

爬蟲基礎入門

Guide introduction

領路人: Allen

- 企業AI 團隊領導人 (熟悉所有機械學習相關領域)
- 機械手臂自動化/瑕疵檢測/ 機器人移動/ 物件偵測&分割模型專家
- 雲端資料串流專家
- 演算法論文復獻/驗證能力

個人AI演算法專利2項

嵌入式開發板演算法佈署經驗 (Nvidia Nano & NX / Raspberry Pi / intel NCS2+ OpenVINO)

雲端平台數據整合經驗 (MLOps & DevOps on AZURE / AWS)

自動化股市回測系統設計經驗

深度學習馬拉松陪跑專家

機械手臂搭配AI演算法經驗

What you learned?

1. 資料流程概述

1.1 專案動機

1.2 數據哪裡找?

2. 如何爬取資料

2.1 資料種類以及策略

2.2 爬取工具介紹

3. 程式基礎及網頁架構介紹

3.1 python 環境架設,架構介紹

3.2 代碼實作講解

3.3 數據分析及圖表練習

3.4 網頁架構概述

3.5 自動化專案封裝

4. 實戰練習

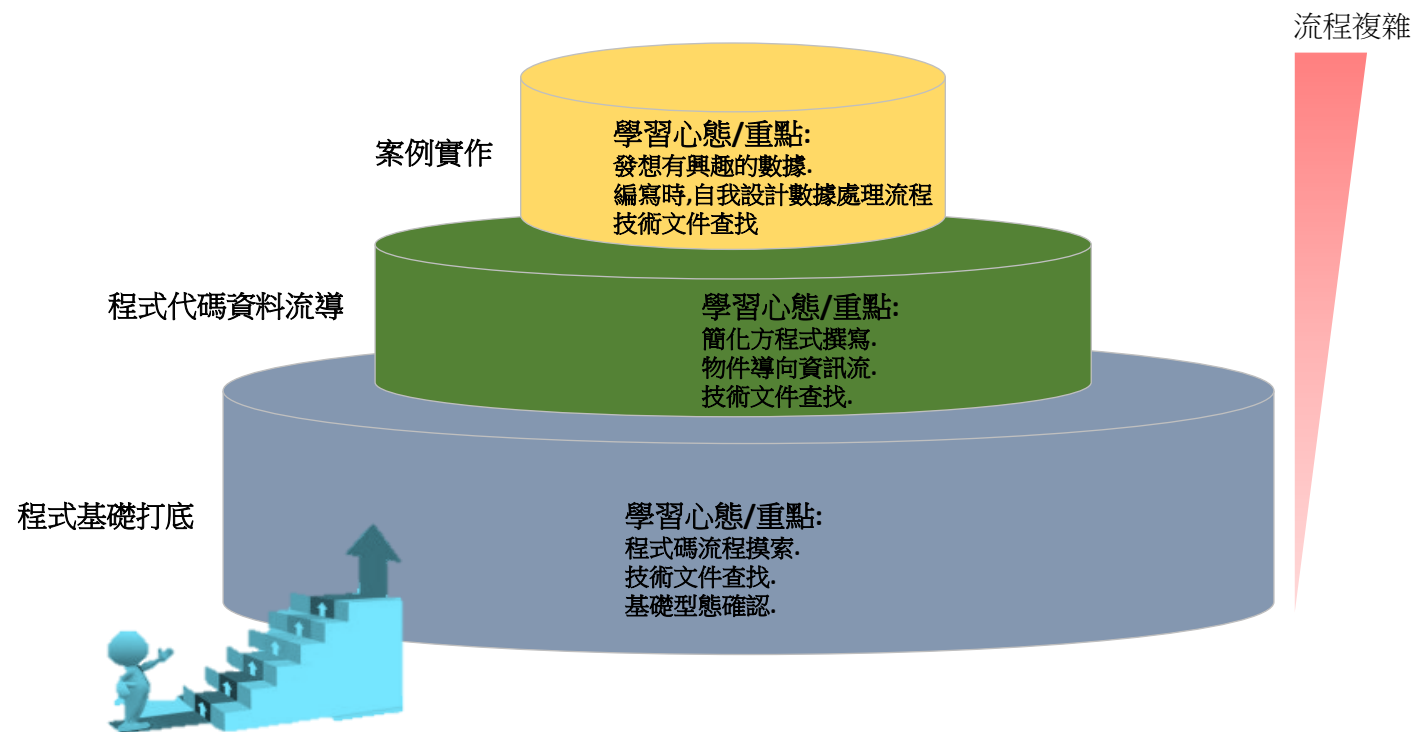
4.1 股市資料爬取

4.2 拍賣網站資料爬取

5. 學員專題實作

5.1 分組作業繳交

What you learned?



多注意老師在查找技術文件的關鍵字.
多舉手提問, 否則後面不容易跟上.
學習結束後, 要不斷的重複coding, 熟悉語法.

What you learned?



	name	price
0	外觀9成新 全機原廠零件 保固6個月【A級福利品】Apple iPhone 12 mini ...	16990
1	琉璃流金+電鍍玻璃手機殼 iPhone 12 Pro / i12 Pro 保護殼 軟邊硬殼 ...	499
2	琉璃流金+電鍍玻璃手機殼 iPhone 12 / i12 保護殼 軟邊硬殼 - 巴黎黑	499
3	琉璃流金+電鍍玻璃手機殼 iPhone 12 Pro / i12 Pro 保護殼 軟邊硬殼 ...	499
4	琉璃流金+電鍍玻璃手機殼 iPhone 11 / i11 保護殼 軟邊硬殼 - 文藝藍	499
5	琉璃流金+電鍍玻璃手機殼 iPhone 12 / i12 保護殼 軟邊硬殼 - 孔雀綠	499
6	琉璃流金+電鍍玻璃手機殼 iPhone 12 / i12 保護殼 軟邊硬殼 - 琉璃粉	499
7	琉璃流金+電鍍玻璃手機殼 iPhone 12 Pro Max / i12 Pro Max 保...	499
8	琉璃流金+電鍍玻璃手機殼 iPhone 12 Pro Max / i12 Pro Max 保...	499
9	琉璃流金+電鍍玻璃手機殼 iPhone 12 / i12 保護殼 軟邊硬殼 - 石英紫	499
10	琉璃流金+電鍍玻璃手機殼 iPhone 12 / i12 保護殼 軟邊硬殼 - 香紗白	499
11	琉璃流金+電鍍玻璃手機殼 iPhone 11 Pro Max / i11 Pro Max 保...	499
12	琉璃流金+電鍍玻璃手機殼 iPhone 12 Pro / i12 Pro 保護殼 軟邊硬殼 ...	499
13	琉璃流金+電鍍玻璃手機殼 iPhone 12 Pro Max / i12 Pro Max 保...	499
14	美國 Case●Mate iPhone 13 Pro Max Tough Clear Plu...	1200
15	▼每日強檔 瘋殺開賣▼珊瑚色★狂降\$3201Apple iPhone XR (128G)-珊瑚色	15299
16	琉璃流金+電鍍玻璃手機殼 iPhone 11 Pro Max / i11 Pro Max 保...	499
17	琉璃流金+電鍍玻璃手機殼 iPhone 11 Pro Max / i11 Pro Max 保...	499
18	琉璃流金+電鍍玻璃手機殼 iPhone 12 / i12 保護殼 軟邊硬殼 - 文藝藍	499
19	琉璃流金+電鍍玻璃手機殼 iPhone 12 Pro / i12 Pro 保護殼 軟邊硬殼 ...	499

What you learned?



stock_crawl.exe

2330_from2020-06_to_2021-08.csv - 記事本

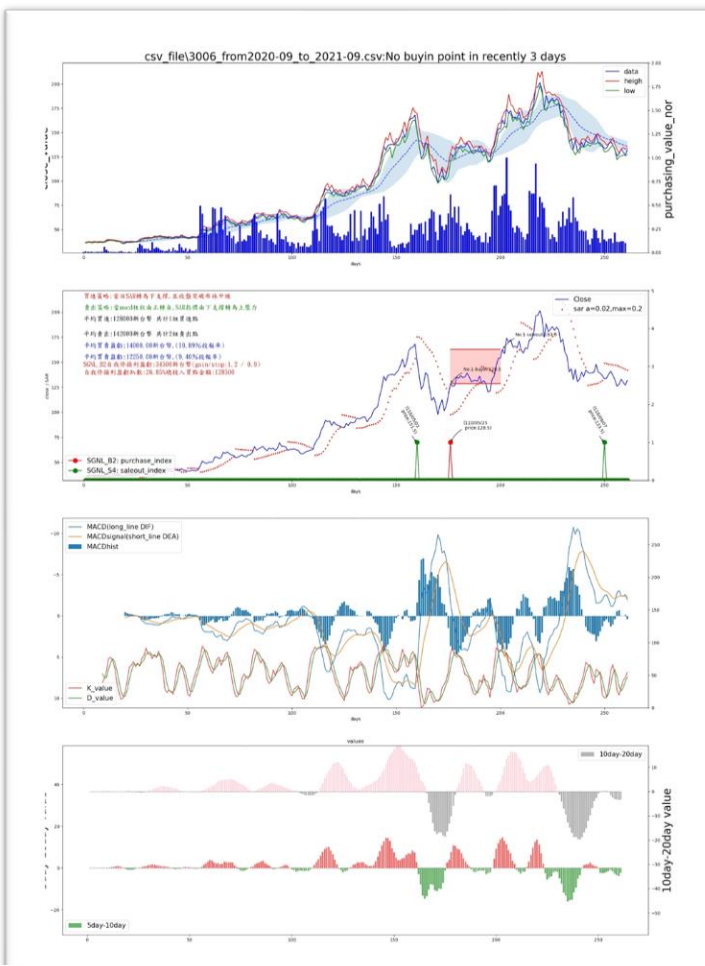
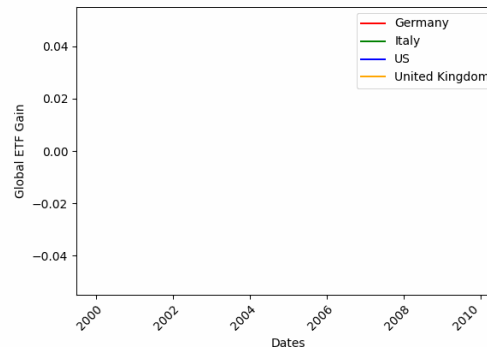
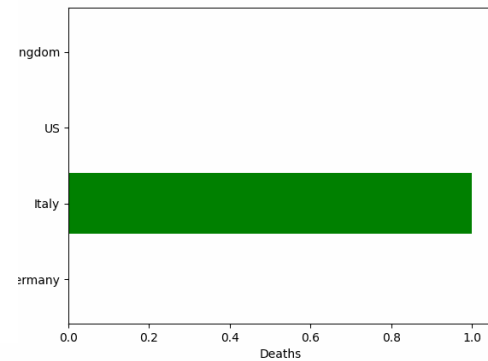
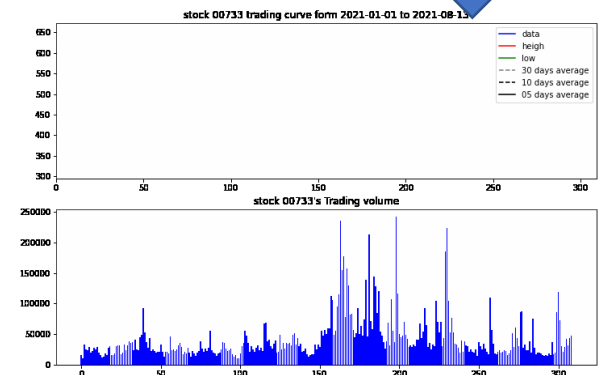
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明

日期,成交股數,成交金額,開盤價,最高價,最低價,收盤價,漲跌價差,成交筆數

0,109/06/01,"37,936,214",11,205,227.094,234.00,296.50,293.50,295.50,+3.50,"15,502"
1,109/06/02,"26,663,587",7,908,476.089,296.00,297.50,296.00,296.50,+1.00,"9,743"
2,109/06/03,"67,894,337",20,354,971.677,300.00,301.00,298.00,301.00,+4.50,"33,062"
3,109/06/04,"47,225,322",14,414,983.522,305.00,306.00,304.00,306.00,+5.00,"24,482"
4,109/06/05,"44,077,262",13,655,618.646,308.50,312.00,308.00,311.50,+5.50,"24,096"
5,109/06/08,"52,042,921",16,500,836.851,316.00,319.00,315.00,318.00,+6.50,"28,581"
6,109/06/09,"37,068,082",11,746,874.522,316.50,319.00,314.00,319.00,+1.00,"19,981"
7,109/06/10,"42,207,259",13,566,733.306,319.00,324.00,318.00,322.50,+3.50,"22,670"
8,109/06/11,"50,612,255",16,340,456.218,325.50,327.00,318.50,320.50,-2.00,"27,061"
9,109/06/12,"46,970,918",14,811,252.573,313.00,317.50,312.50,316.00,-4.50,"25,263"
10,109/06/15,"31,032,404",15,955,585.136,316.00,317.50,308.50,309.50,-6.50,"15,959"
11,109/06/16,"40,725,216",12,823,172.756,317.00,317.00,314.00,315.00,+5.50,"18,836"
12,109/06/17,"34,692,968",10,922,017.056,316.50,317.00,313.50,315.00,0.00,"15,959"
13,109/06/18,"35,443,474",11,123,108.346,314.50,315.00,313.00,314.50,+0.00,"11,897"
14,109/06/19,"48,130,280",15,102,300.186,314.00,314.50,312.00,314.50,0.00,"12,915"
15,109/06/22,"37,374,616",11,722,688.924,314.50,316.50,312.00,312.00,-2.50,"18,836"
16,109/06/23,"41,300,084",12,981,121.544,316.00,316.50,312.50,315.00,+3.00,"15,959"
17,109/06/24,"55,464,024",17,641,164.202,319.00,320.00,316.00,317.50,+2.50,"27,086"
18,109/06/29,"56,672,892",17,695,228.688,314.00,315.00,310.00,312.00,-5.50,"30,558"
19,109/06/30,"49,241,860",15,386,428.040,313.50,314.00,311.00,313.00,+1.00,"16,165"
20,109/07/01,"29,685,342",9,400,765.756,315.00,318.00,314.00,317.50,+4.50,"15,617"
21,109/07/02,"27,789,178",12,119,073.560,319.00,322.00,318.00,322.00,+4.50,"18,355"
22,109/07/03,"55,656,987",18,280,314.743,327.00,330.50,326.50,329.50,+7.50,"30,558"
23,109/07/06,"57,256,220",19,200,718.243,332.50,338.00,332.00,338.00,+8.50,"30,558"
24,109/07/07,"59,046,323",20,143,258.406,343.50,346.00,338.00,338.50,+0.50,"15,617"



report.html



Introduction yourself

Name?

各位學員介紹自己, 以及嗜好

Career?

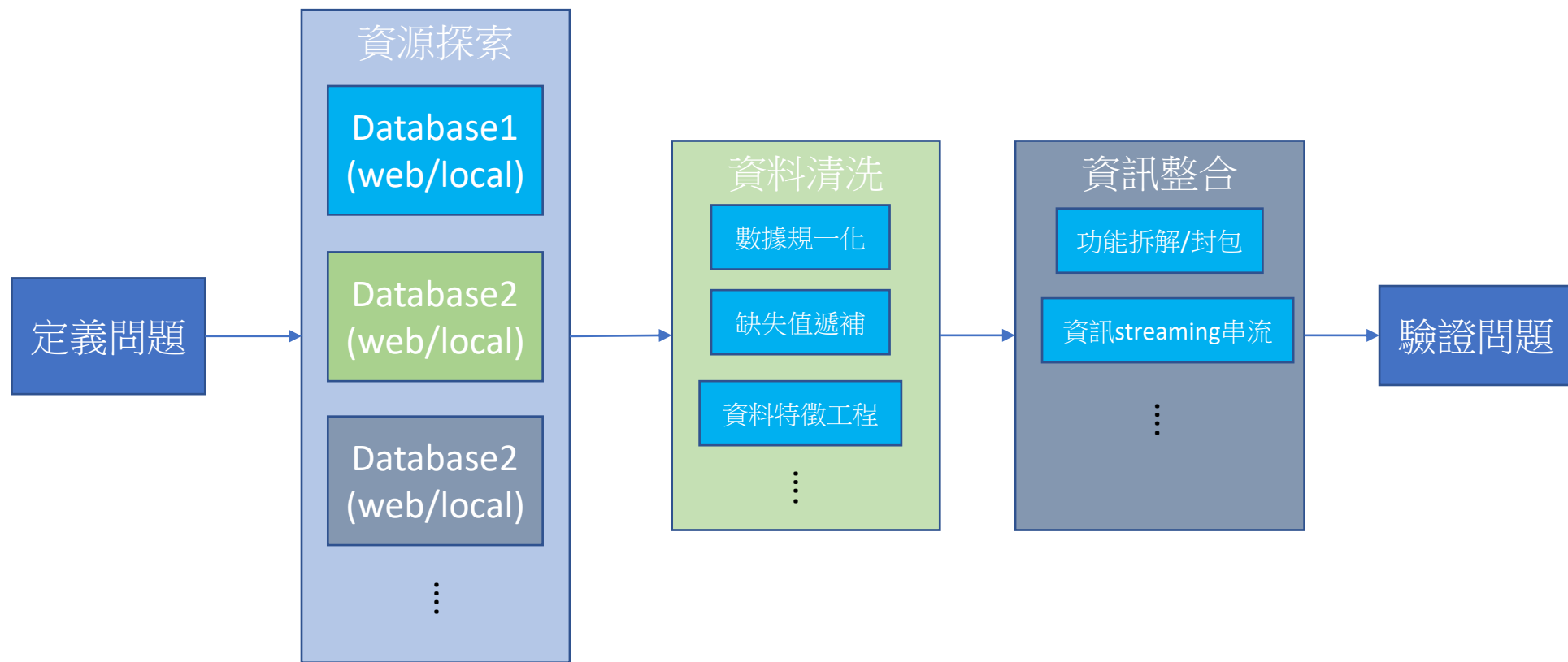
工作類型, 會處理到的分析工具, 或者數據是什麼類型? 目前對於程式基礎有哪些?

Any ideal?

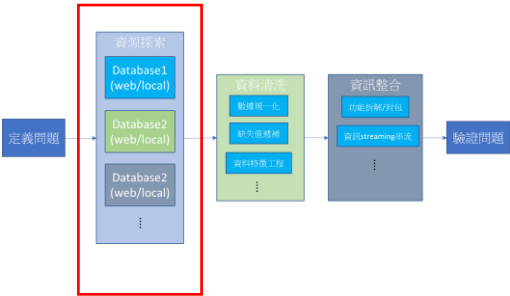
花了時間上課, 希望得到什麼?

爬蟲基礎入門-Part 1

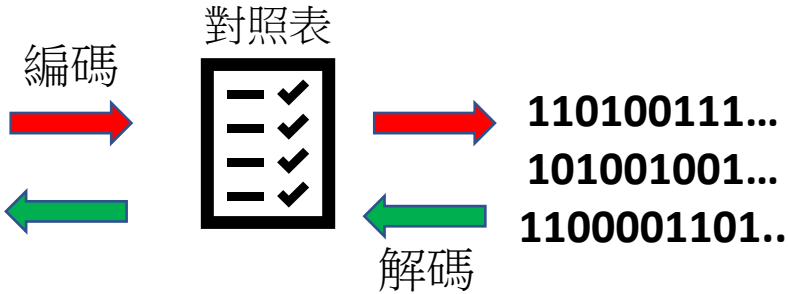
FLOW



資料流程概述-何謂編碼



```
ws.on("message", m => {  
  let a = m.split(" ");  
  switch(a[0]){  
    case "connect":  
      if(a[1]){  
        if(!clients.has(a[1])){  
          ws.send("connected");  
          ws.id = a[1];  
        }else{  
          ws.id = a[1];  
          clients.set(a[1], {client: {position: {x: 0, y: 0, z: 0}}});  
          ws.send("connected");  
        }  
      }  
    }  
  }  
});
```



ASCII(英文)
BIG5(繁中)
GBK(簡中)

...

ASCII(英文) 對照表

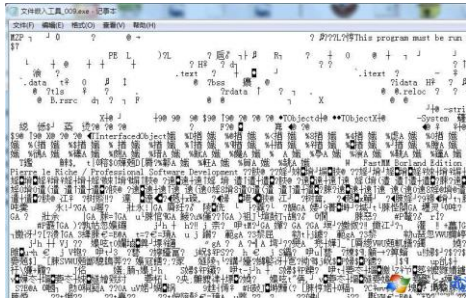
BIG5(繁中)對照表

亂碼

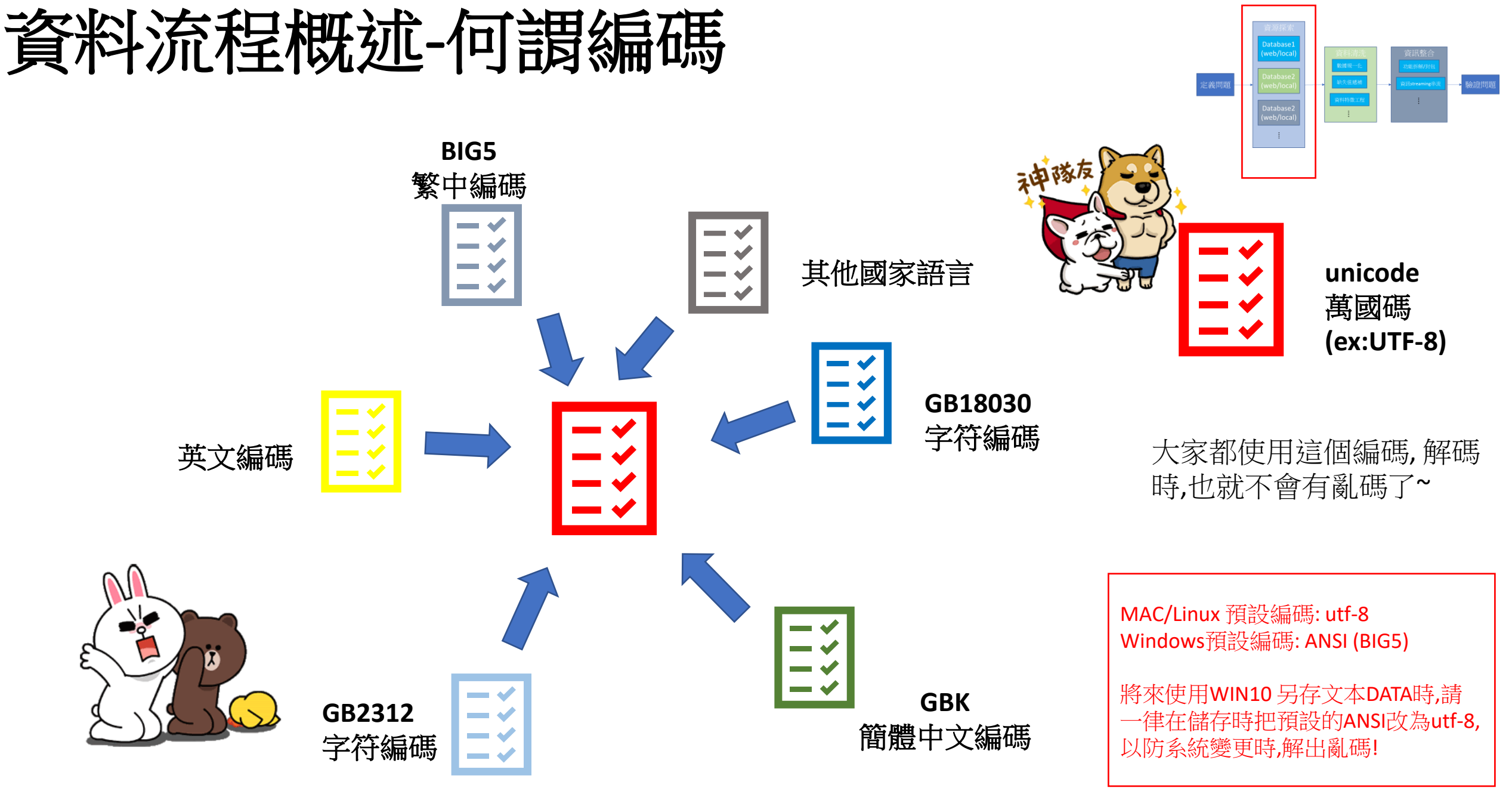
```
ws.on("message", m => {  
  let a = m.split(" ");  
  switch(a[0]){  
    case "connect":  
      if(a[1]){  
        if(!clients.has(a[1])){  
          ws.send("connected");  
          ws.id = a[1];  
        }else{  
          ws.id = a[1];  
          clients.set(a[1], {client: {position: {x: 0, y: 0, z: 0}}});  
          ws.send("connected");  
        }  
      }  
    }  
  }  
});
```



110100111...
101001001...
1100001101..

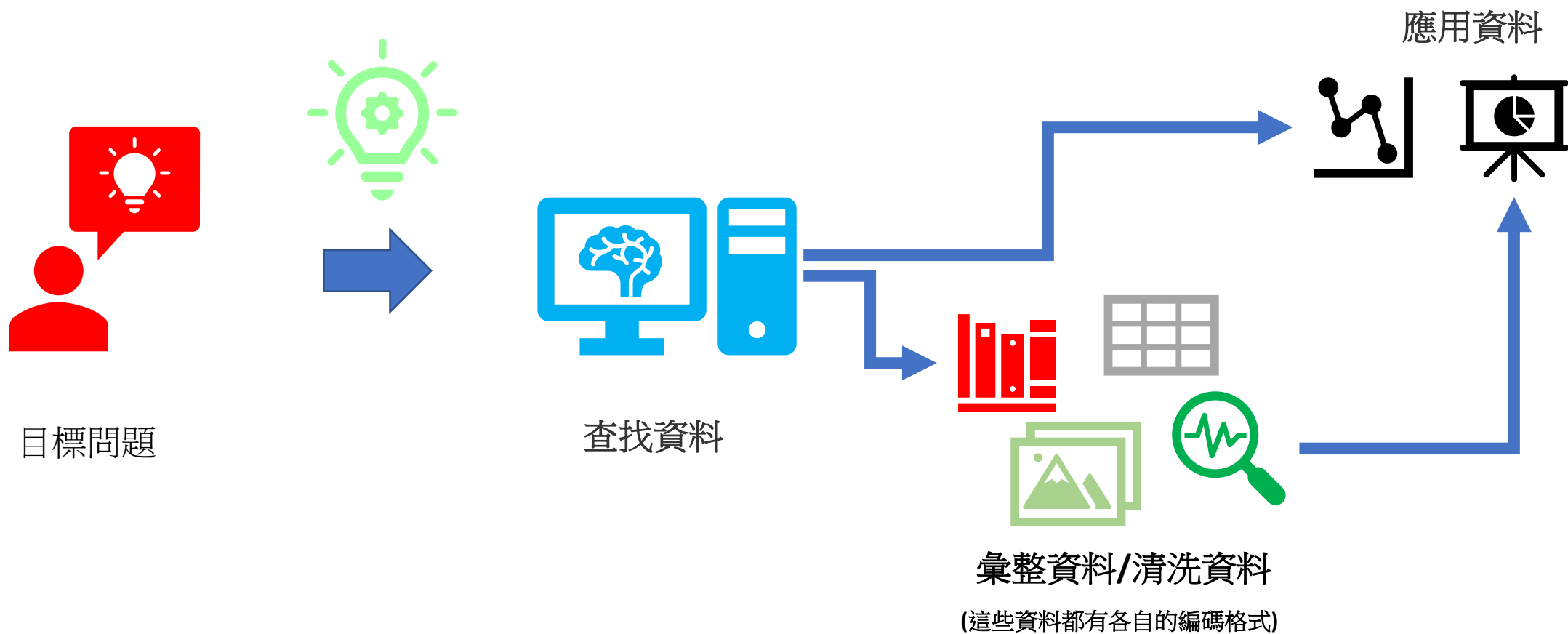


資料流程概述-何謂編碼



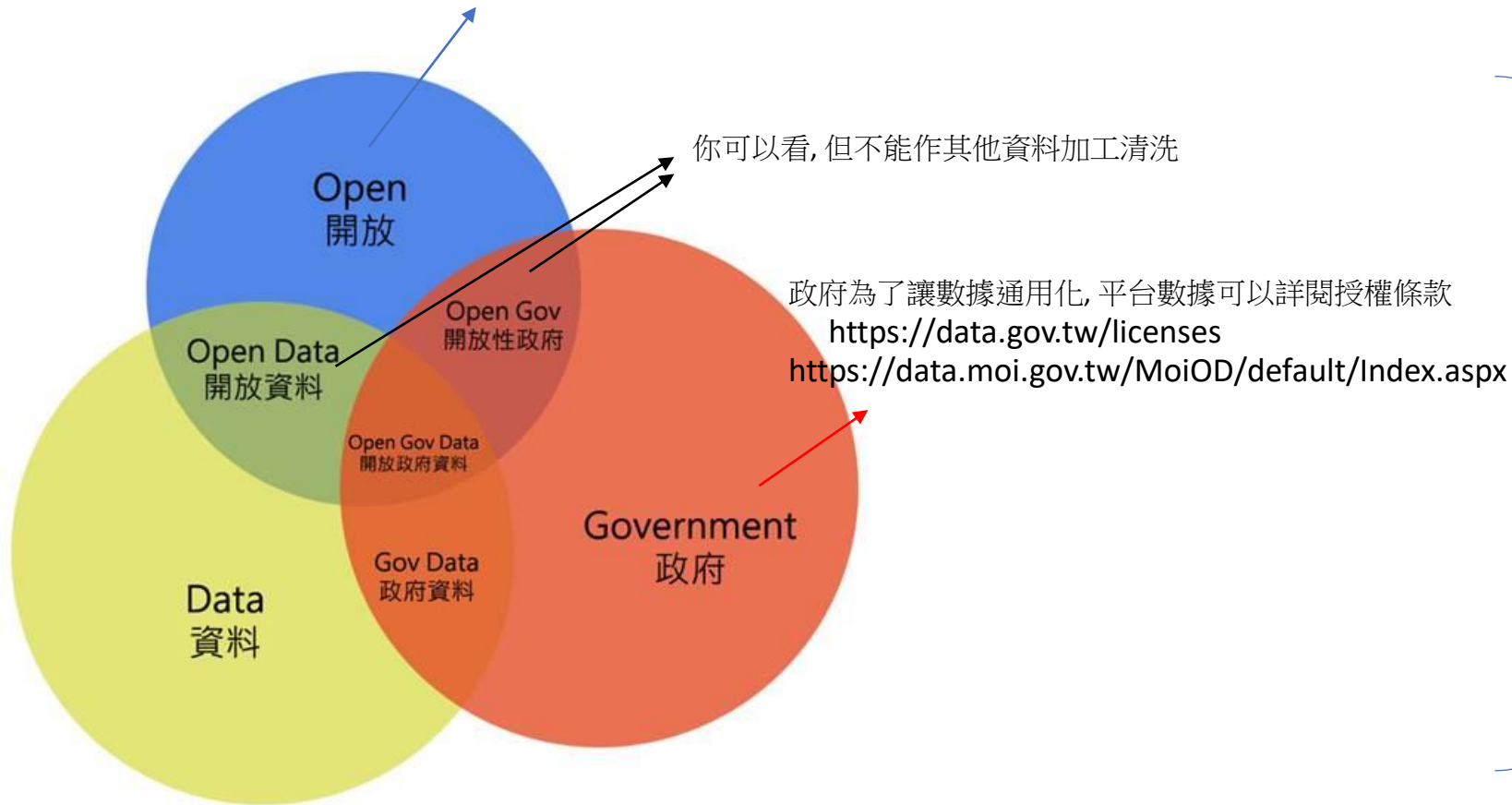
大家都使用這個編碼, 解碼時, 也就不會有亂碼了~

資料流程概述



數據哪裡找?

不受著作權、專利權，以及其他管理機制所限制，可以開放給社會公眾，任何人都可以自由出版使用



json

```
[{"name": "Tom",  
  "lastname": "Chen",  
  "report": [{"subject": "Math", "score": 80}, {"subject": "English", "score": 90}]}
```

CSV

xml

```
<!DOCTYPE html>  
<html>  
  <head>...</head>  
  <body> == $0  
    <div id="sysDialog" title="系統公告"></div>  
    <script type="text/javascript">...</script>  
    <div id="container">...</div>  
    <div tabindex="-1" role="dialog" class="ui-dialog ui-corner-all ui-widget ui-wi  
dget-content ui-front ui-draggable ui-resizable" aria-describedby="disclaimerDi  
v" aria-labelledby="ui-id-1" style="display: none; position: absolute;">...</div>  
    <div tabindex="-1" role="dialog" class="ui-dialog ui-corner-all ui-widget ui-wi  
dget-content ui-front ui-draggable ui-resizable" aria-describedby="privacyDiv"  
aria-labelledby="ui-id-2" style="display: none; position: absolute;">...</div>  
    <ul id="ui-id-3" tabindex="0" class="ui-menu ui-widget ui-widget-content ui-aut  
ocomplete ui-front" style="display: none;"></ul>  
    <div role="status" aria-live="assertive" aria-relevant="additions" class="ui-he  
lper-hidden-accessible"></div>  
  </body>  
</html>
```

常用資料格式:

CSV

CSV (Comma Separated Values) 逗號分隔值，是一種常見的資料格式，使用逗號將不同欄位做為分隔。可以使用一般的文字編輯器以原始格式開啟，也可以使用 excel 或 number 等試算表軟體以表格方式開啟。



Data.csv

```
Year,Make,Model,Description,Price
1997,Ford,E350,"ac, abs, moon",3000.00
1999,Chevy,"Venture ""Extended Edition""",,,4900.00
1999,Chevy,"Venture ""Extended Edition, Very Large""",,5000.00
1996,Jeep,Grand Cherokee,"MUST SELL!
air, moon roof, loaded",4799.00
```

有換行符號, 實際轉到python時, 若編碼部正確, 會導致誤判

Utf-8, big5, utf-8, ACSII... 編碼未限制, 但通常是用utf-8

JSON

JSON (JavaScript Object Notation) 一種輕量級資料交換格式。其內容由屬性和值所組成，因此也有易於閱讀和處理的優勢。JSON是獨立於程式語言的資料格式。



Data.json


```
[
  {
    "text": "This is the text",
    "color": "dark_red",
    "bold": "true",
    "strikethrough": "true",
    "clickEvent": {
      "action": "open_url",
      "value": "zh.wikipedia.org"
    },
    "hoverEvent": {
      "action": "show_text",
      "value": {
        "extra": "something"
      }
    }
  },
  {
    "translate": "item.dirt.name",
    "color": "blue",
    "italic": "true"
  }
]
```

常用資料格式：

XML

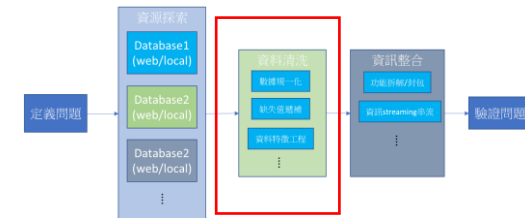
可延伸標記式語言（英語：Extensible Markup Language，簡稱：XML）是一種標記式語言。XML是從標準通用標記式語言（SGML）中簡化修改出來的。它主要用到的有可延伸標記式語言、可延伸樣式語言（XSL）。

`<?xml version="1.0" encoding="UTF-8"?>`. 版本/編碼要記得!

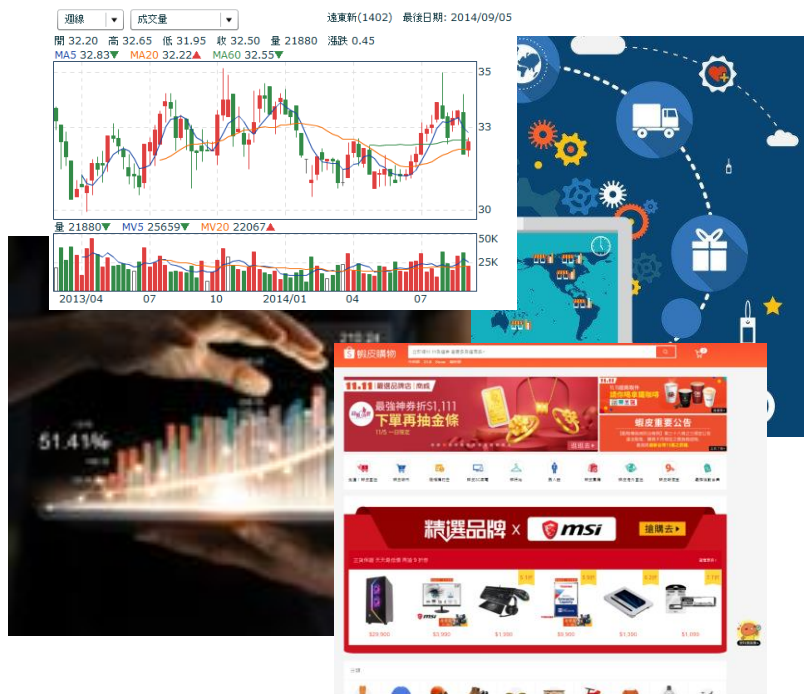


```
<?xml version="1.0"?>
<小纸条>
  <收件人>大元</收件人>
  <發件人>小張</發件人>
  <主題>問候</主題>
  <具體內容>早啊，飯吃了沒？ </具體內容>
</小纸条>
```

找不到,但就在眼前, 怎麼辦?



你想要的結果呈現

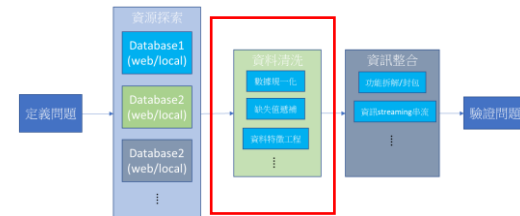


資料爬取/清洗



各位正要學的, 其實就是一種資料清洗/整理個方式

何謂數據規一化



常見的特徵處理手法:

Normalization

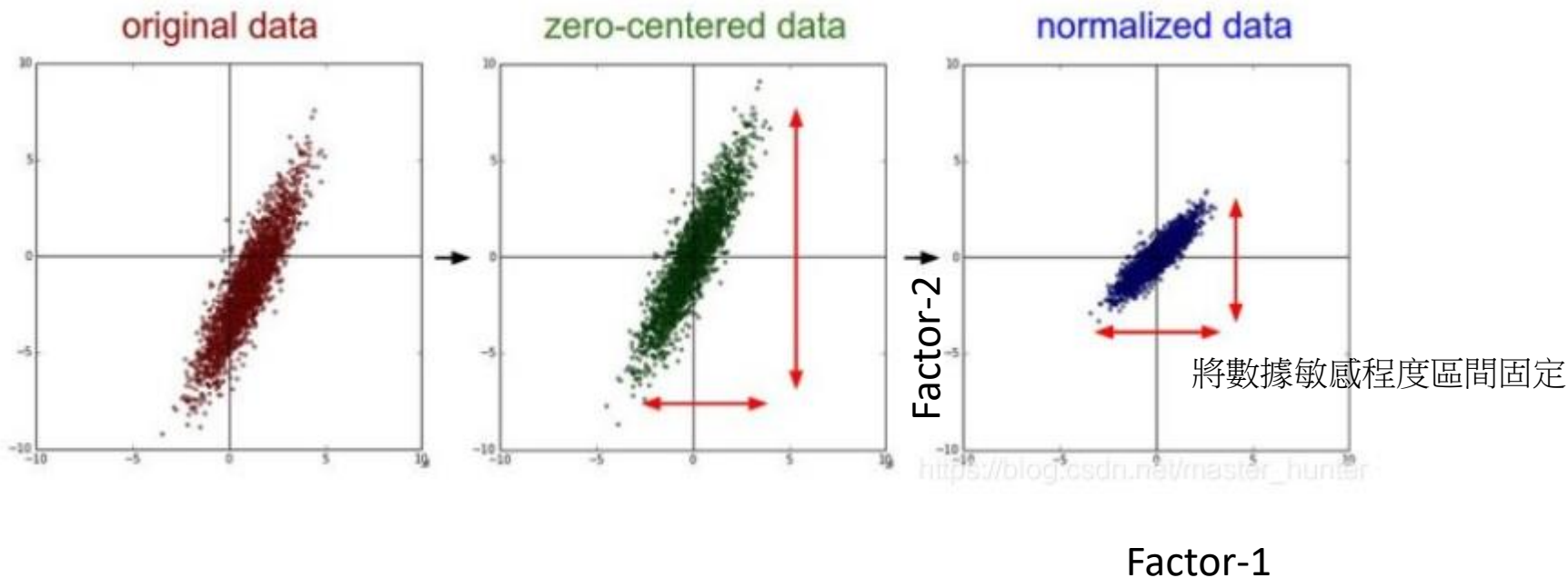
$$X_{new} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

Mean normalization

$$X_{new} = \frac{X_i - \text{mean}(X)}{X_{max} - X_{min}}$$

Standardization

$$X_{new} = \frac{X_i - \mu}{\sigma}$$



假設某位學生第一次段考數學成績是57分，國文成績是73分；而全班的數學平均分數是43分，標準差7分，全班的國文平均分數是78分，標準差5分，是否能說此生的國文考的比數學好呢？
解答：不能單以表面的73分比57分高，就說國文考的比數學好！

數學標準化成績是 $\frac{57-43}{7}=2$

國文標準化成績是 $\frac{78-73}{5}=1$

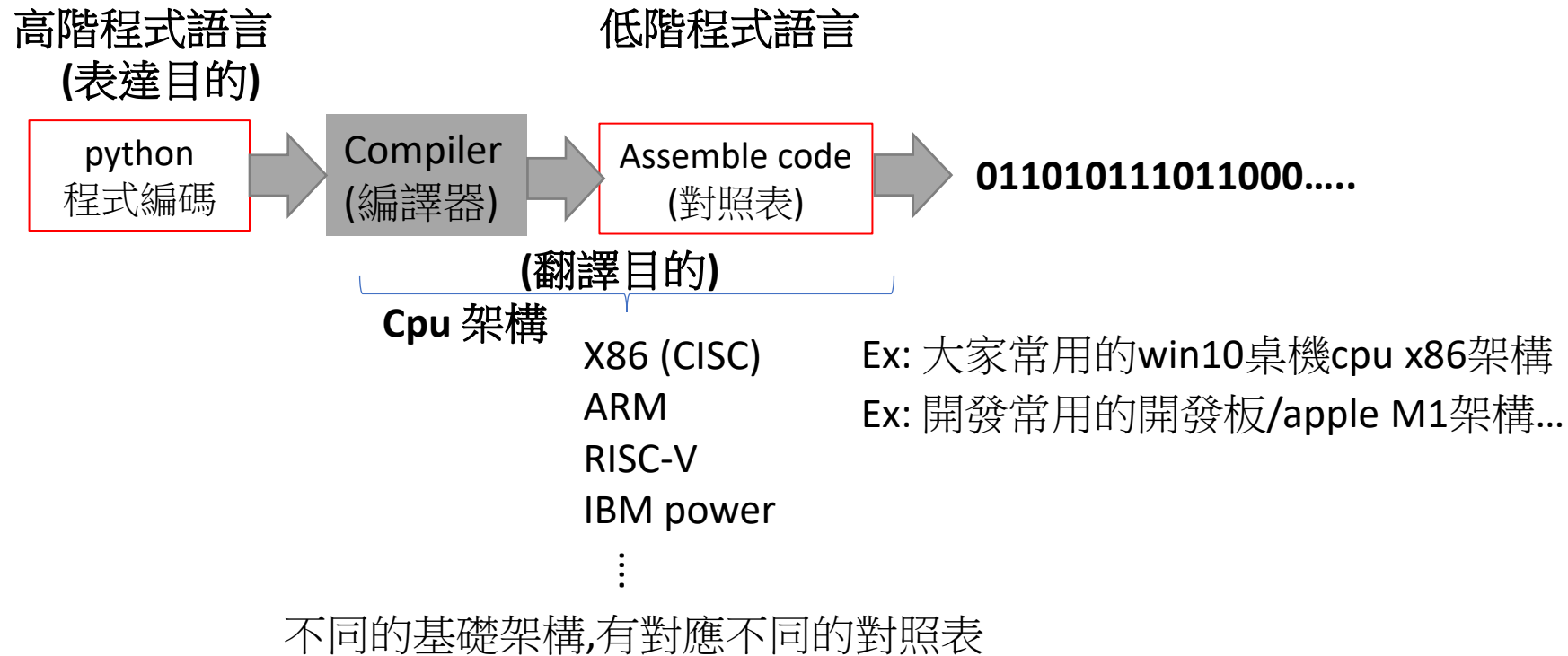
所以應是數學成績比國文成績好

爬取工具介紹-程式語言



Jan 2021	Jan 2020	Change	Programming Language	Rat	
1	2	▲	C	17.	
2	1	▼	Java	11.96%	-4.93%
3	3		Python	11.72%	+2.01%
4	4		C++	7.56%	+1.99%
5	5		C#	3.95%	-1.40%
6	6		Visual Basic	3.84%	-1.44%
7	7		JavaScript	2.20%	-0.25%
8	8		PHP	1.99%	-0.41%
9	18	▲▲	R	1.90%	+1.10%
10	23	▲▲	Groovy	1.84%	+1.23%

爬取工具介紹-程式語言



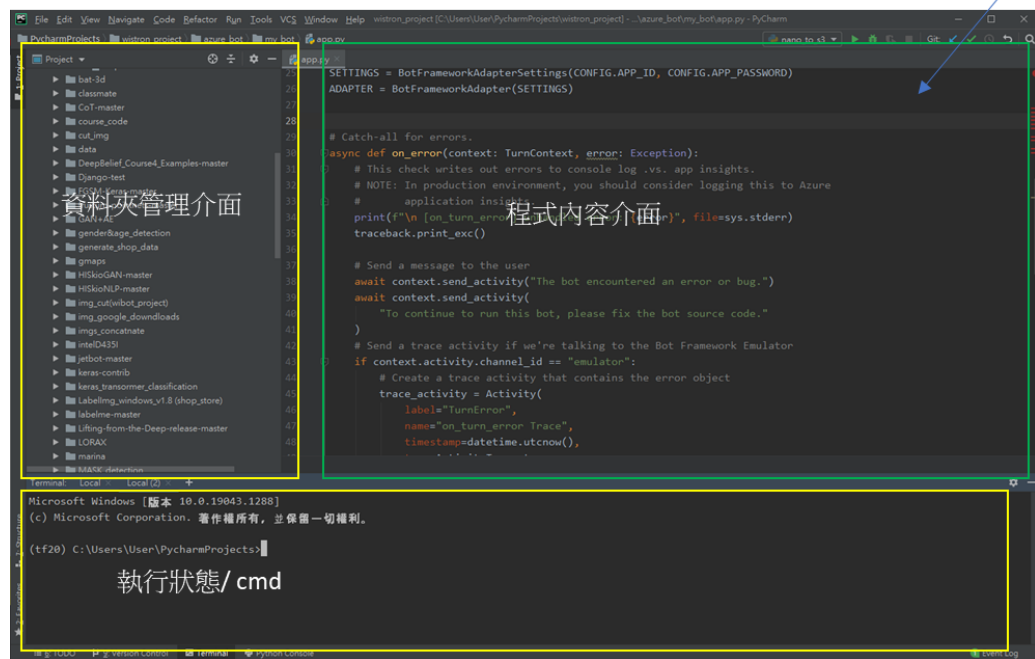
所以對我們來說, **python**, 我們只在意高階語言的內容, 至於編譯成電腦懂得信息, 則在一開始安裝系統環境時, 就已經順帶安裝**python**於各個架構的編譯器了學~

爬取工具介紹-IDE工具

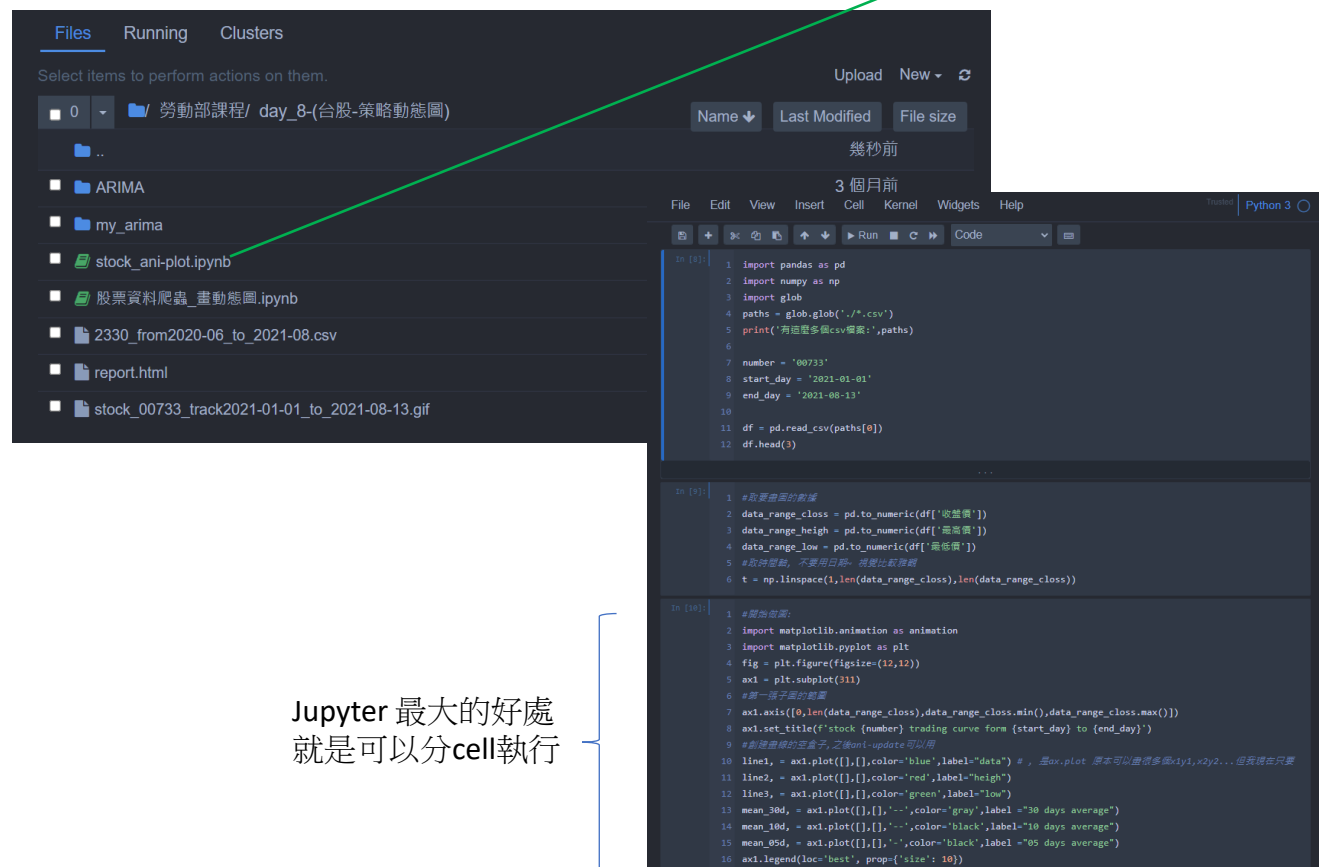


<https://www.jetbrains.com/pycharm/>

跑你的 py檔

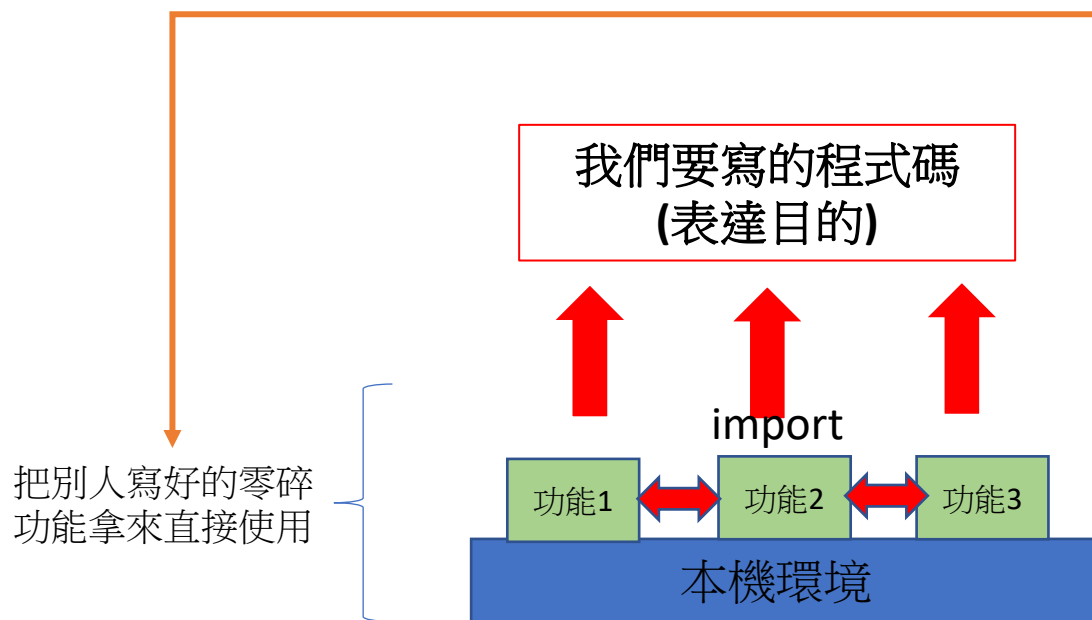


原本的py檔, 在jupyter 要轉為ipynb格式才能分段執行



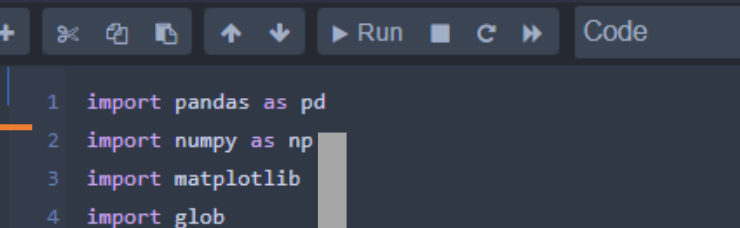
Jupyter 最大的好處
就是可以分cell執行

有了程式碼, 那麼什麼是版本控管?



Pip: 只要你有安裝python 就會自帶這個安裝小幫手

- `pip3 install pandas==2.0.1` (win10 使用pip 安裝工具範例)
- `sudo apt install python3-pandas==2.0.1` (ubuntu 使用apt 安裝工具範例)



The screenshot shows a Jupyter Notebook interface. At the top, there is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. Below the menu bar is a toolbar with icons for saving, adding, deleting, and running code. The main area displays a code cell with the following Python code:

```
In [8]: 1 import pandas as pd
2 import numpy as np
3 import matplotlib
4 import glob
5 paths = glob.glob('*.csv')
6 # print('有這麼多個檔案:', paths)
7 number = '00733'
8 start_day = '2021-01-01'
9 end_day = '2021-08-31'
10
11 df = pd.read_csv(paths[0])
12 df.head(3)
```

名稱	修改日期	類型	大小
__pycache__	2020/10/13 下午 03:00	檔案資料夾	
_config	2020/10/13 下午 03:00	檔案資料夾	
_libs	2020/10/13 下午 03:00	檔案資料夾	
api	2020/10/13 下午 03:00	檔案資料夾	
arrays	2020/10/13 下午 03:00	檔案資料夾	
compat	2020/10/13 下午 03:00	檔案資料夾	
core	2020/10/13 下午 03:00	檔案資料夾	
errors	2020/10/13 下午 03:00	檔案資料夾	
io	2020/10/13 下午 03:00	檔案資料夾	
plotting	2020/10/13 下午 03:00	檔案資料夾	
tests	2020/10/13 下午 03:00	檔案資料夾	
tseries	2020/10/13 下午 03:00	檔案資料夾	
util	2020/10/13 下午 03:00	檔案資料夾	
init.py	2020/10/8 上午 12:35	Python File	11 KB
_testing.py	2020/10/8 上午 12:35	Python File	89 KB
_typing.py	2020/10/8 上午 12:35	Python File	4 KB
_version.py	2020/10/8 上午 12:35	Python File	1 KB
confTest.py	2020/10/8 上午 12:35	Python File	31 KB
testing.py	2020/10/8 上午 12:35	Python File	1 KB

每一個別人寫好的功能都都是一個資料夾的功能包

Coding time

環境重點:

1. 安裝python 3.6 或3.7 /設定到環境變數中(win10)
2. 安裝pycharm IDE, 並且pip 一系列之後會用到的工具包
3. 安裝anaconda3 可以做虛擬環境的版本控制 (進階知識)

Coding 重點:

1. 學習資料匯入匯出/encode 格式
2. Import 工具的應用
3. If/else判斷式撰寫
4. List / dir / tuple / set / 資料集型態對應到的工具使用
5. For in 循環 while 循環 介紹

QA