.7941	9.0809	54	470702	8523065625
1	5.9567	4.7968	13	186555	1757726713
2	8.3078	7.0407	74	322648	1137537235
3	7.3665	2.5165	65	704587	6567393236
4	4.0961	1.1307	31	472130	7440663949
5	3.8099	1.9586	75	178065	6289802927
6	6.3336	4.372	13	666829	9931249544
7	5.7409	6.7697	85	369002	5662813655
8	4.3114	6.941	3	166384	8471780938
9	6.3414	0.0758	65	400060	1253803156
10	2.0173	4.8627	6	21353	8684462954
11	8.7101	2.9442	73	153493	2159916487
12	0.8829	1.3445	64	574488	7652380351
13	2.4336	8.06	62	238054	8234363596
14	6.155	1.9774	8	325411	2272949794
15	7.6219	9.6208	65	321519	4740742194
16	3.2494	3.2096	75	777982	2123587484
17	0.7084	8.9051	69	320633	8016758016

   e. 推薦類似路徑的人下一個地點，推薦類似路徑的人朋友
   f. 使用者的資料
   g. 上網看影片:classification, clustering, association rule找出重要的球員

h. https://sites.google.com/site/yangdingqi/home/foursquare-dataset(打卡資料-->可用於旅遊地點推薦)
i. ask老師 資訊內容 dataset長怎樣

7. Social Cluster analysis(Analysing "How ISIS Uses Twitter" using social network cluster analysis)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	name	username	descriptio	location	followers	numbersta	time	tweets													
2	GunsandC	GunsandC	ENGLISH TRANSL		640	49	1/6/2015 2	ENGLISH TRANSLATION: 'A MESSAGE TO THE TRUTHFUL IN SYRIA - SHEIKH ABU MUHAMMED AL MAQDISI: http://t.co/73x														
3	GunsandC	GunsandC	ENGLISH TRANSL		640	49	1/6/2015 2	ENGLISH TRANSLATION: SHEIKH FATIH AL JAWLANI 'FOR THE PEOPLE OF INTEGRITY, SACRIFICE IS EASY' http://t.co/uqqz														
4	GunsandC	GunsandC	ENGLISH TRANSL		640	49	1/6/2015 2	ENGLISH TRANSLATION: FIRST AUDIO MEETING WITH SHEIKH FATIH AL JAWLANI (HA): http://t.co/TgXT1GdGw7 http://t.co/2														
5	GunsandC	GunsandC	ENGLISH TRANSL		640	49	1/6/2015 2	ENGLISH TRANSLATION: SHEIKH NASIR AL WUHAYSHI (HA), LEADER OF AQAP: 'THE PROMISE OF VICTORY': http://t.co/3qg														
6	GunsandC	GunsandC	ENGLISH TRANSL		640	49	1/6/2015 2	ENGLISH TRANSLATION: AQAP: 'RESPONSE TO SHEIKH BAGHDADIS STATEMENT 'ALTHOUGH THE DISBELIEVERS DISLIK														
7	GunsandC	GunsandC	ENGLISH TRANSL		640	49	1/6/2015 2	THE SECOND CLIP IN A DA'WAH SERIES BY A SOLDIER OF JN: Video Link :http://t.co/EPaPRlph5W http://t.co/4VUYszairt														
8	GunsandC	GunsandC	ENGLISH TRANSL		640	49	1/6/2015 2	ENGLISH TRANSCRIPT : OH MURABIT! : http://t.co/hujLj9KGkG http://t.co/t9IxMtBVGK														
127	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 2	#مدينة سارع بين جنود خلافة كيستهم جهڒيون بين بلاك بونيوولاية_حلب#														
128	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 2	السرمدين														
129	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 2	/xGLf2TcqF5/s/ZMTR														
130	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 2	@AK47_PK: This is how coalition Air Raid Destroyed humanity, An infant injured in an Coalition strike at Ar Rutbah, August 6...														
131	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 2	15 year old US crusader sent to hell with I fag in his mouth by islamic state in kafra mazra3ah. Hilarious http://t.co/D2/s/CP7e.														
132	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 2	A Canadian warplane mistakenly broadcast its location over Islamic State controlled territory http://t.co/Vi8cNgOjCR/s/dHC														
133	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 2	Couple of the tweets gives you an idea of the situation at that time.. http://t.co/Y6bNsEVSuS/s/G2Wb														
134	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 6	What happened to all the AQAP leaders? Story of a mubahala By abu bilal al Yemeni (IS Yemen) http://t.co/iTMDnlY6ZP/s/xa5U														
135	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 1	Islamic state Wilaya alanbar Dead apostates prisoners executed West of Khalidiya http://t.co/usEiyB0xqN/s/FWxZ http://t.co/ppqVRVRukU/s/8														
136	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 1	@update_my_infos: #الخلافة_دمشق#الدولة_الشام في مدينة القريتين_ولاية														
137	IS_BAQI"	BaqiyaIs			25	471	9/5/2015 1	/s/mLcEن														

8. 股票投資組合
https://archive.ics.uci.edu/ml/datasets/Stock+portfolio+performance
9. 股票數據
https://www.kaggle.com/minatverma/nse-stocks-data + ++
10. 台積電良率分析 +
11. Dcard爬蟲+
https://github.com/leVirve/dcard-spider
12. 暫定:美西旅遊路徑推薦(Foursquare 打卡資料)
user id, 經緯度, 地點id(類型), 時間,
https://sites.google.com/site/yangdingqi/home/foursquare-dataset
~有台灣ㄇ?
13. instagram hashtag-->業配文
https://www.instagram.com/developer/authorization/
~applications no longer accepted是什麼意思?不給爬ㄌ嗎

2018/3/23 NBA 冠軍預測
資料https://www.kaggle.com/c/mens-machine-learning-competition-2018/data
由於NBA每個季度都會有球員交易 往往先發球員組合每年都會換, 每一年都是一個全新的開始
1球隊的球員組合?每支隊伍比賽大多固定先發 X

2 兩隊每節分數比較 評估哪支隊伍較有勝率： 整場比賽節奏快慢 體力消耗 個人分數占團體分數過高

3 網路輿論

應該朝比賽時不確定因素發想？球員耐壓性 主客優勢(觀眾的音量)

NBA例行賽：82場比賽 4月結束, 每個分區的前八名 進入季後賽

目前有30支球隊

季後賽 [編輯]

西區季後賽			NBA總冠軍		東區季後賽		
第一輪	準決賽	組決賽	總決賽		組決賽	準決賽	第一輪
[1]金州勇士 4 [8]波特蘭拓荒者 0	[1]金州勇士 4 [5]猶他爵士 0	[1]金州勇士 4 [2]聖安東尼奧馬刺 0	金州勇士 4	克里夫蘭騎士 1	[1]波士頓塞爾蒂克 1 [2]克里夫蘭騎士 4	[1]波士頓塞爾蒂克 4 [4]華盛頓巫師 3	[1]波士頓塞爾蒂克 4 [8]芝加哥公牛 2
[4]洛杉磯快艇 3 [5]猶他爵士 4							[4]華盛頓巫師 4 [5]亞特蘭大老鷹 2
[3]休士頓火箭 4 [6]奧克拉荷馬雷霆 1	[3]休士頓火箭 2 [2]聖安東尼奧馬刺 4					[3]多倫多暴龍 0 [2]克里夫蘭騎士 4	[3]多倫多暴龍 4 [6]密爾瓦基公鹿 2
[2]聖安東尼奧馬刺 4 [7]曼斐斯灰熊 2							[2]克里夫蘭騎士 4 [7]印第安那溜馬 0

• 粗體字為勝隊。

Rank	NAME	TEAM	OFFENSE	DEFENSE	TOTAL	Wins
1	Chris Paul	LAC	6.26	3.25	9.51	5.61
2	Kevin Durant	GS	5.5	1.89	7.39	4.85
3	Jimmy Butler	CHI	5.49	1.42	6.91	4.4
4	James Harden	HOU	7.66	-1.12	6.54	4.59
5	Russell Westbrook	OKC	6.76	-0.8	5.96	4.47
6	Stephen Curry	GS	6.72	-0.99	5.73	3.99
7	Kawhi Leonard	SA	5.14	0.51	5.65	3.67
8	DeMarcus Cousins	SAC	5.01	0.44	5.45	3.68
9	Giannis Antetokounmpo	MIL	2.48	2.85	5.33	3.37
10	Kyle Lowry	TOR	5.15	-0.01	5.14	3.87
11	Draymond Green	GS	1.23	3.79	5.02	3.43
12	Blake Griffin	LAC	2.6	2.37	4.97	3.66
13	Anthony Davis	NO	2.35	2.3	4.65	3.77
14	LeBron James	CLE	3.85	0.33	4.18	2.93
15	Paul Millsap	ATL	0.12	3.83	3.95	2.64
16	Trevor Ariza	HOU	2.96	0.93	3.89	3.03
17	Kevin Love	CLE	2.68	1.11	3.79	2.59
18	Kemba Walker	CHA	4.61	-0.83	3.78	2.86
19	Otto Porter Jr.	WSH	2.71	1.04	3.75	2.63
20	Mike Conley	MEM	3.7	0.01	3.71	2.24

2018/03/15 ~jwhuang

有比較明確的題目方向了嗎?

~組長:預計分組討論, 下禮拜(3/22)組內提案

- 社群媒體分析
- 籃球預測

2018/03/02 討論

To-do:
1.社群媒體、股票、空汙
2.大數據比賽相關題目參考
3.我們的優勢?(改善別人已經做過的結果, 或是加入其他 idea)
4.datasets(從 dataset 裡面找到想要分析的 feature 或想要研究的方向)

想法:股票結合社群?

加入 Twitter 情緒來預測股票趨勢
相關 paper:*method? dataset?*

我認為無法實做, 手邊沒有data, 也無法得知他們演算法的細節
效果只在特定的情緒表現佳, 有的還很差
簡單來說, 許多細節不清楚, 容易產生很多疑問

Twitter mood predicts the stock market
https://arxiv.org/pdf/1010.3003.pdf
dataset: http://terramood.informatics.indiana.edu/data 好像不見力 的確

IMPROVING PREDICTION OF STOCK MARKET INDICES BY ANALYZING THE PSYCHOLOGICAL STATES OF TWITTER USERS
https://goo.gl/UmECRb

助教建議:需先定義那些情緒會影響股價?是股價影響情緒還是情緒影響股價這點也很重要。
以情緒影響股價 paper中的做法有定義多種情緒，但最後結果發現只有一至兩種可以較為準確預測道瓊指數變化

空汙參考資料:https://qqaq.ee.ncku.edu.tw
汙染源境外／境內汙染比例？
.....(更具體想要解決的問題)
助教建議:須先了解哪些因素會造成境內汙染及境外汙染是如何傳遞到台灣的。(地形、風向、風力、地理位置等等)
別的社群媒體的資料應用
.....(？)

Dataset
(maybe) from Kaggle:https://www.kaggle.com/competitions

比賽題目參考(都有**dataset**了):

- 趨勢科技 T-Brain AI 實戰吧 (台灣)
  垃圾郵件、惡意程式
- 2018"雲移杯(大陸)
  景区突发事件互联网传播分析及可視化
- Google Cloud & NCAA® ML Competition 2018(美國)
  預測籃球賽冠軍
- 天文數據挖掘大賽(大陸)
  自己找pattern對天體區分
  ex.例如选择特定波长或波段上的光谱流量值等作为特征，并运用算法对各种天体进行区分。
- March電路'18(IBM)
  看不懂比賽要幹嘛？？？
- 國人就醫的共病案例數預測(工研院)
  健保資料
- 設備維護預測(工研院)
- Alexathon (Amazon)
  做語音辨識什麼的？？？

- 2018 Data Science Bowl

  用影像找cell裡面的nuclei，及早找到癌症
- FashionAI全球挑战赛—服饰关键点定位（阿里巴巴）

  阿里巴巴的服飾圖像和使用者購買行為
- Shinkenchiku Residential Design Competition 2018（日本）

  用ai設計房子
- Google Landmark Recognition Challenge (Google)

  訓練從圖片辨識landmark
- ICPR 2018 挑战赛

  文本識別
- The General Video Game AI Competition – Learning Track (Deep Mind)

  做ai遊戲
- World Computer Chess Championships 2018
- Ms. Pac-Man Vs. Ghost Team Competition 2018

  小鬼吃糖果的遊戲
- 阿里巴巴一大堆大賽題目有趣有趣
- 台灣渣打金融科技 校園創意挑戰賽
- 共創急智網
- Microsoft Imagine Cup 2018

  思考問題 找方法解決

  https://imagine.microsoft.com/zh-tw/Country/TW (第一輪截止:3/31)

助教建議:這些比賽都有很明確的要解決之問題及完整的資料庫提供，建議可以先研究前人有甚麼問題沒有解決，然後大家一起討論解決方法。

助教建議:建議可以去看"data mining concepts and techniques"第三版這本書的以下章節(網路上好像有PDF檔)

- Chapter 6 Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods 243
- Chapter 8 Classification: Basic Concepts 327
- Chapter 10 Cluster Analysis: Basic Concepts and Methods 443

可以更了解重要特徵要怎麼找