l.pulipuli.info/22/nsysu

中文自然語言處理動手玩

http://l.pulipuli.info/22/nsysu

政治大學人工智慧與數位教育中心

陳勇汀 研究員

pudding@nccu.edu.tw

2022年

課程網頁

1.pulipuli.info/22/nsysu

陳勇汀 (布丁) 講者簡介

現職:

- 國立政治大學人工智慧與數位教育中心 研究員

學歷:

- 國立政治大學圖書資訊與檔案學研究所 博士

專長:

- 自然語言處理
- 人工智慧與資料探勘
- 數位人文

布丁布丁吃什麼？
http://blog.pulipuli.info

課程大綱

1. 自然語言處理基本流程
2. 語義向量 (Embedding)
3. 語義向量實作
4. 分類應用 : 問答判斷
    - 實作 : 設計問答資料集
5. 分群應用
    - 實作 : 用Python實作分群
6. 結語

聽演講就是來聊天！

slido

Join at

slido.com

#495 224

聽演講就是來聊天！

隨時提問OK

可匿名好安心

https://www.sli.do/

Part 1.

自然語言處理基本流程

電腦也吃花生？略吃

電腦也吃花生？略吃

非結構化資料

電腦也吃花生？略吃

非結構化資料

結構化資料

長度cm	重量g
2.98	0.74
3.5	0.76

特徵萃取

文本資料的特徵萃取

非結構化資料

兄弟們！來！來！
來和他們一拼！
憑我們有這一身，
　我們有這雙腕，
休怕他毒氣、機關槍！
休怕他飛機、炸裂彈！
來！和他們一拼！

——賴和《南國哀歌》

結構化資料

包含詞彙：

- 兄弟
- 們
- 來
- 和
- 他們
- 一
- 拼
- 憑
- 我們
- 有
- 這
- 一身
- 這雙
- 腕
- 休
- 怕
- 他
- 毒氣
- 機關槍
- 飛機
- 炸裂彈

文本資料的特徵萃取

非結構化資料

5 休怕他毒氣、機關槍！

6 休怕他飛機、炸裂彈！

結構化資料 (詞袋模型)

文本編號	休	怕	他	毒氣	機關槍	飛機	炸裂彈
5	1	1	1	1	1	0	0
6	1	1	1	0	0	1	1

1=有這個字; 0=沒有這個字

文本模型到電腦的「理解」

結構化資料
(詞袋模型)

文本編號	他	毒氣	機關槍
5	1	1	1
6	1	0	0

我站在雲林

https://tenor.com/view/dance-android-robot-gif-15459115

自然語言處理流程

1. 分解 — 最小意義單位的分析
2. 轉換 — 統一不同的用詞
3. 移除 — 刪除多餘的資訊
4. 向量化 — 轉成電腦可讀的結構化文本模型

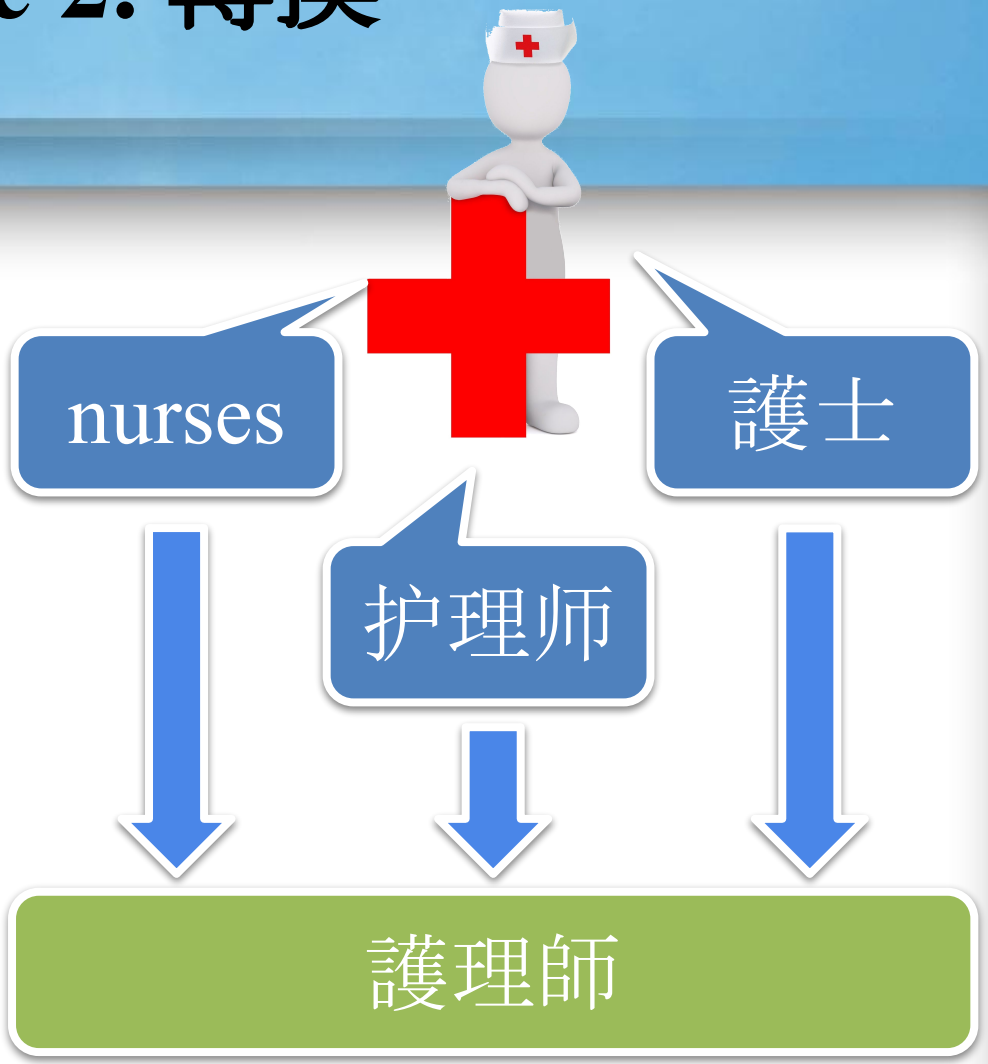
Phase 1分解
最小意義單位分析方式

- 空格
- n-gram：n字詞，將文本拆成以n個字組成的詞彙
- 詞典法：以既有詞典為基礎，配合演算法來決定斷詞方式

斷開魂結！
斷開鎖鏈！
斷開一切的牽連！

Phase 2. 轉換

1. 翻譯
   a. 跨語言的翻譯
   b. 英文的大寫轉換小寫
   c. 中文的繁簡字轉換
2. 英文的詞幹化
3. 主題詞表轉換



Phase 3. 移除

- 移除標點符號
- 移除數字
- 移除HTML標籤
- 移除非主要使用語言
  - 英文研究者常常移除英文以外的語言
- 移除停用詞

下雨天，留客天，留，我不？留！XDD

下雨天 留客天 留 我不 留

Phase 4. 文字向量化
文本模型選擇

	Bag of Words 詞袋	TF 詞頻	TF-IDF 詞頻-反文件頻	one-hot
同文件頻率	X	O	O	O
多文件獨特性	X	X	O	X
字順	X	X	X	O
用途	解讀巨量資料	解讀少量資料	搜尋引擎	自然語言處理

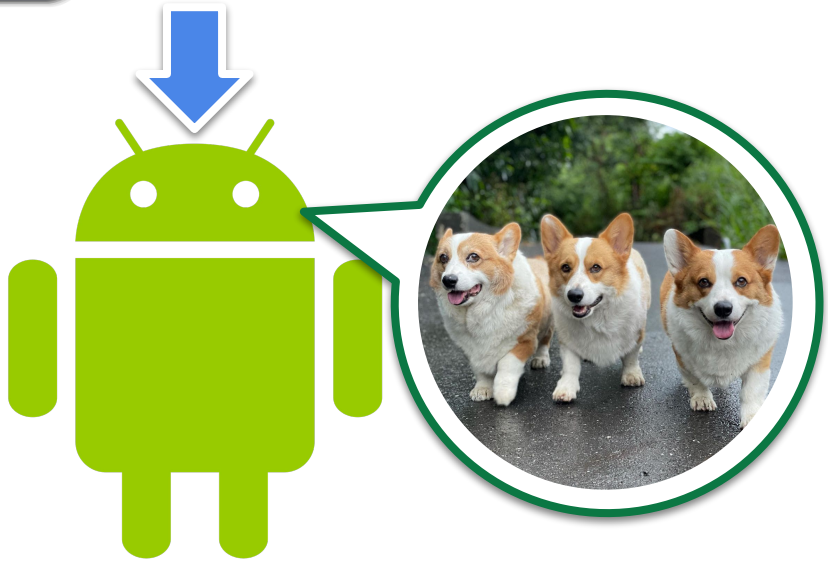
結構化資料 (詞袋模型)

文本編號	他	毒氣	機關槍
5	1	1	1
6	1	0	0

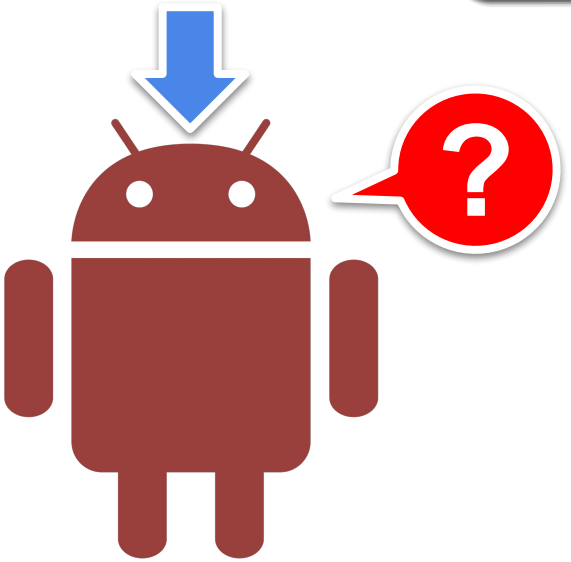
這樣就夠了，嗎？

缺點1：只能接受精確的詞彙

狗 狗 狗



德國牧羊犬

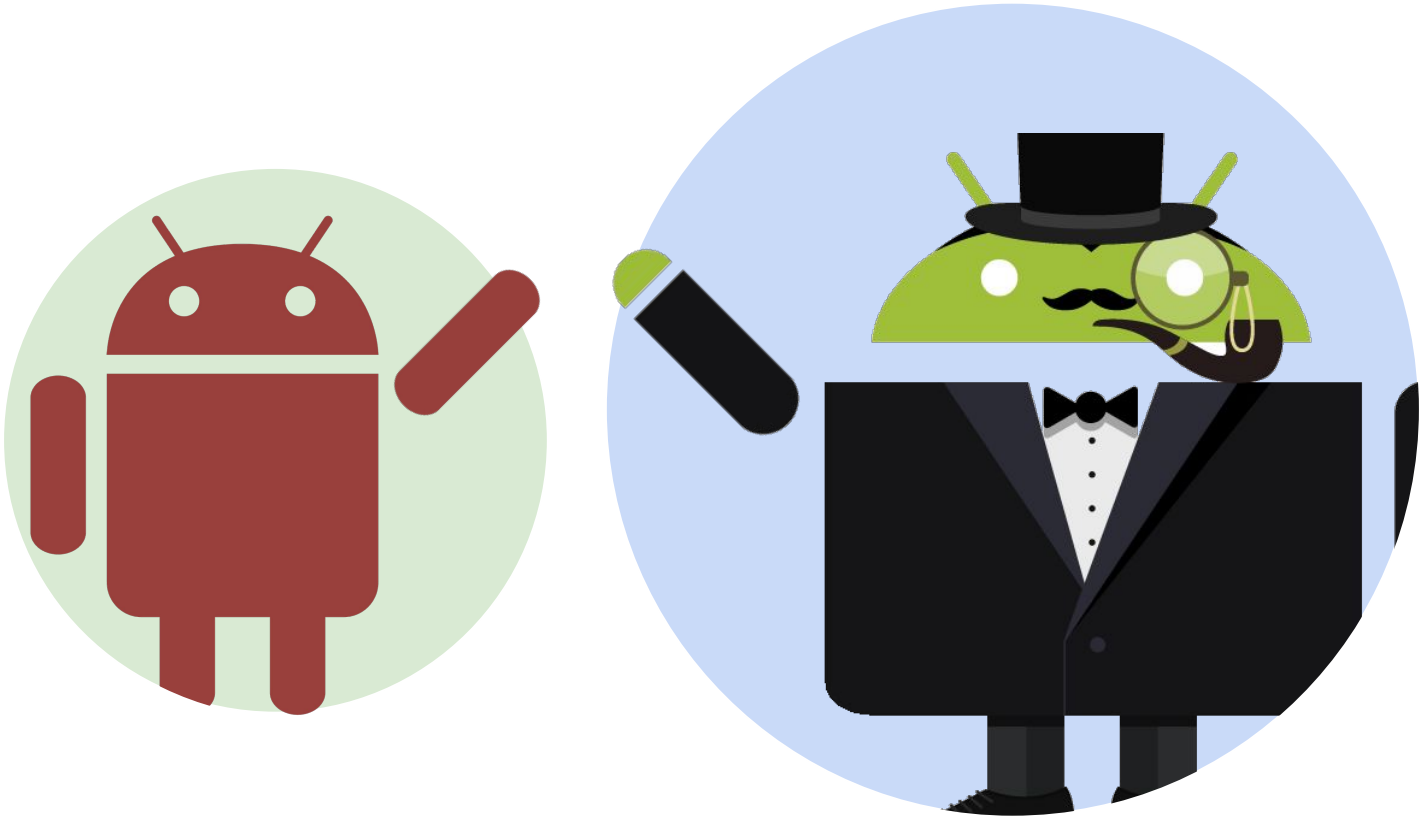


缺點2：脈絡遺失

山上到處是盛開的杜鵑

樹林裡傳來了杜鵑的叫聲

https://www.easyatm.com.tw/wiki/%E5%A4%9A%E7%BE%A9%E8%A9%9E



embedding

交給語義向量吧！

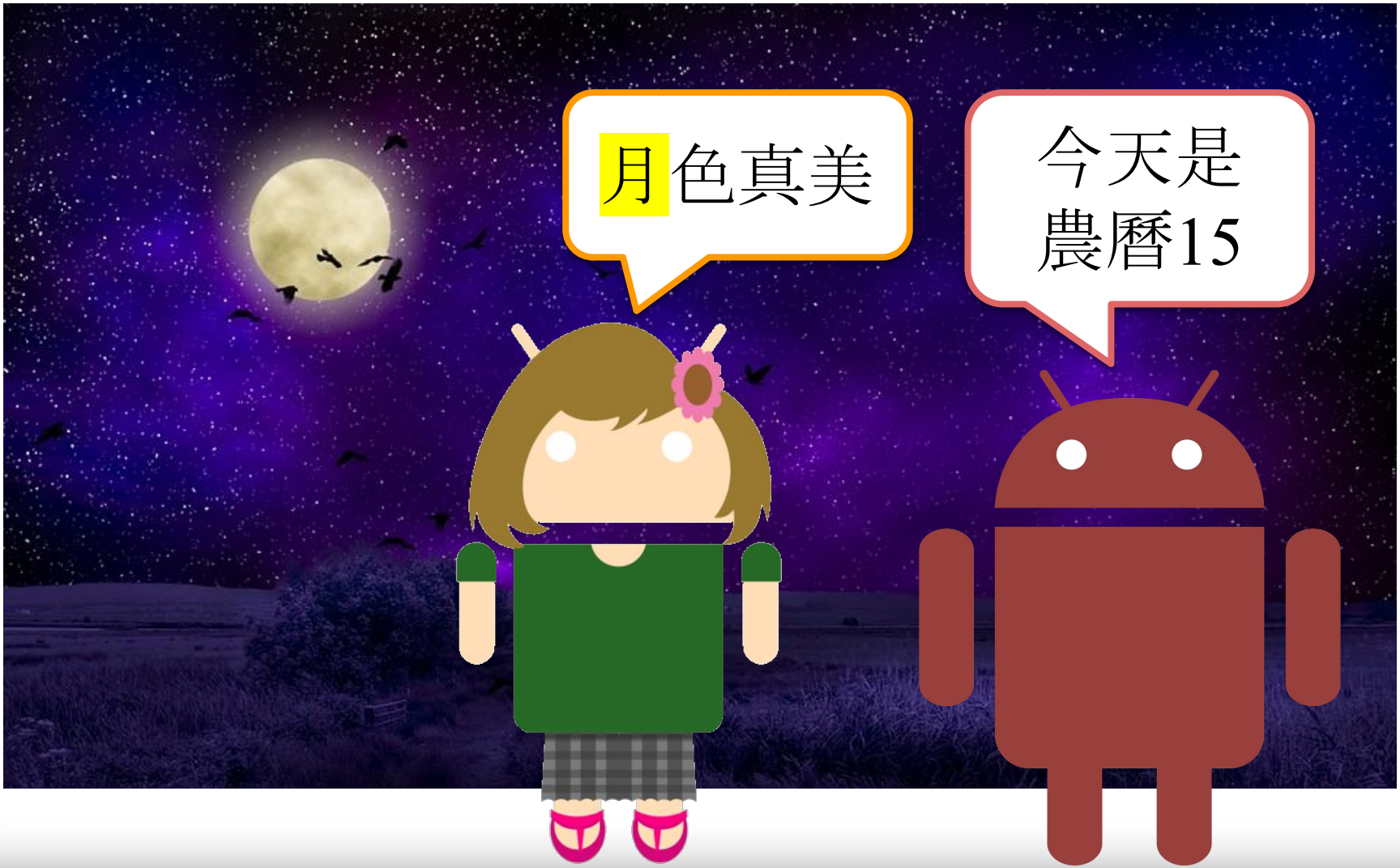
Part 2.

語義向量 Embedding

情境題

https://pixabay.com/illustrations/moon-moonlight-night-sky-dark-2372573/

詞袋模型的反應



https://pixabay.com/illustrations/moon-moonlight-night-sky-dark-2372573/

語義向量的反應

https://pixabay.com/illustrations/moon-moonlight-night-sky-dark-2372573/

怎麼做到的？

「I love you」翻譯做「月色真美」

日本文豪

夏目漱石

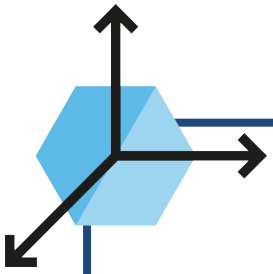
“日本人はそんなことは言わない。月が綺麗ですねとでも訳しておけ。”

日本人不會那樣說，翻譯成月色真美之類的即可


Pre-trained Model
預建置模型

語義向量的產生

月色真美



[-.04, -.05, .05, 0.05, 0,

.....

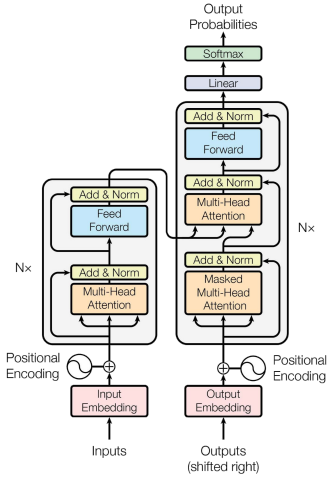
.06 , -0.04, -.05, .03, .06]

Universal Sentence Encoder (USE)

資料來源

- Wikipedia
- 網路新聞
- 網路問答
- 論壇

深度學習網路架構

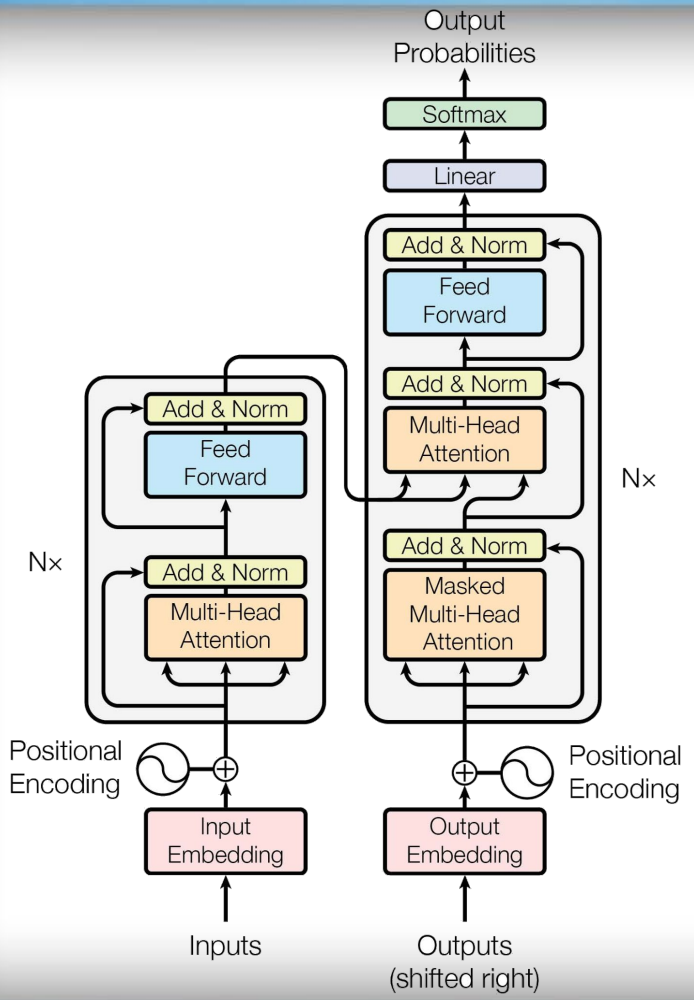


Transformer

預建置模型

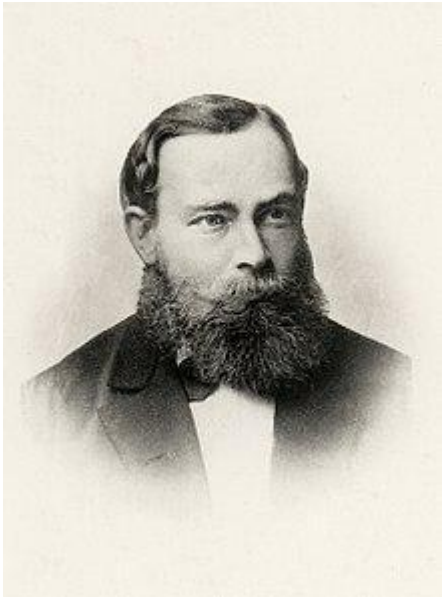
(Cer et al., 2018)

深度學習網路架構
Transformer



- 利用Transformer架構編碼器捕捉語義向量
  - 不同的**文字**
  - 文字的前後順序 (雙向)
    ➡ **文法**
- 語義向量 (sentence embedding)以512維度的向量呈現

the principle of compositionality
複合性原理



Gottlob Frege

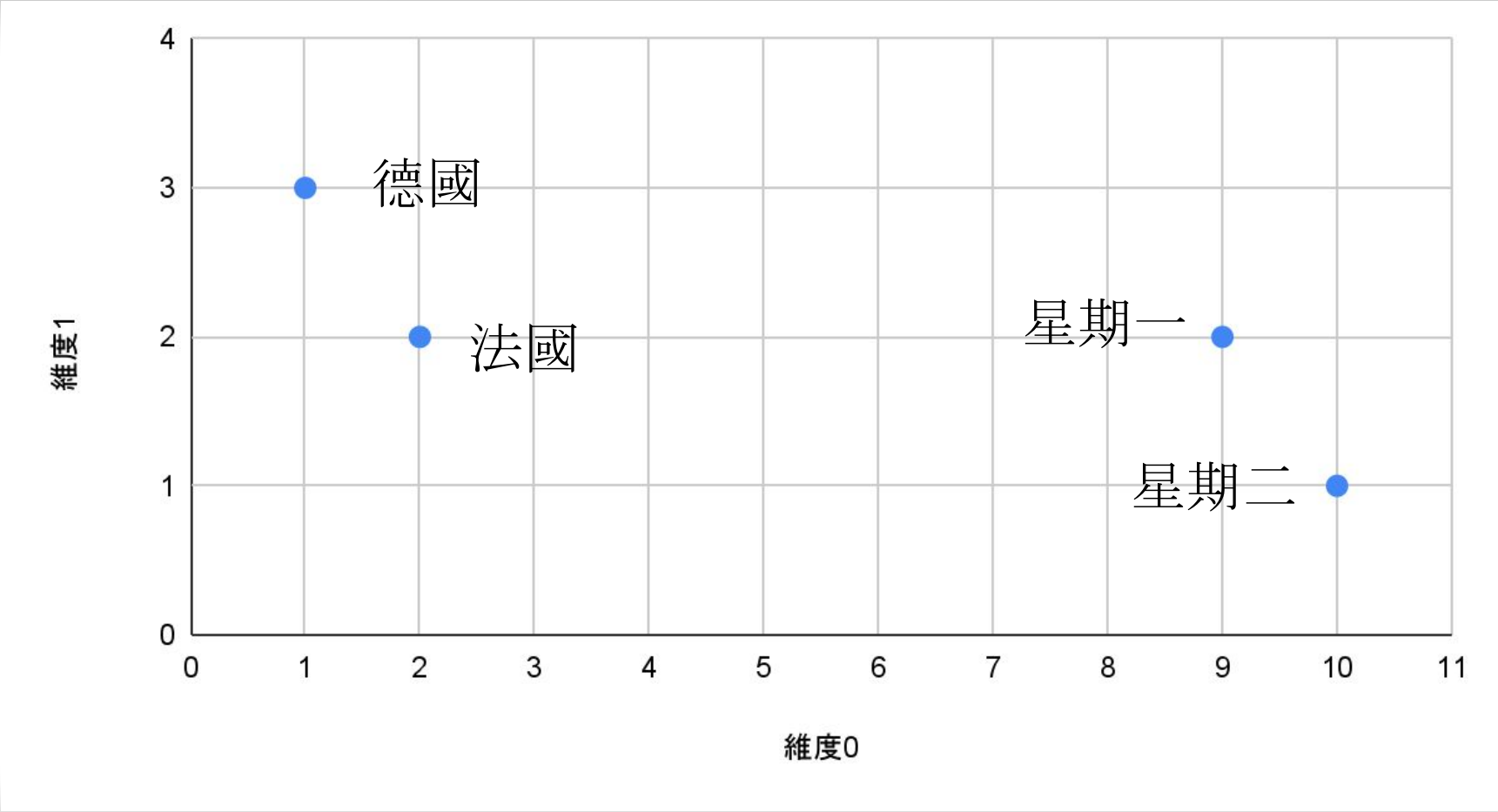
(1848-1925)

- 一個複雜表達式的意義 (=向量)是由
  - 其各組成部分的意義(=文字)以及
  - 用以結合它們的規則 (=文法)來決定的。

簡化版語義向量的例子

文字	語義向量	
	維度0	維度1
德國	[1,	3]
法國	[2,	2]
星期一	[9,	2]
星期二	[10,	1]

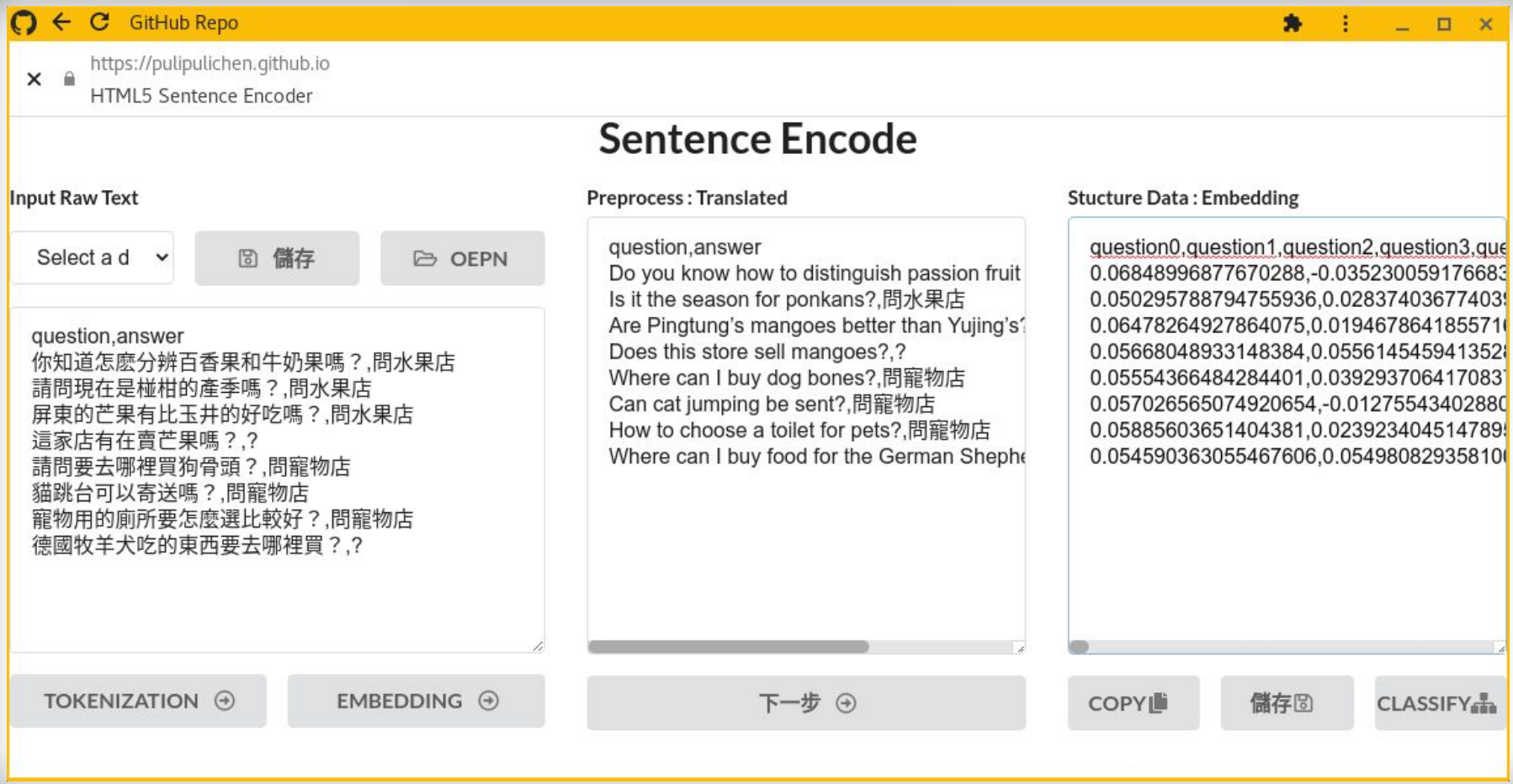
簡化版語義向量的例子



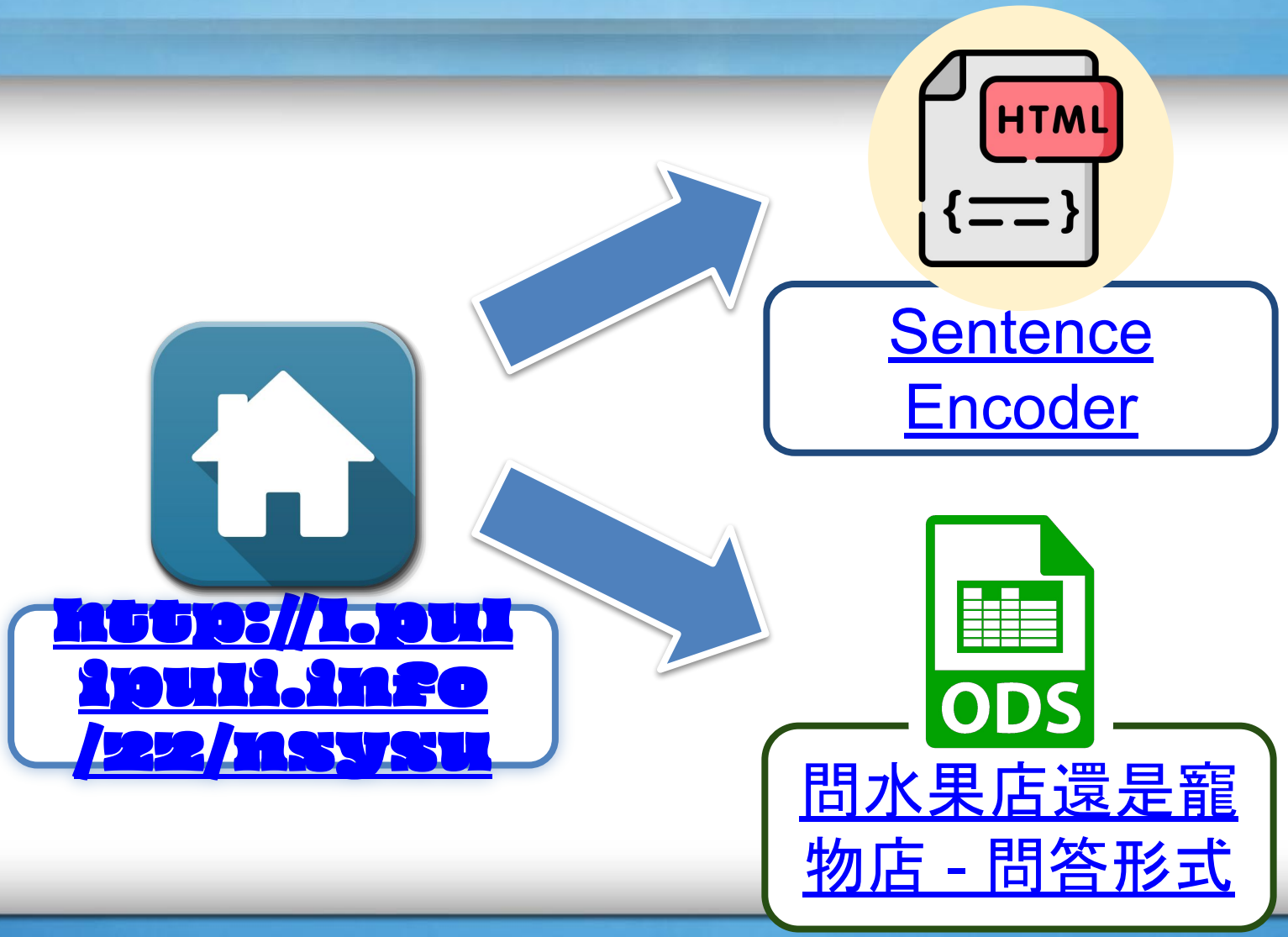
Part 3.

語義向量實作

Sentence Encoder



課程網頁取得教材

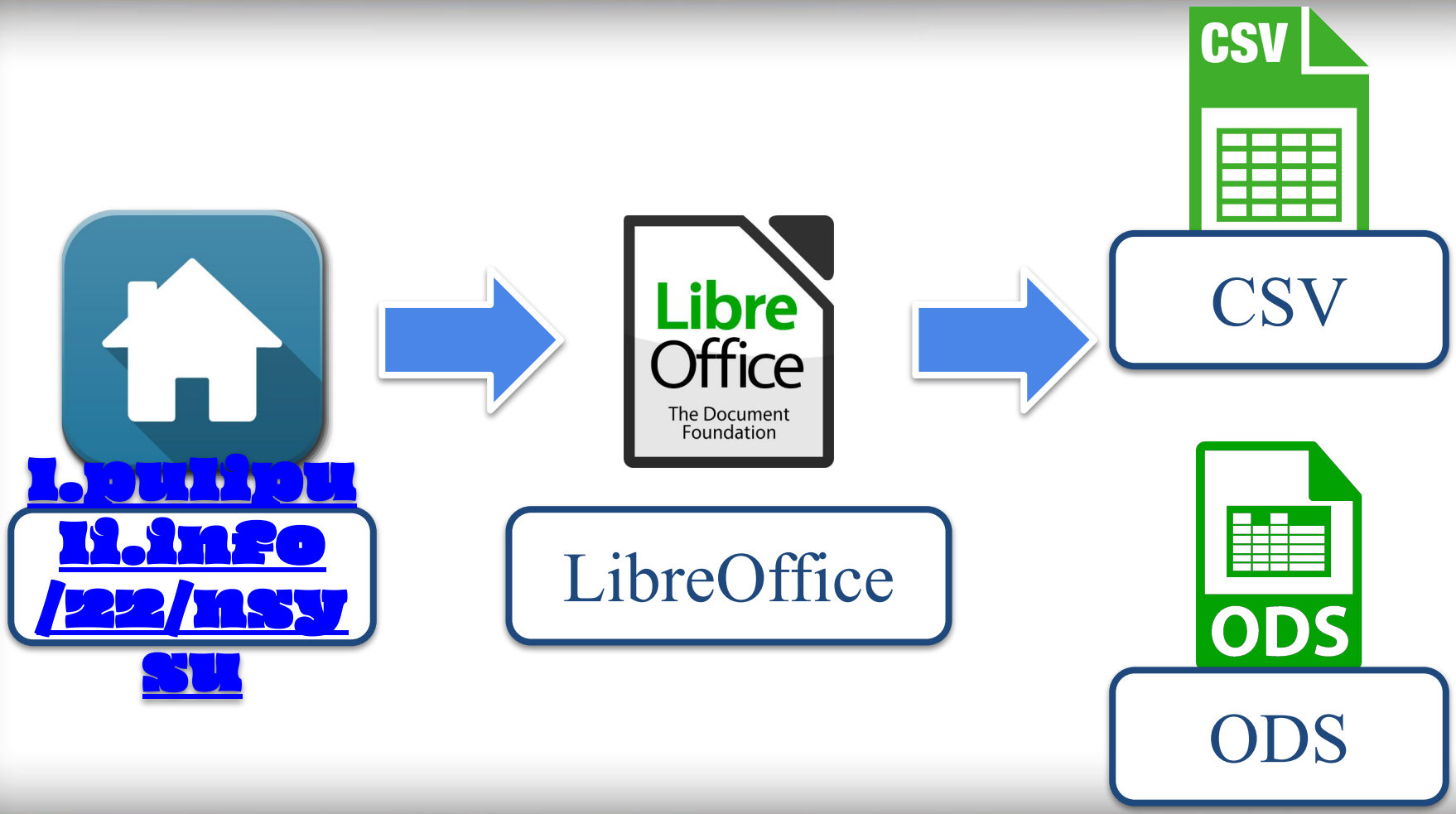

HTML
Sentence Encoder
http://l.pui.ipui.info/22/nsysu
ODS
問水果店還是寵物店 - 問答形式

辦公室套裝的自由軟體
LibreOffice

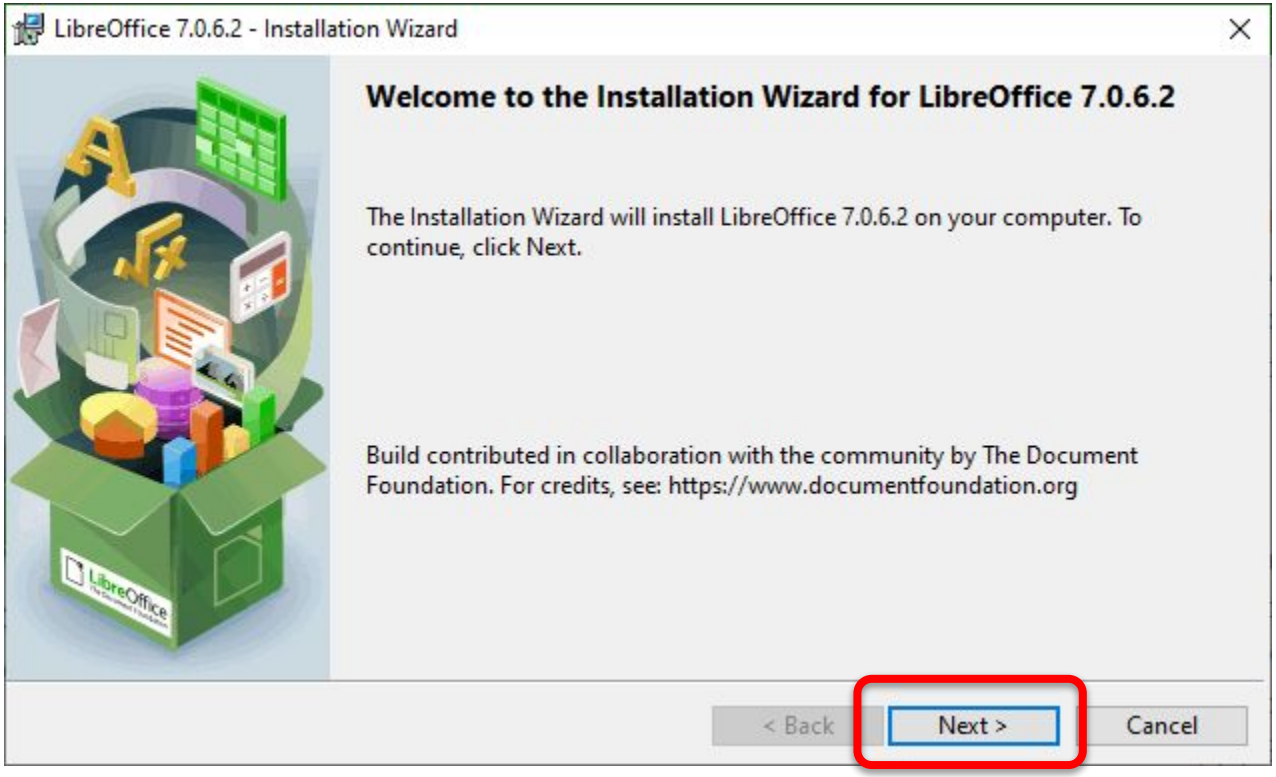
- LibreOffice辦公室套裝軟體的試算表工具
- LibreOffice是跨平臺的開放自由軟體，是編輯開放文件格式(ODF)的最佳選擇
- 開放文件格式包含文件(ODT)、試算表(ODS)、投影片(ODP)等多種類型格式
- 開放文件格式是我國政府的主要通用格式

https://zh-tw.libreoffice.org/download/libreoffice-fresh/

LibreOffice下載



LibreOffice安裝



觀察資料集

question	answer
你知道怎麼分辨百香果和牛奶果嗎？	問水果店
請問現在是椪柑的產季嗎？	問水果店
屏東的芒果有比玉井的好吃嗎？	問水果店
這家店有在賣芒果嗎？	?
請問要去哪裡買狗骨頭？	問寵物店
貓跳台可以寄送嗎？	問寵物店
寵物用的廁所要怎麼選比較好？	問寵物店
德國牧羊犬吃的東西要去哪裡買？	?

觀察資料集

文本

分類

question	answer
你知道怎麼分辨百香果和牛奶果嗎？	問水果店
請問現在是椪柑的產季嗎？	問水果店
屏東的芒果有比玉井的好吃嗎？	問水果店
這家店有在賣芒果嗎？	？
請問要去哪裡買狗骨頭？	問寵物店
貓跳台可以寄送嗎？	問寵物店
寵物用的廁所要怎麼選比較好？	問寵物店
德國牧羊犬吃的東西要去哪裡買？	？

已知案例

未知案例

Sentence Encoder操作

Tokenization ⇨ 詞袋模型

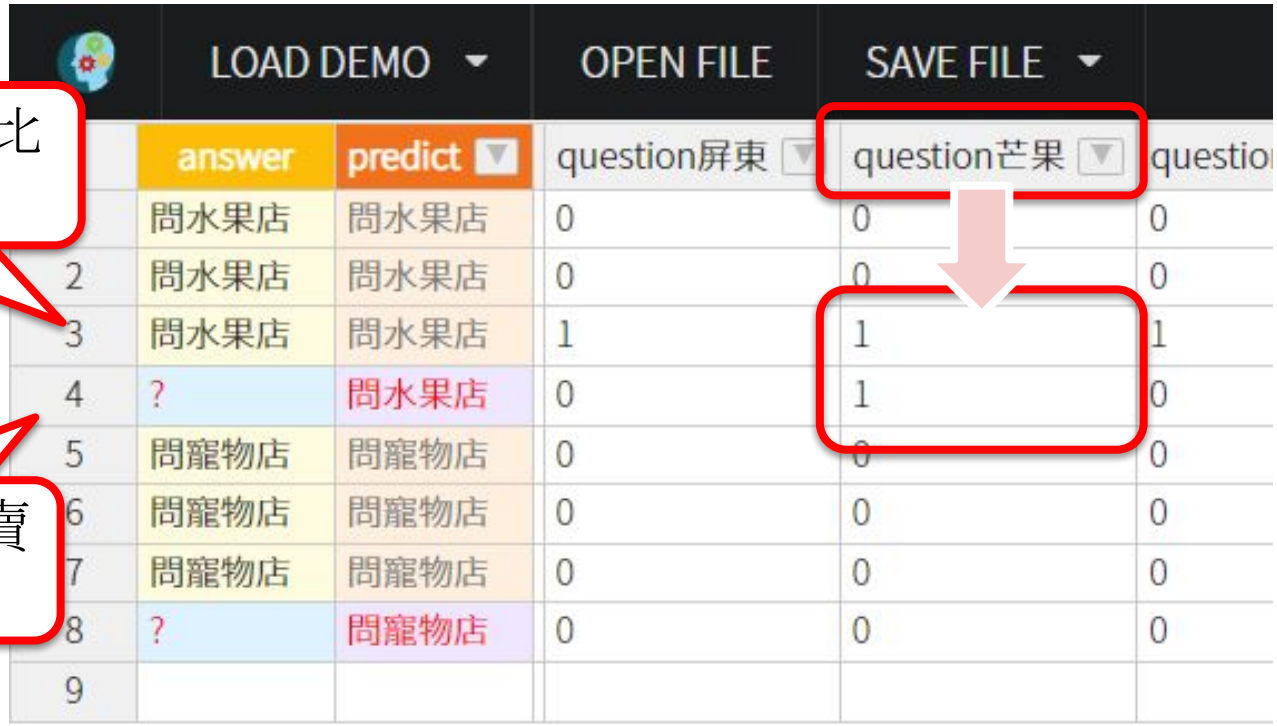
LOAD DEMO ▾ OPEN FILE SAVE FILE ▾

	answer	predict	question屏東	question芒果	question玉井	question好吃	question這
1	問水果店	問水果店	0	0	0	0	0
2	問水果店	問水果店	0	0	0	0	0
3	問水果店	問水果店	1	1	1	1	0
4	?	問水果店	0	1	0	0	1
5	問寵物店	問寵物店	0	0	0	0	0
6	問寵物店	問寵物店	0	0	0	0	0
7	問寵物店	問寵物店	0	0	0	0	0
8	?	問寵物店	0	0	0	0	0
9							

Tokenization ⇨ 詞袋模型

3. 屏東的芒果有比玉井的好吃嗎？

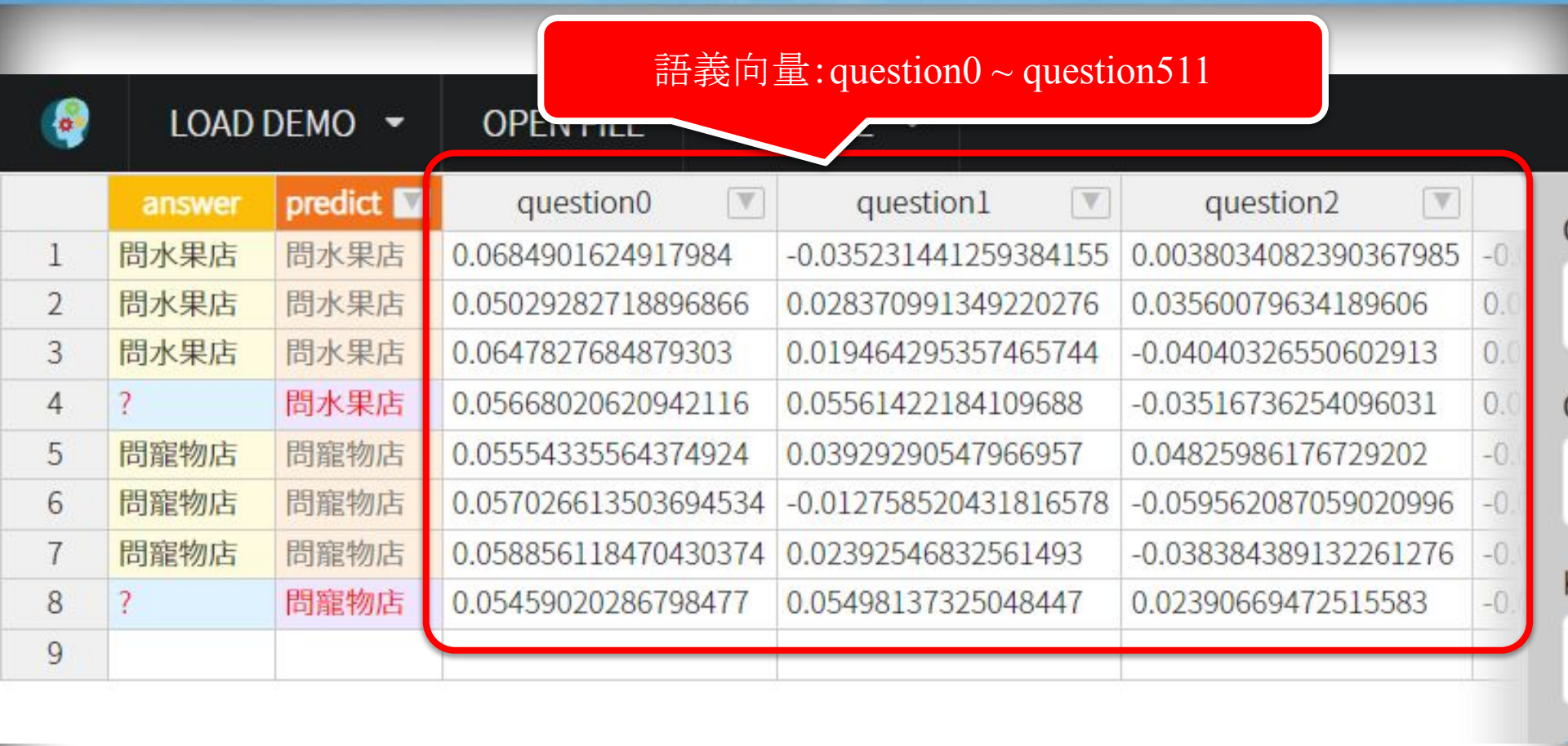
4. 這家店有在賣芒果嗎？



LOAD DEMO | OPEN FILE | SAVE FILE

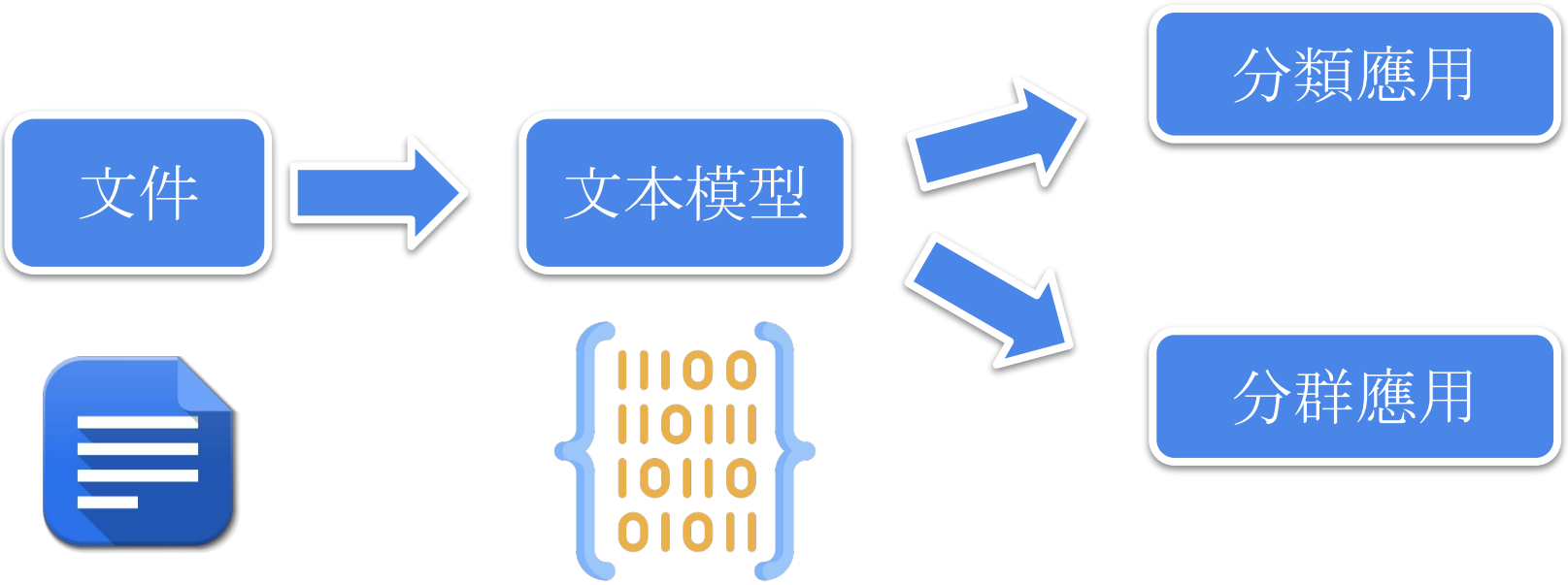
	answer	predict	question屏東	question芒果	questio
1	問水果店	問水果店	0	0	0
2	問水果店	問水果店	0	0	0
3	問水果店	問水果店	1	1	1
4	?	問水果店	0	1	0
5	問寵物店	問寵物店	0	0	0
6	問寵物店	問寵物店	0	0	0
7	問寵物店	問寵物店	0	0	0
8	?	問寵物店	0	0	0
9					

Embedding ⇨ 語義向量



	answer	predict	question0	question1	question2	
1	問水果店	問水果店	0.0684901624917984	-0.035231441259384155	0.0038034082390367985	-0.
2	問水果店	問水果店	0.05029282718896866	0.028370991349220276	0.03560079634189606	0.
3	問水果店	問水果店	0.0647827684879303	0.019464295357465744	-0.04040326550602913	0.
4	?	問水果店	0.05668020620942116	0.05561422184109688	-0.03516736254096031	0.
5	問寵物店	問寵物店	0.05554335564374924	0.03929290547966957	0.04825986176729202	-0.
6	問寵物店	問寵物店	0.057026613503694534	-0.012758520431816578	-0.059562087059020996	-0.
7	問寵物店	問寵物店	0.058856118470430374	0.02392546832561493	-0.038384389132261276	-0.
8	?	問寵物店	0.05459020286798477	0.05498137325048447	0.02390669472515583	-0.
9						

應用



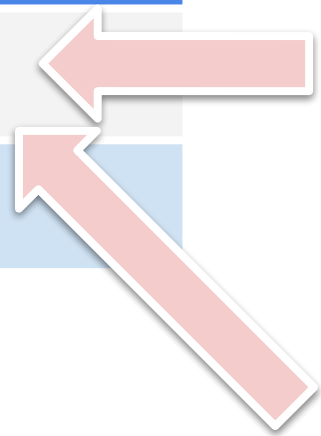
Part 4.

分類應用:問答判斷

未知案例 ⇦ 已知案例

未知案例

question	answer
這家店有在賣芒果嗎？	？
如何挑給德國牧羊犬吃的飼料？	？



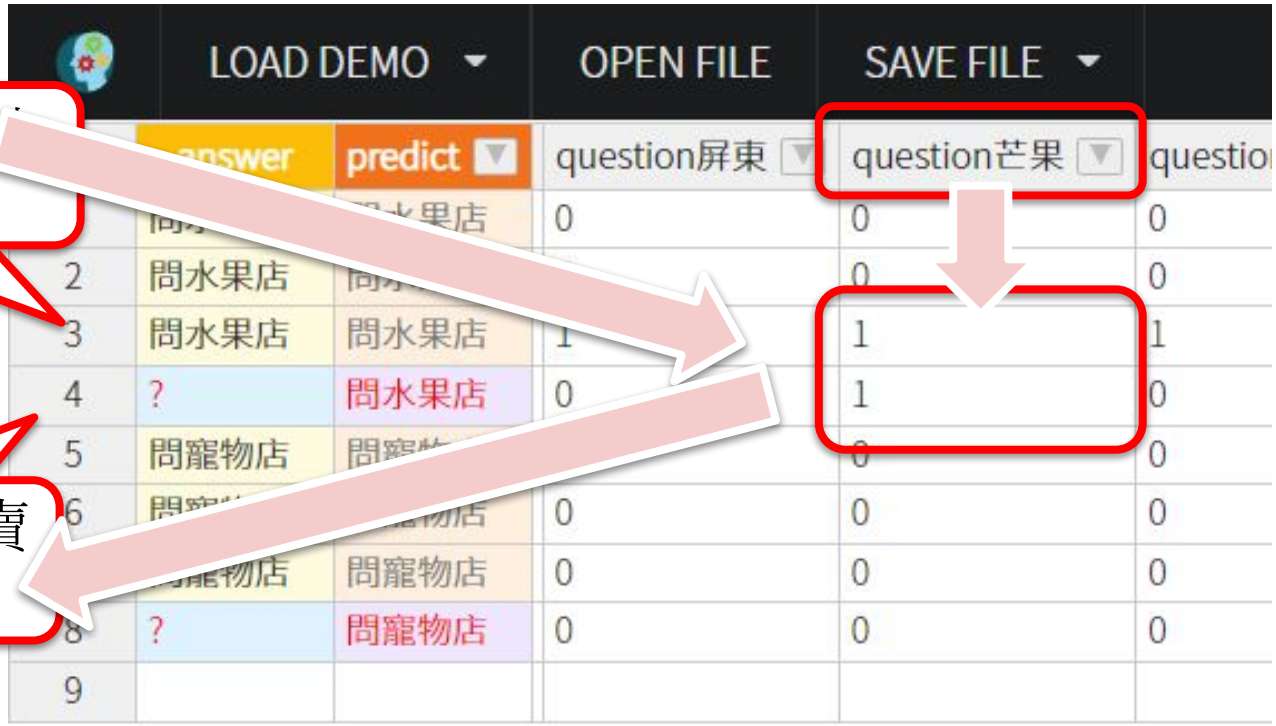
已知案例

question	answer
你知道怎麼分辨百香果和牛奶果嗎？	問水果店
請問現在是椪柑的產季嗎？	問水果店
屏東的芒果有比玉井的好吃嗎？	問水果店
請問要去哪裡買狗骨頭？	問寵物店
貓跳台可以寄送嗎？	問寵物店
寵物用的廁所要怎麼選比較好？	問寵物店

詞袋模型的相近

3. 屏東的芒果有比玉井的好吃嗎？

4. 這家店有在賣芒果嗎？

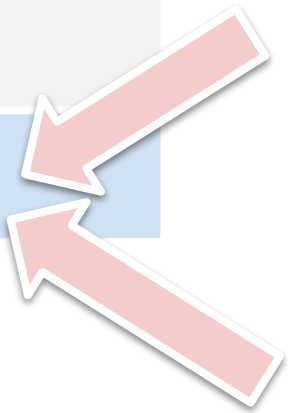


	answer	predict	question屏東	question芒果	questio
		水果店	0	0	0
2	問水果店			0	0
3	問水果店	問水果店	1	1	1
4	?	問水果店	0	1	0
5	問寵物店	問寵		0	0
6	問寵	物店	0	0	0
	寵物店	問寵物店	0	0	0
8	?	問寵物店	0	0	0
9					

未知案例 ⇦ 已知案例

未知案例

question	answer
這家店有在賣芒果嗎？	？
如何挑給德國牧羊犬吃的飼料？	？



已知案例

question	answer
你知道怎麼分辨百香果和牛奶果嗎？	問水果店
請問現在是椪柑的產季嗎？	問水果店
屏東的芒果有比玉井的好吃嗎？	問水果店
請問要去哪裡買狗骨頭？	問寵物店
貓跳台可以寄送嗎？	問寵物店
寵物用的廁所要怎麼選比較好？	問寵物店

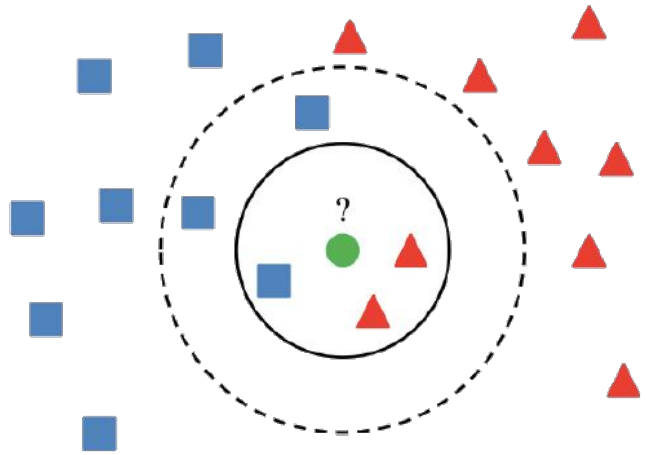
用語義向量來計算相似度

LOAD DEMO ▾ OPEN FILE SAVE FILE ▾

	answer	predict	question0	question1	question2	
1	問水果店	問水果店	0.0684901624917984	-0.035231441259384155	0.0038034082390367985	-0.
2	問水果店	問水果店	0.05029282718896866	0.028370991349220276	0.03560079634189606	0.
3	問水果店	問水果店	0.0647827684879303	0.019464295357465744	-0.04040326550602913	0.
4	?	問水果店	0.05668020620942116	0.05561422184109688	-0.03516736254096031	0.
5	問寵物店	問寵物店	0.05554335564374924	0.03929290547966957	0.04825986176729202	-0.
6	問寵物店	問寵物店	0.057026613503694534	-0.012758520431816578	-0.059562087059020996	-0.
7	問寵物店	問寵物店	0.058856118470430374	0.02392546832561493	-0.038384389132261276	-0.
8	?	問寵物店	0.05459020286798477	0.05498137325048447	0.02390669472515583	-0.
9						

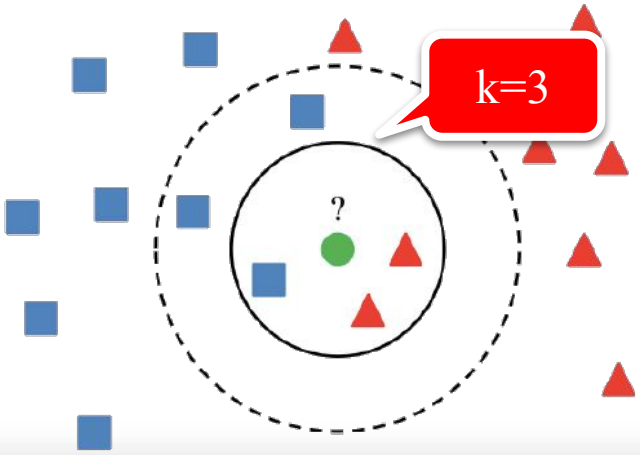
K-Nearest Neighbors (K-NN)

- k-最近鄰分類法是為未知案例找尋**相似**的已知案例，以此決定未知案例的分類
- 相似度的計算常見使用歐幾里得距離
- 廣泛應用於圖形辨識領域



KNN步驟

1. 選擇用來判斷分類的候選數量K (最小值為1)
2. 對於未知案例, 選出相似度最高的K個候選已知案例
3. 統計K個候選已知案例的分類數量
4. 將候選已知案例數量最多的分類, 指派給未知案例

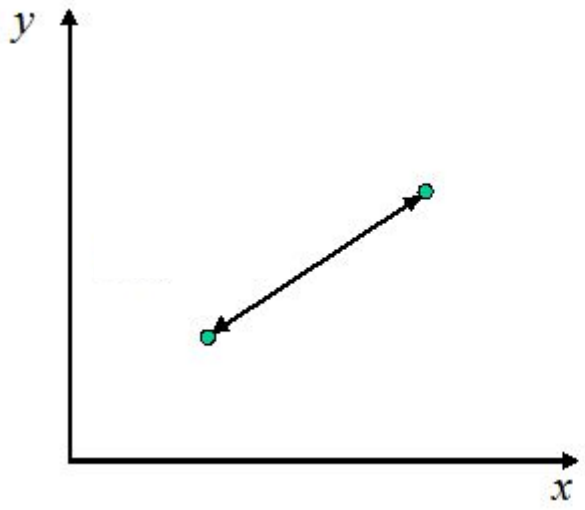


https://www.slideshare.net/akhileshjoshi123/knn-classification?qid=139ec64f-3f01-467b-8f7c-968bc8e4a70e&v=&b=&from_search=5

相似度的計算

歐基里得距離 (Euclidean distance)

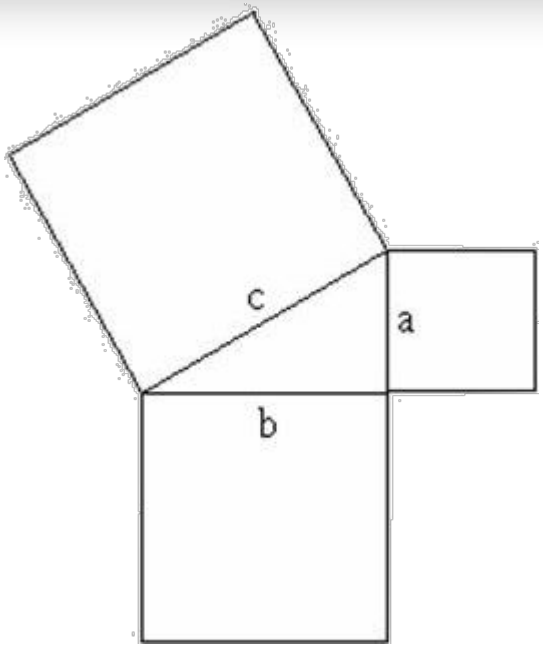
資料點$x_i=<x_{i1},x_{i2},\ldots,x_{ik}>$和資料點$x_j=<x_{j1},x_{j2},\ldots,x_{jk}>$之間的歐基里得距離算法

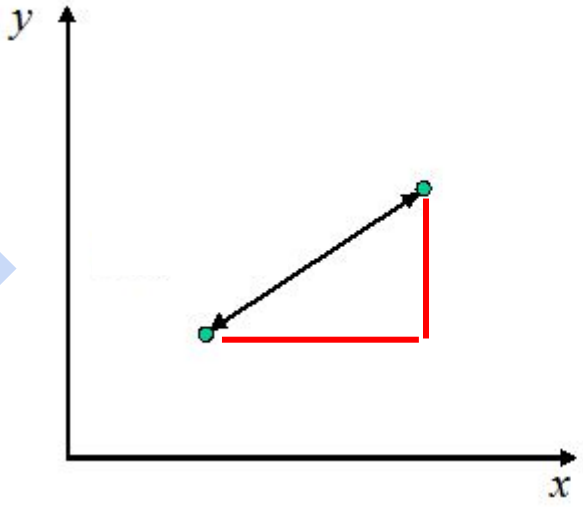


$$d_2(x_i, x_j) = \left(\sum_{d=1}^{k} |x_{id} - x_{jd}|^2\right)^{1/2}$$

$$= \|x_i - x_j\|_2 \left(\|x_i - x_j\|\right)$$

畢氏定理⇨歐基里得距離

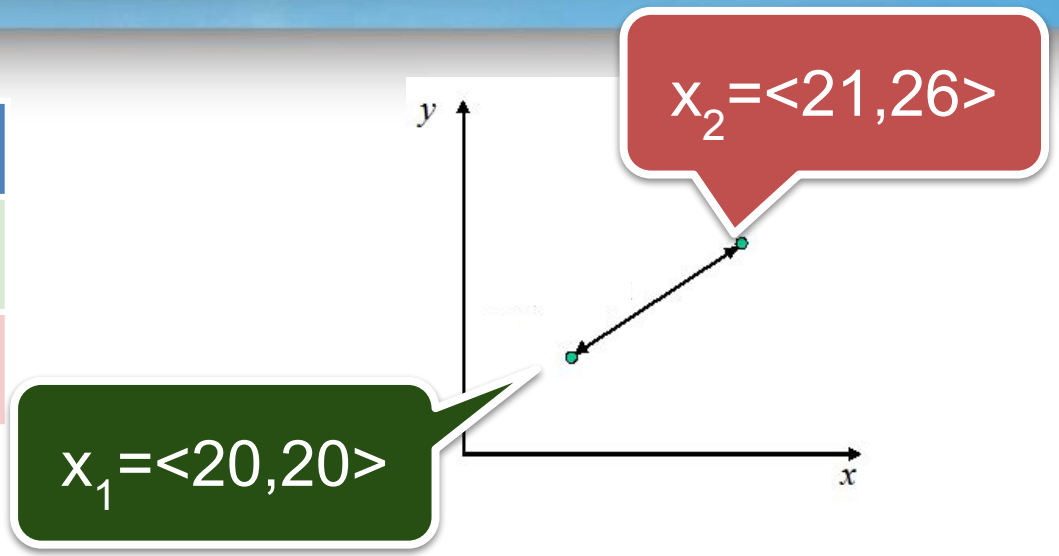




$c^2=a^2+b^2$

歐基里得距離計算

句子	維度1	維度2
1	20	20
2	21	26



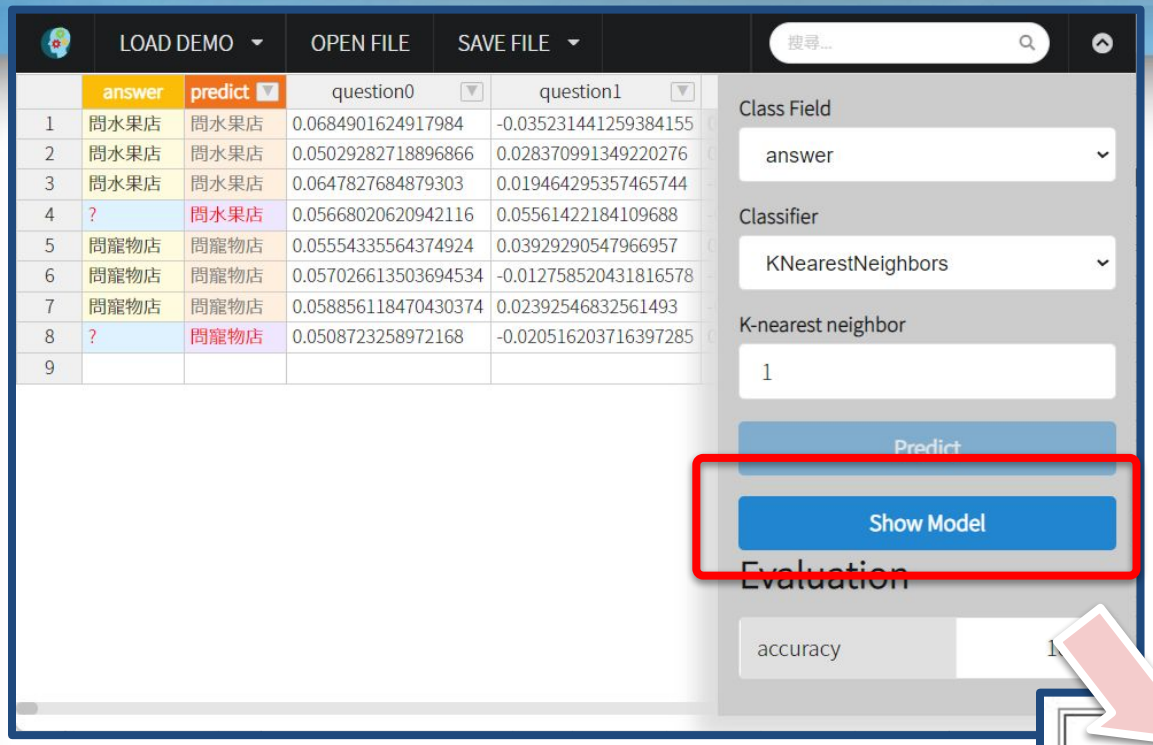
句子x_1=[20, 20] 與

句子x_2=[21, 26]

之間的歐基里得距離為：

$$d_2(x_1, x_2) = \sqrt{(21-20)^2 + (26-20)^2} \approx 6$$

相似度計算結果



	answer	predict	question0	question1
1	問水果店	問水果店	0.0684901624917984	-0.035231441259384155
2	問水果店	問水果店	0.05029282718896866	0.028370991349220276
3	問水果店	問水果店	0.0647827684879303	0.019464295357465744
4	?	問水果店	0.0566802062094211 6	0.05561422184109688
5	問寵物店	問寵物店	0.05554335564374924	0.03929290547966957
6	問寵物店	問寵物店	0.057026613503694534	-0.012758520431816578
7	問寵物店	問寵物店	0.058856118470430374	0.02392546832561493
8	?	問寵物店	0.0508723258972168	-0.020516203716397285
9				

歐幾里得距離

正規化

- 距離0 : 100%
- 距離最遠 : 0%

Unknowns		Neighbors 1	2	3	5	6	7
4	問水果店	22%	9%	30%	10%	0%	3%
8	問寵物店	10%	0%	5%	33%	15%	29%

與未知案例相似的已知案例

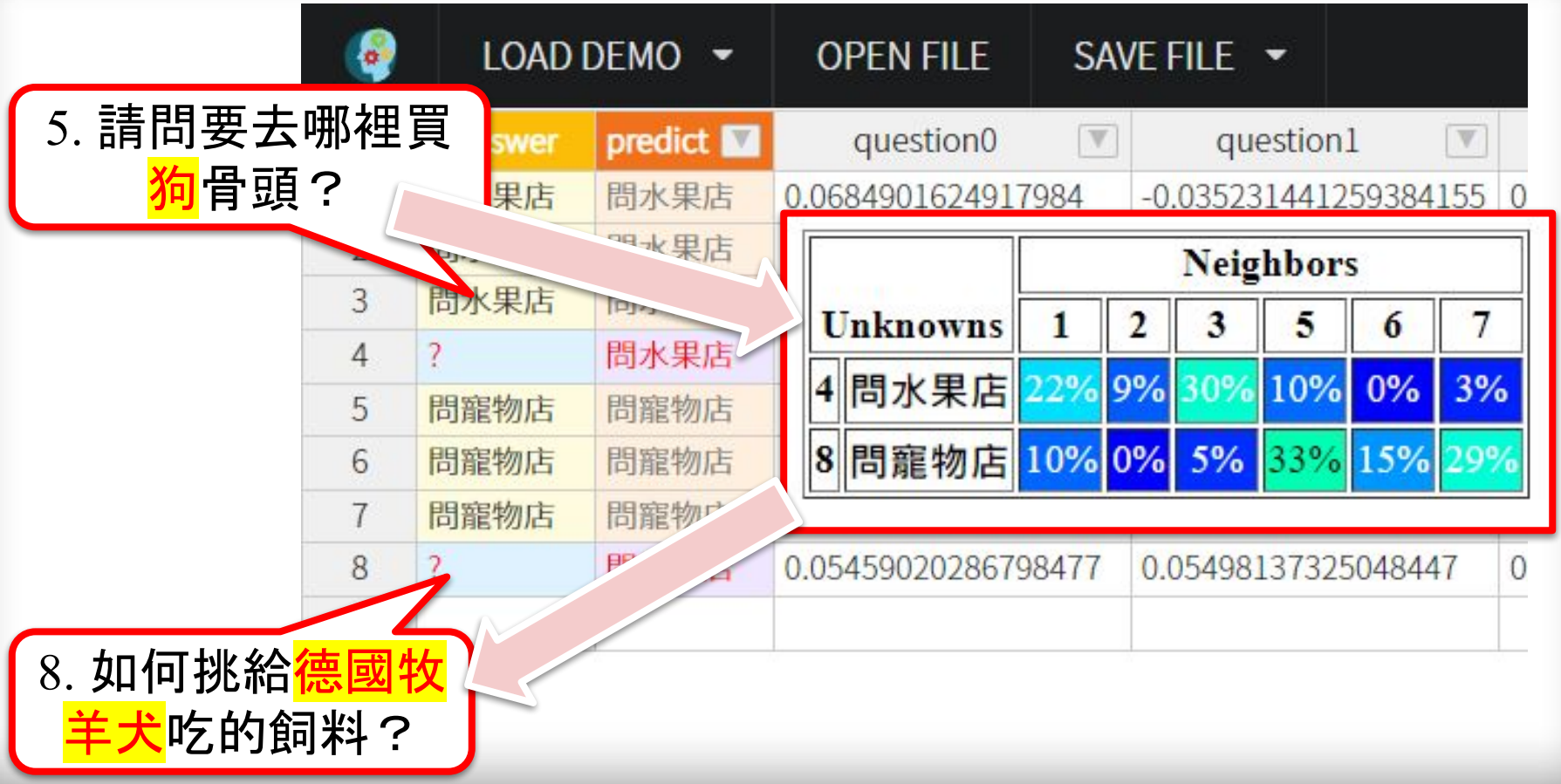
詞袋模型

Unknowns		Neighbors					
		1	2	3	5	6	7
4	問水果店	0%	0%	23%	5%	16%	11%
8	問寵物店	0%	0%	10%	5%	15%	10%

語義向量

Unknowns		Neighbors					
		1	2	3	5	6	7
4	問水果店	22%	9%	30%	10%	0%	3%
8	問寵物店	10%	0%	5%	33%	15%	29%

語義向量的相近


LOAD DEMO
OPEN FILE
SAVE FILE
5. 請問要去哪裡買狗骨頭？
8. 如何挑給德國牧羊犬吃的飼料？
swer
predict
question0
question1
果店
問水果店
0.0684901624917984
-0.035231441259384155
0
3
問水果店
4
?
問水果店
5
問寵物店
問寵物店
6
問寵物店
問寵物店
7
問寵物店
8
?
0.05459020286798477
0.05498137325048447
0
Neighbors
Unknowns
1
2
3
5
6
7
4 問水果店
22%
9%
30%
10%
0%
3%
8 問寵物店
10%
0%
5%
33%
15%
29%

今天來點兔子吧！

挑戰時刻來臨了

換你設計問答資料集！

http://l.pulipuli.info/22/nsysu

設計問答資料集

1. 建立問答資料集
   a. 圖書館常見問題
   b. 蝦皮協助中心
   c. 104常見問題
2. 產生語義向量的文本模型
3. 測試語義向量+KNN能不能正確分類問題的答案

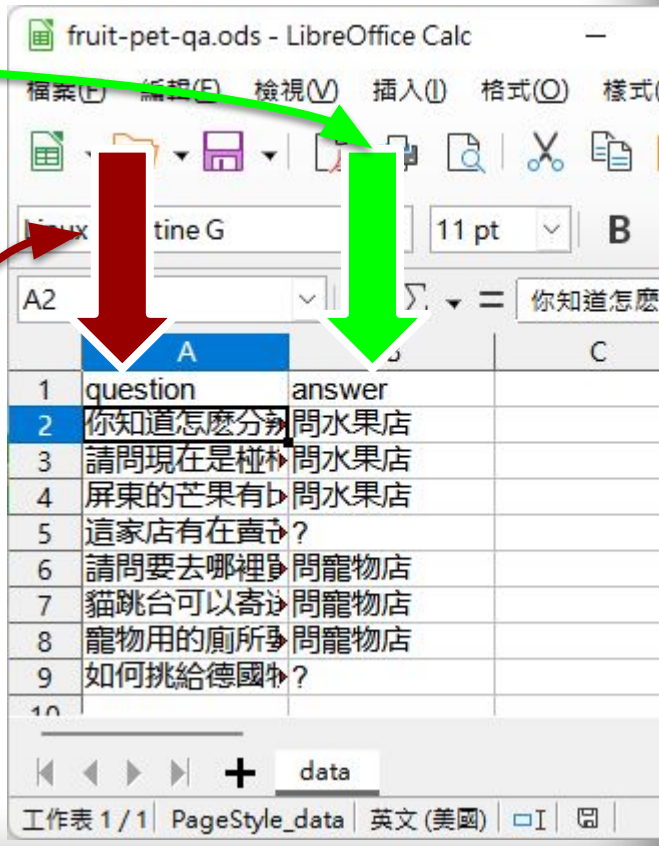
編輯問答資料集 (1/3)

- http://1.pull.pull.info/22/nsysu

問水果店還是寵物店 - 問答形式

編輯問答資料集 (2/3)

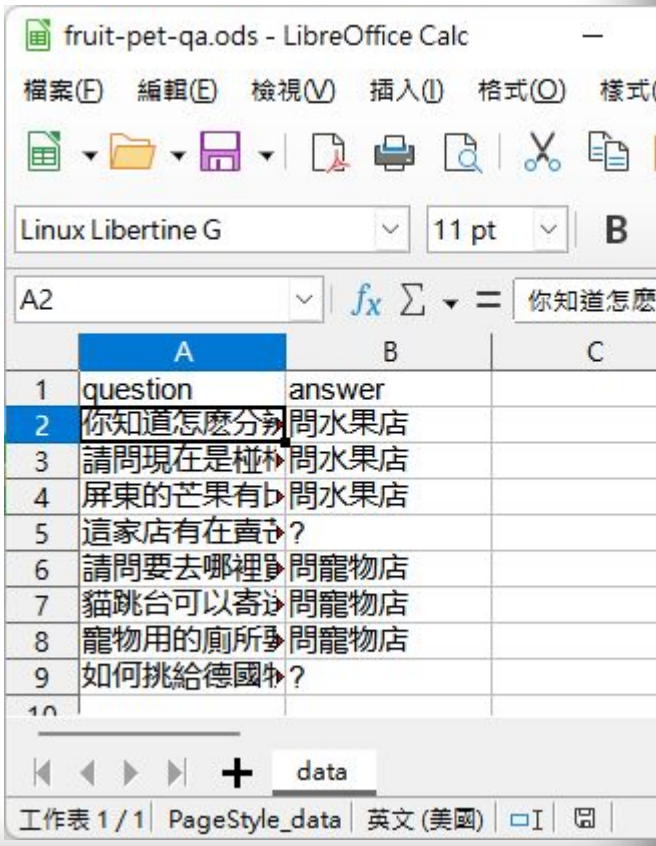

分類
帳號安全與其他
訂單與付款
退貨退款
賣家相關
寄件物流
蝦皮商城
文章
[退貨交寄] 申請退貨後，請問商品怎麼退回？
[訂單取消] 當付款後取消訂單/賣家同意退貨款
[退貨交寄] 買家如何確認退貨的配送進度？
【蝦皮安心退】退貨退款懶人包


fruit-pet-qa.ods - LibreOffice Calc
檔案(F) 編輯(E) 檢視(V) 插入(I) 格式(O)
11 pt
A2
question
answer
問水果店
問寵物店
data
工作表 1 / 1
PageStyle_data
英文 (美國)

編輯問答資料集 (3/3)

問答資料集建議最小條件:

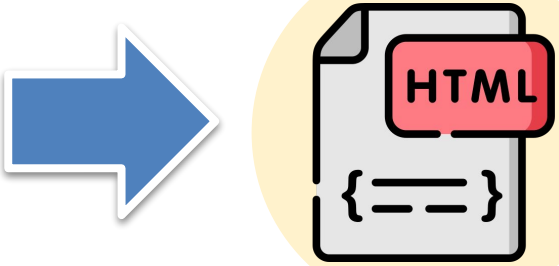
- 2個分類, 各4個問題, 共8個問答
- 每個分類個別將其中一個問題的答案改為「?」, 表示「未知案例」



測試看看語義向量+KNN演算法能不能正確判斷問題的分類答案

測試問答資料集

http://1.pull.pull.info/22/nsysu



Sentence Encoder

LOAD DEMO

	answer	predict	
1	問水果店	問水果	068
2	問水果店	問水果	050
3	問水果店	問水果店	0.064
4	?	問水果店	0.056
5	問寵物店	問寵物店	0.055
	問寵物店	問寵物店	0.057
	問寵物店	問寵物店	0.058
	?	問寵物店	0.054

學習單:設計問答資料集

http://1.pui
ipui1.info
/22/nsysu

學習單:設計問
答資料集

設計問答資料集

換你囉

設計問答資料集 - 中文自然語言處理動手玩

演講內容請見 http://l.pulipuli.info/22/nsysu

請使用LibreOffice Calc編輯:
下載網址: https://zh-tw.libreoffice.org/download/libreoffice-fresh/

問答資料集範本:
https://docs.google.com/spreadsheets/d/1eHKhjn3Wv48NlRfj0n1DJCmI43XtIj9Zqkt1mWu1gFg/export?format=ods

問答資料集參考資料來源
- 圖書館常見問題 https://lis.nsysu.edu.tw/p/412-1001-17831.php
- 蝦皮協助中心 https://help.shopee.tw/tw/s/
- 104常見問題 https://www.104.com.tw/faq/match-letter

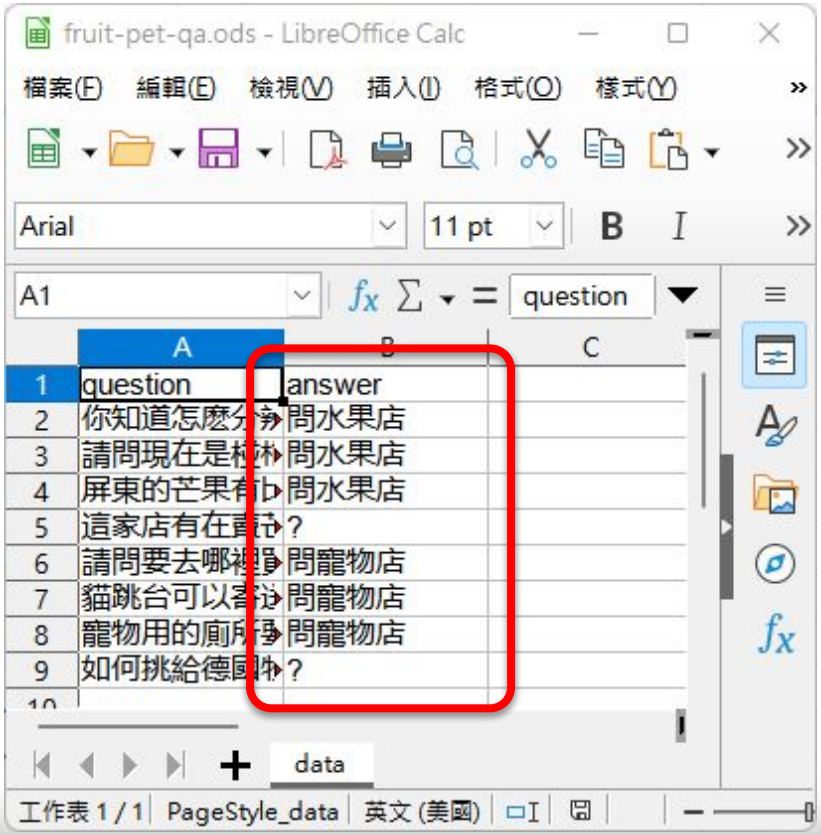
測試你的問答資料集:
https://pulipulichen.github.io/HTML5-Sentence-Encoder/

Part 5.

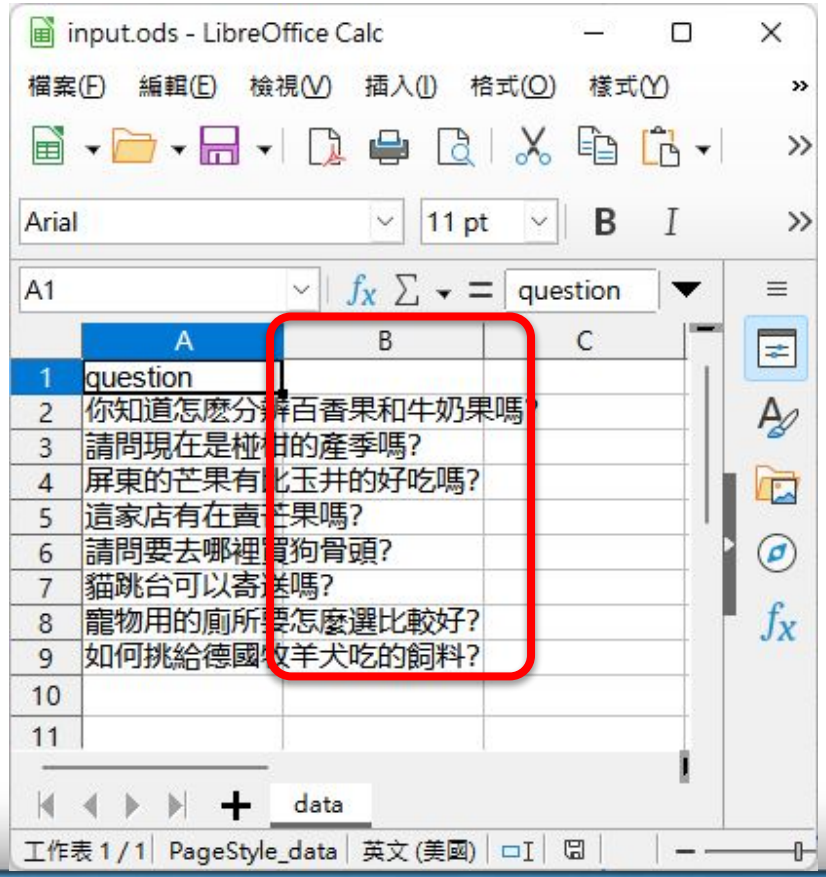
分群應用

未知分類的文本

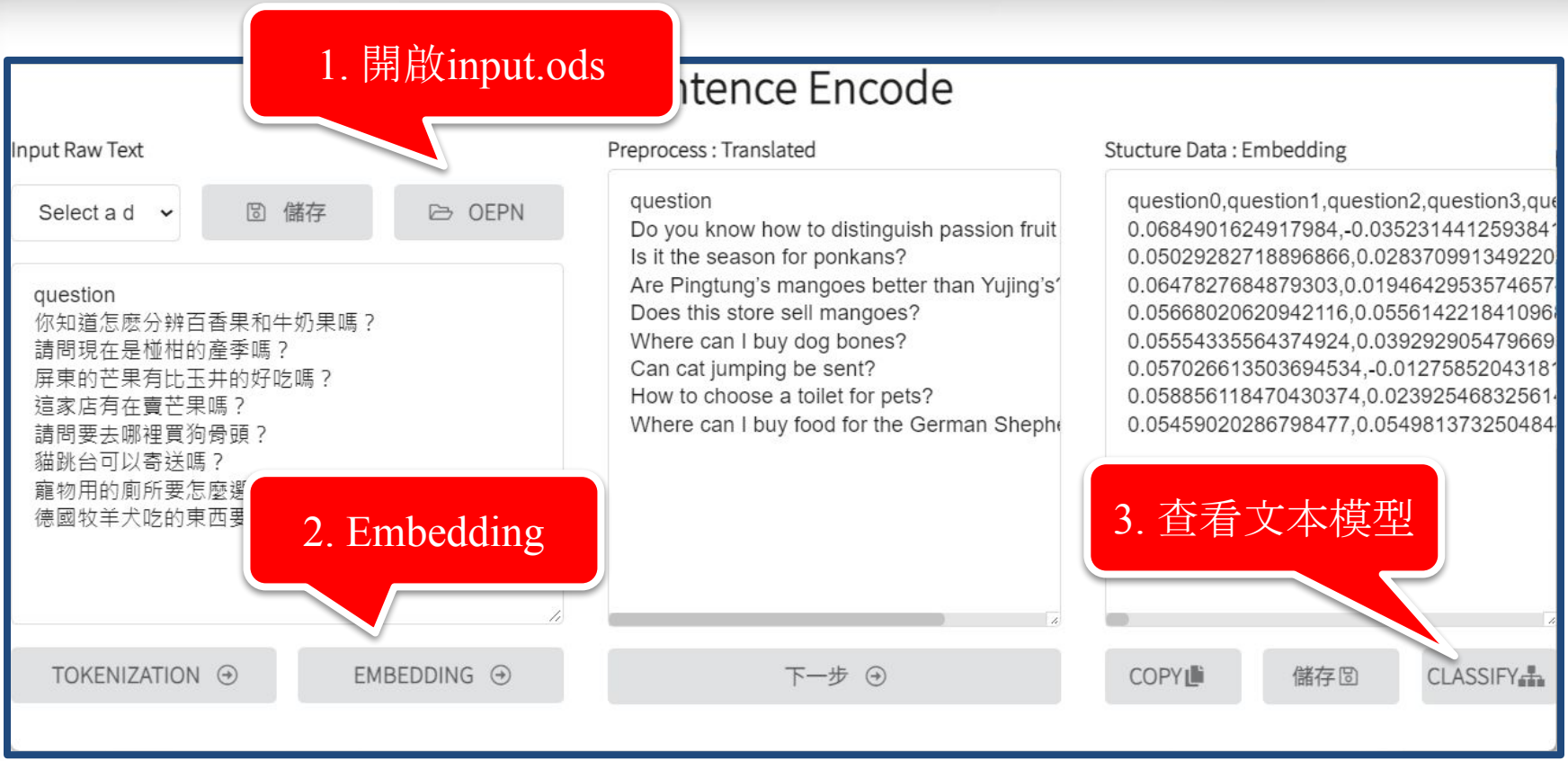
有分類欄位⇨分類



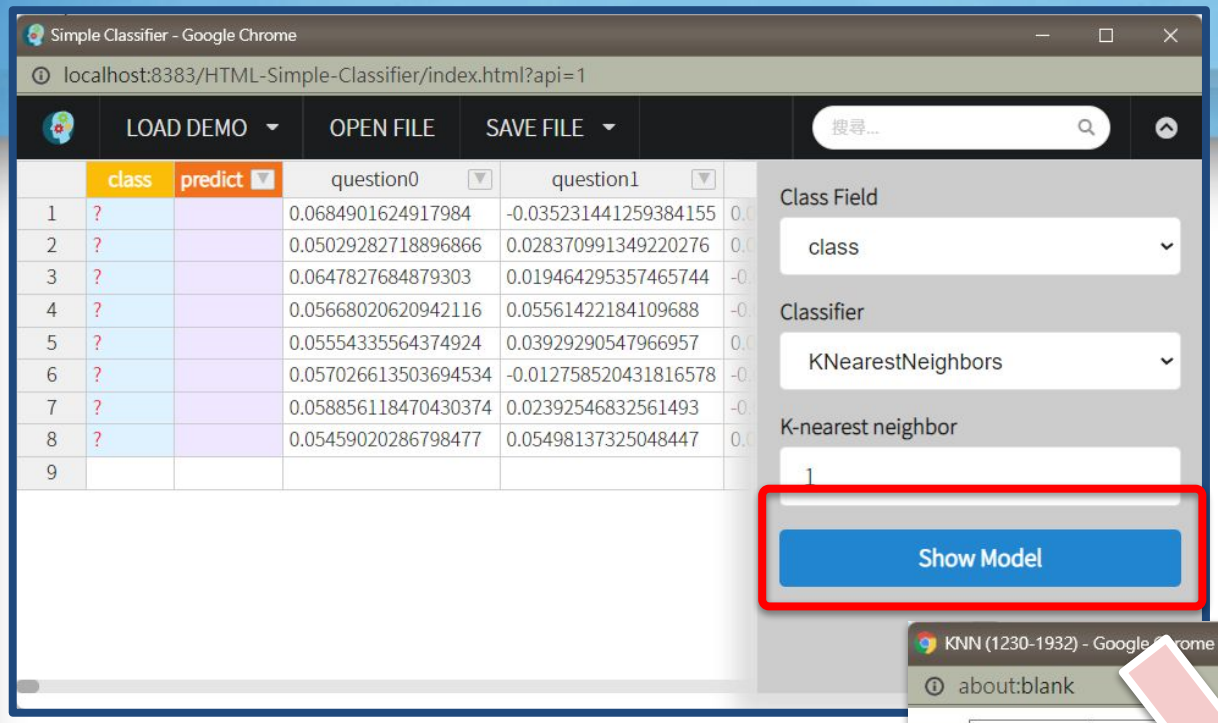
沒有分類欄位⇨分群

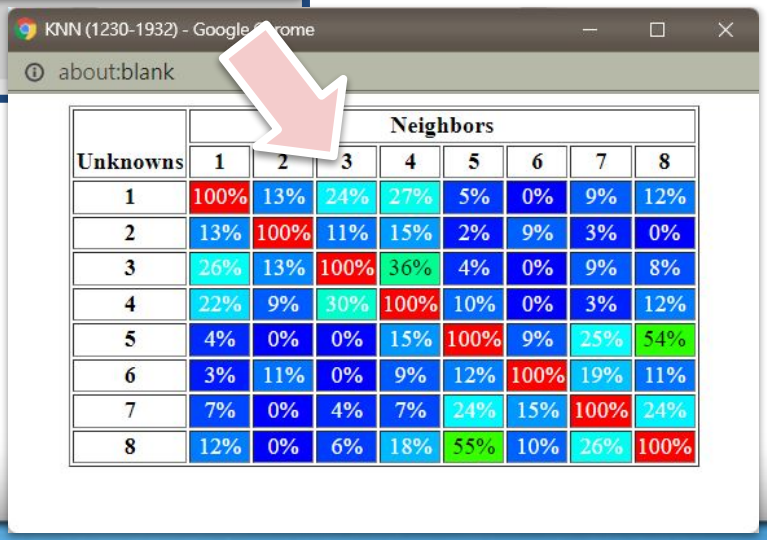


觀察文本相似度矩陣 (1/3)



觀察文本相似度矩陣 (2/3)





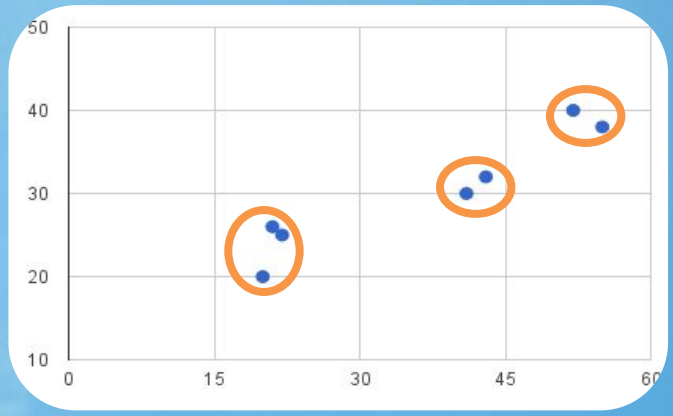
Unknowns	Neighbors 1	2	3	4	5	6	7	8
1	100%	13%	24%	27%	5%	0%	9%	12%
2	13%	100%	11%	15%	2%	9%	3%	0%
3	26%	13%	100%	36%	4%	0%	9%	8%
4	22%	9%	30%	100%	10%	0%	3%	12%
5	4%	0%	0%	15%	100%	9%	25%	54%
6	3%	11%	0%	9%	12%	100%	19%	11%
7	7%	0%	4%	7%	24%	15%	100%	24%
8	12%	0%	6%	18%	55%	10%	26%	100%

觀察文本相似度矩 (3/3)

KNN (1230-1932) - Google Chrome

about:blank

Unknowns	Neighbors							
	1	2	3	4	5	6	7	8
1	100%	13%	24%	27%	5%	0%	9%	12%
2	13%	100%	11%	15%	2%	9%	3%	0%
3	26%	13%	100%	36%	4%	0%	9%	8%
4	22%	9%	30%	100%	10%	0%	3%	12%
5	4%	0%	0%	15%	100%	9%	25%	54%
6	3%	11%	0%	9%	12%	100%	19%	11%
7	7%	0%	4%	7%	24%	15%	100%	24%
8	12%	0%	6%	18%	55%	10%	26%	100%



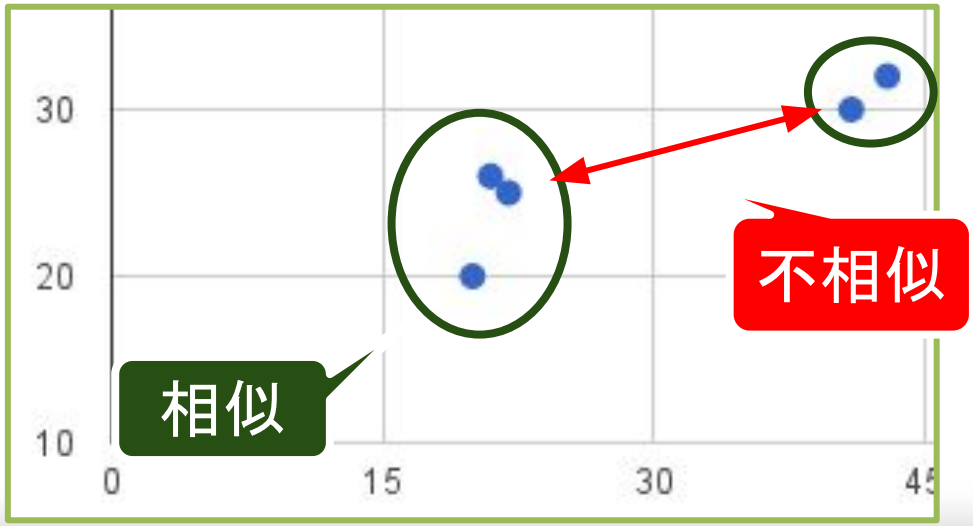
分群演算法

K Means

K平均法

分群的概念與目的

- 將資料集合中的資料記錄(資料點)加以分群成數個群集(cluster)
- 使得每個群集中的資料點間相似程度高於與其它群集中資料點的相似程度
- 從群集結果推論出有用、隱含、令人感興趣的特性和現象



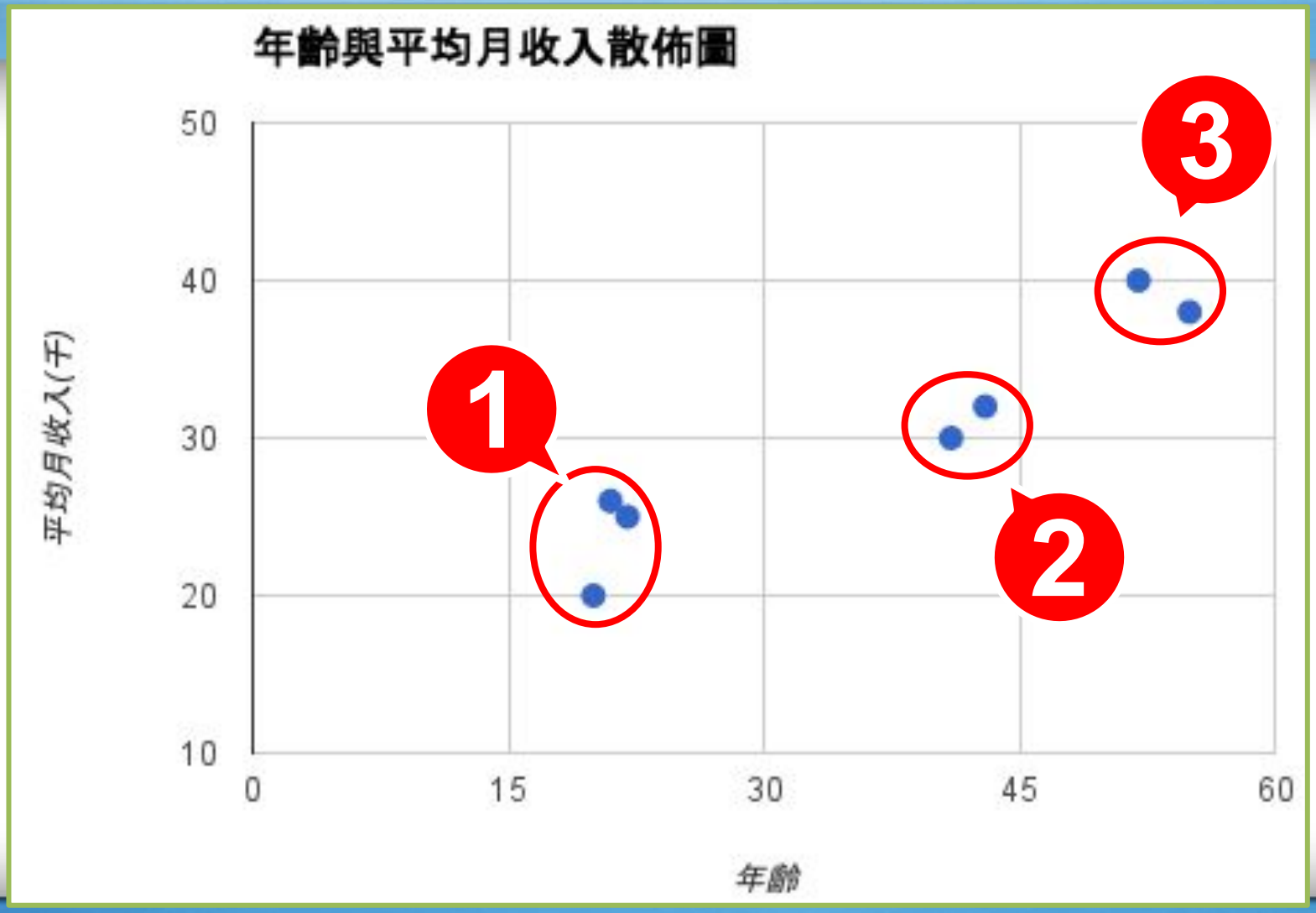
K平均法
演算法流程

輸入：**資料集合、使用者定義之群集數量*k***

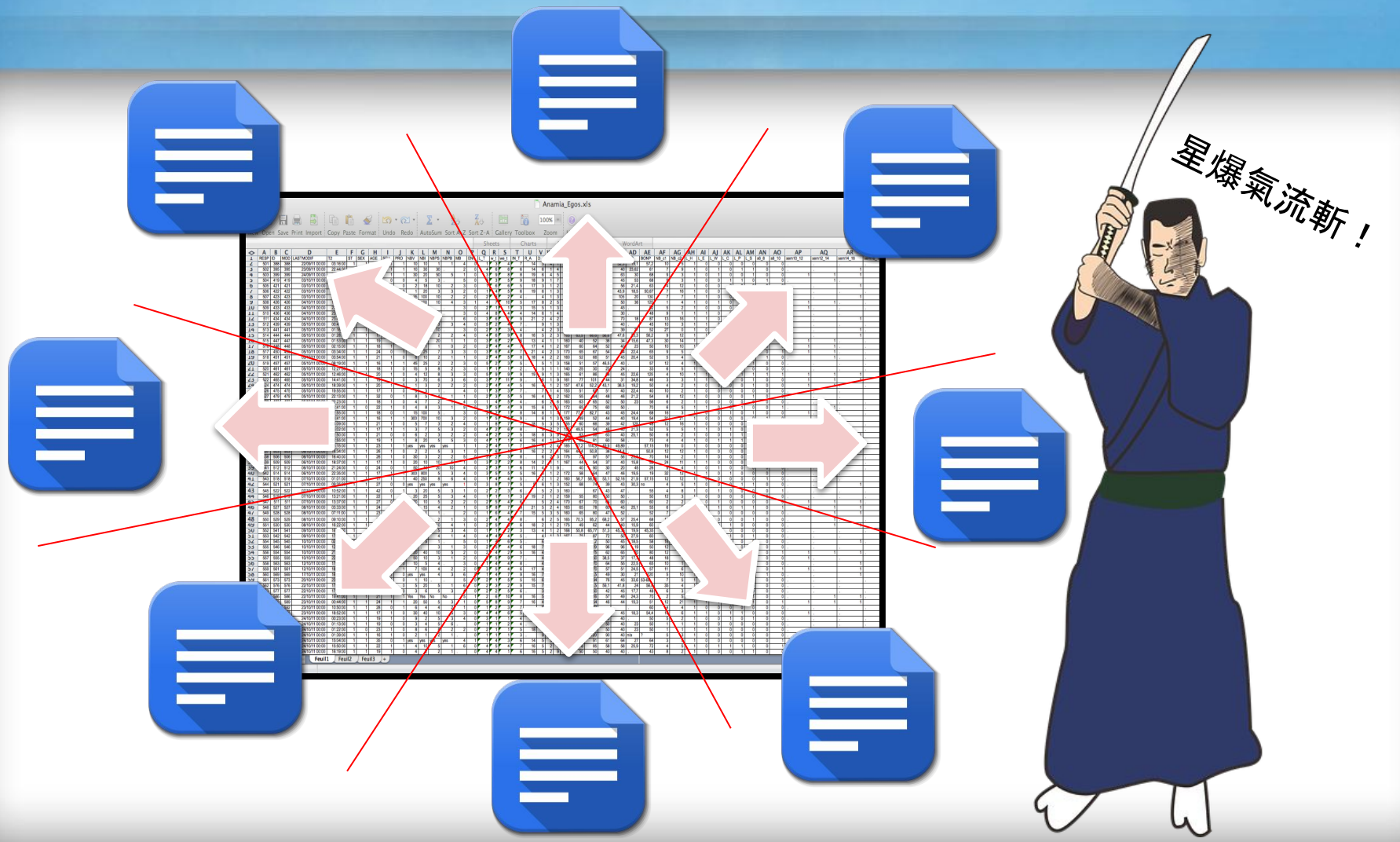
輸出：***k*個互不交集的群集**

1. 隨機從資料集合中選擇任*k*個資料點當作起始*k*群的群集中心
2. 利用相似度計算公式，將資料點分別歸屬到距其最近之群集中心所屬的群集，形成*k*個群集。
3. 利用各群集中所含的資料點，重新計算各群集之群集中心點
4. 條件判斷：
   a. 假如由步驟3所得到各群之群集中心與之前所計算之群集中心相同，則表示分群結果已穩定，並結束此處理程序並輸出各群結果
   b. 否則回到步驟2繼續執行

「k」=3，3個分群

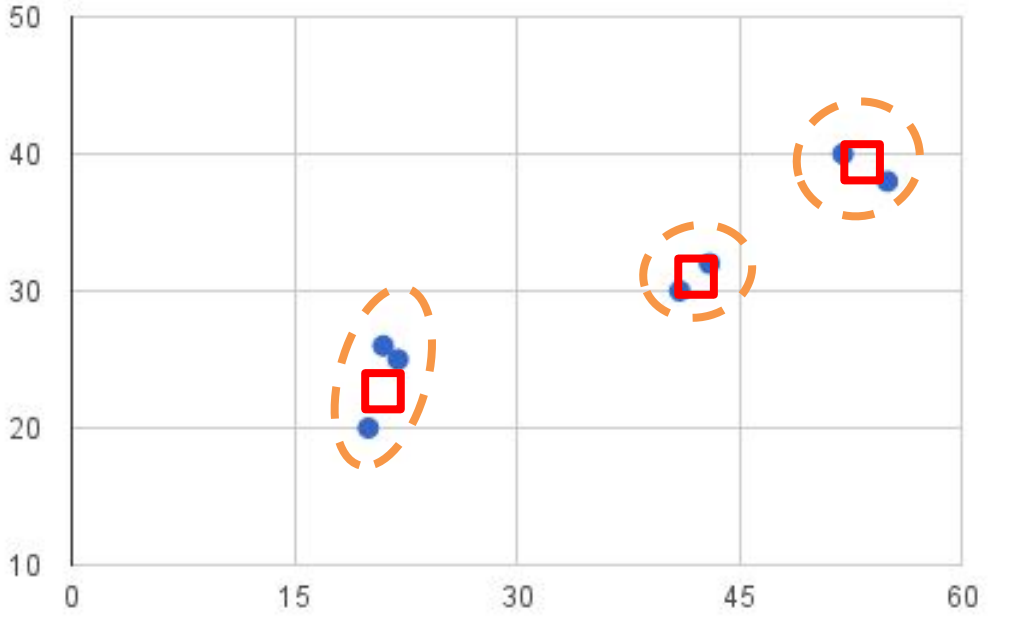


如何選擇分群數量K值？

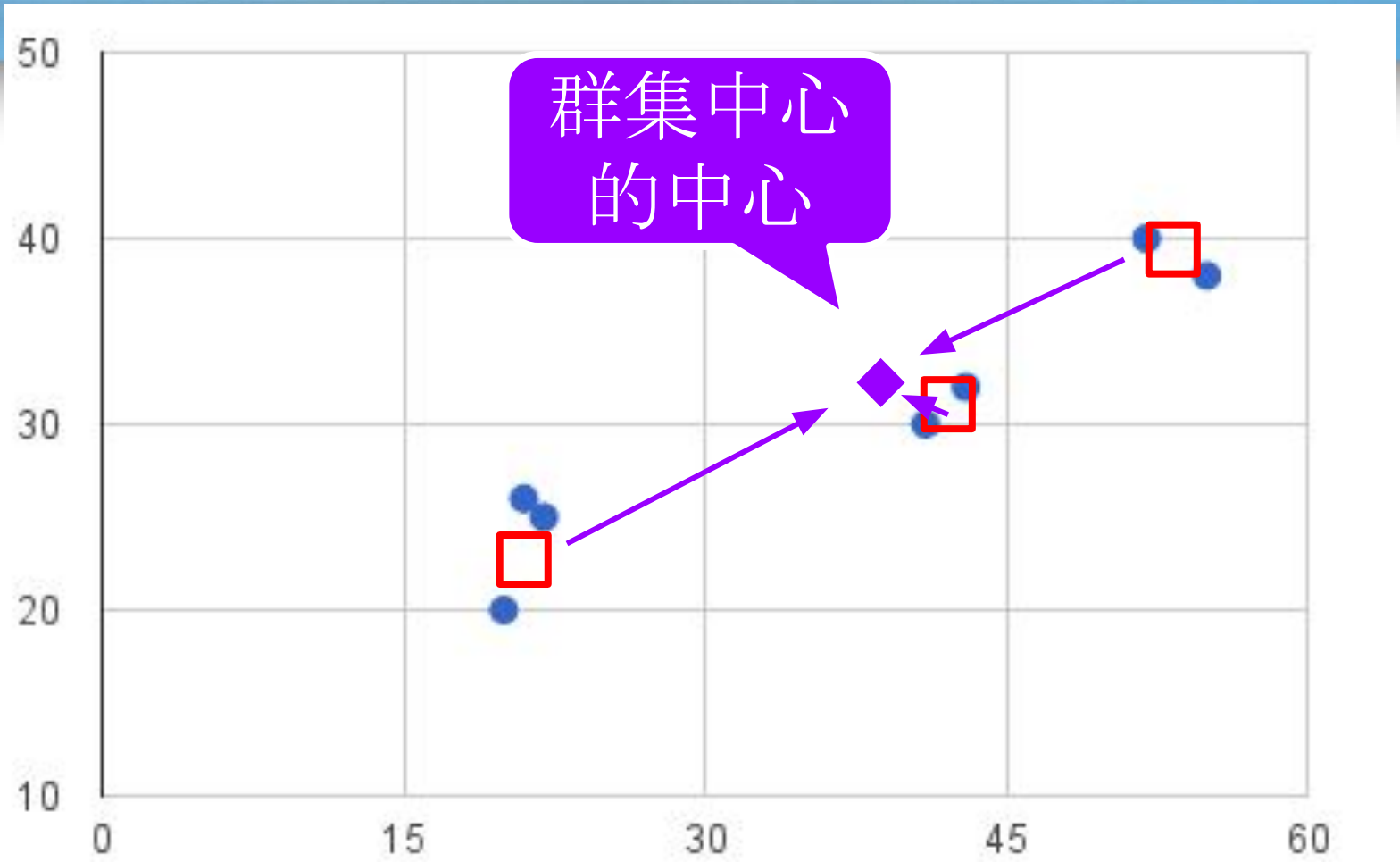


評估分群品質
CH指標 (Calinski-Harabasz)

$$CH(K) = \frac{[\text{trace } \mathbf{B} / K - 1]}{[\text{trace } \mathbf{W} / N - K]} \quad \text{for} \quad K \in \mathbb{N}$$

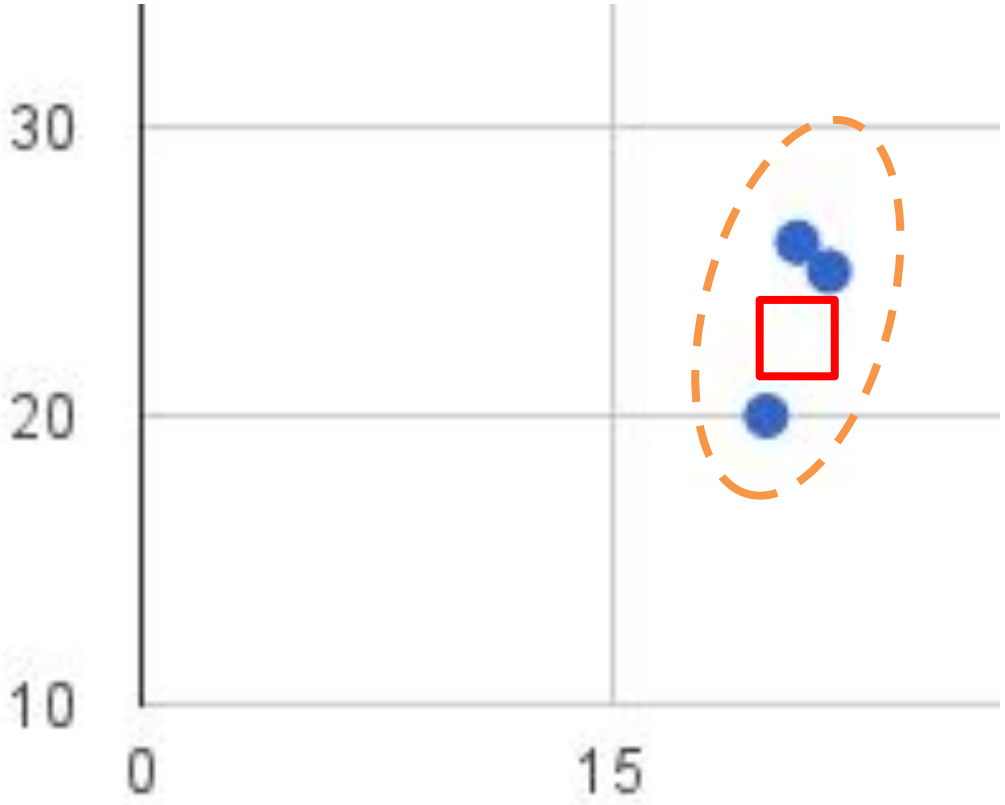


CH指標
trace B：各群之間的距離 (越大越好)

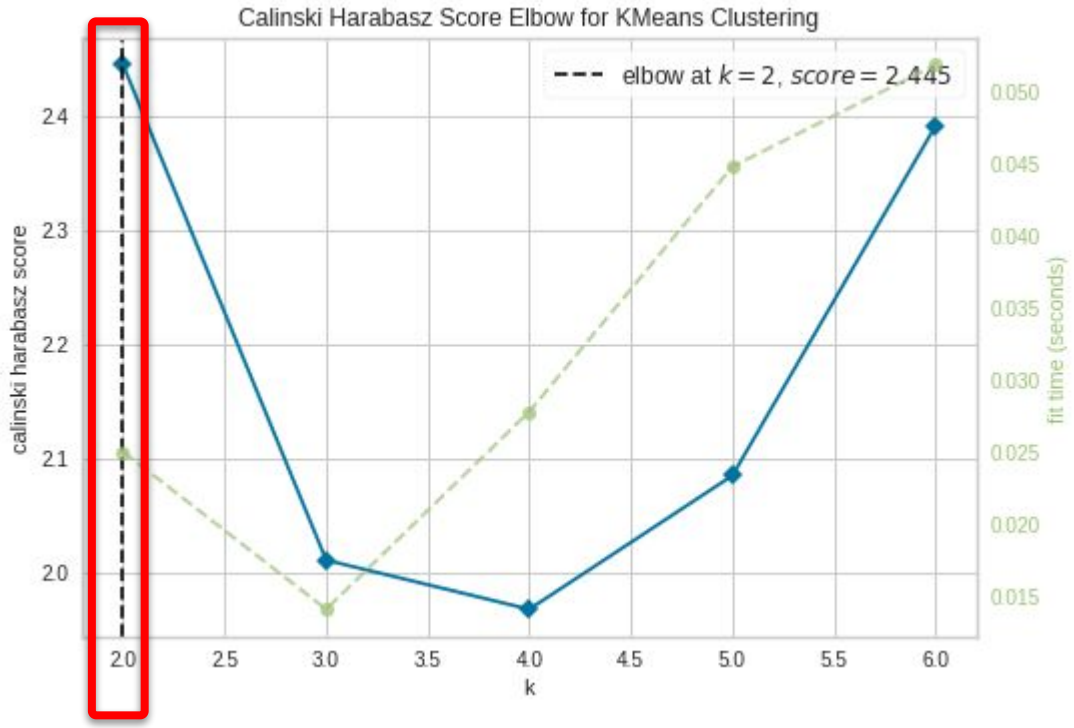


CH指標

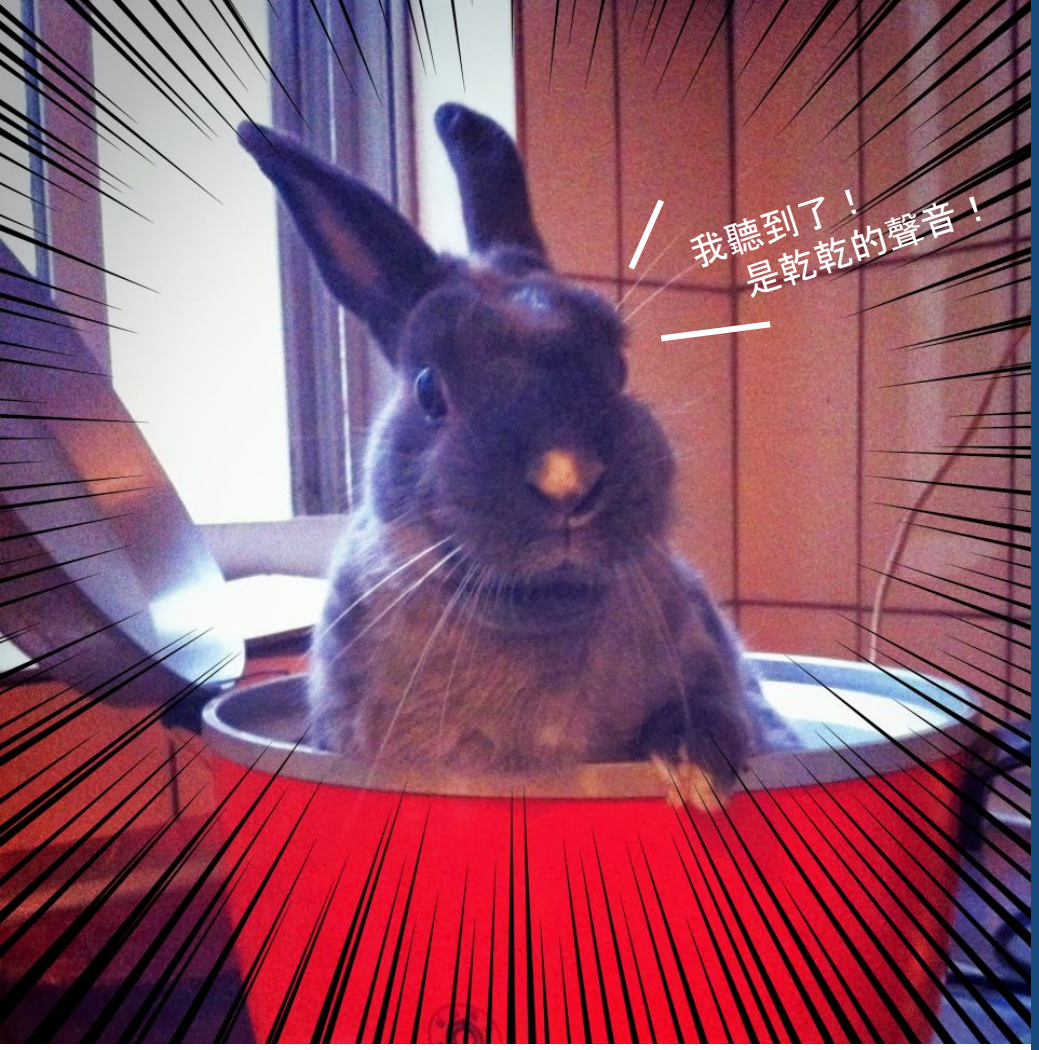
trace W:群內各點的距離 (越小越好)



搭配Elbow Method決定最佳K值



https://www.scikit-yb.org/en/latest/api/cluster/elbow.html



食戟之兔 貳之鍋

勇者的挑戰永無止盡

用Python 實作分群！

http://l.pulipuli.info/22/nsysu

用Python實作分群

1. 下載「未知分類形式」的input.ods資料集
2. 開啟「Collaboratory」
3. 上傳input.ods檔案
4. 開啟分群腳本, 複製腳本內容
5. 到「Collaboratory」貼上並執行
6. 觀察執行結果
7. 下載output.ods
8. 觀察分群結果

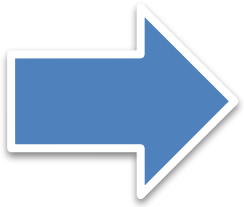
下載input.ods資料集

http://1.pull.pull.info/22/nsysu

問水果店還是寵物店 - 未知分類形式

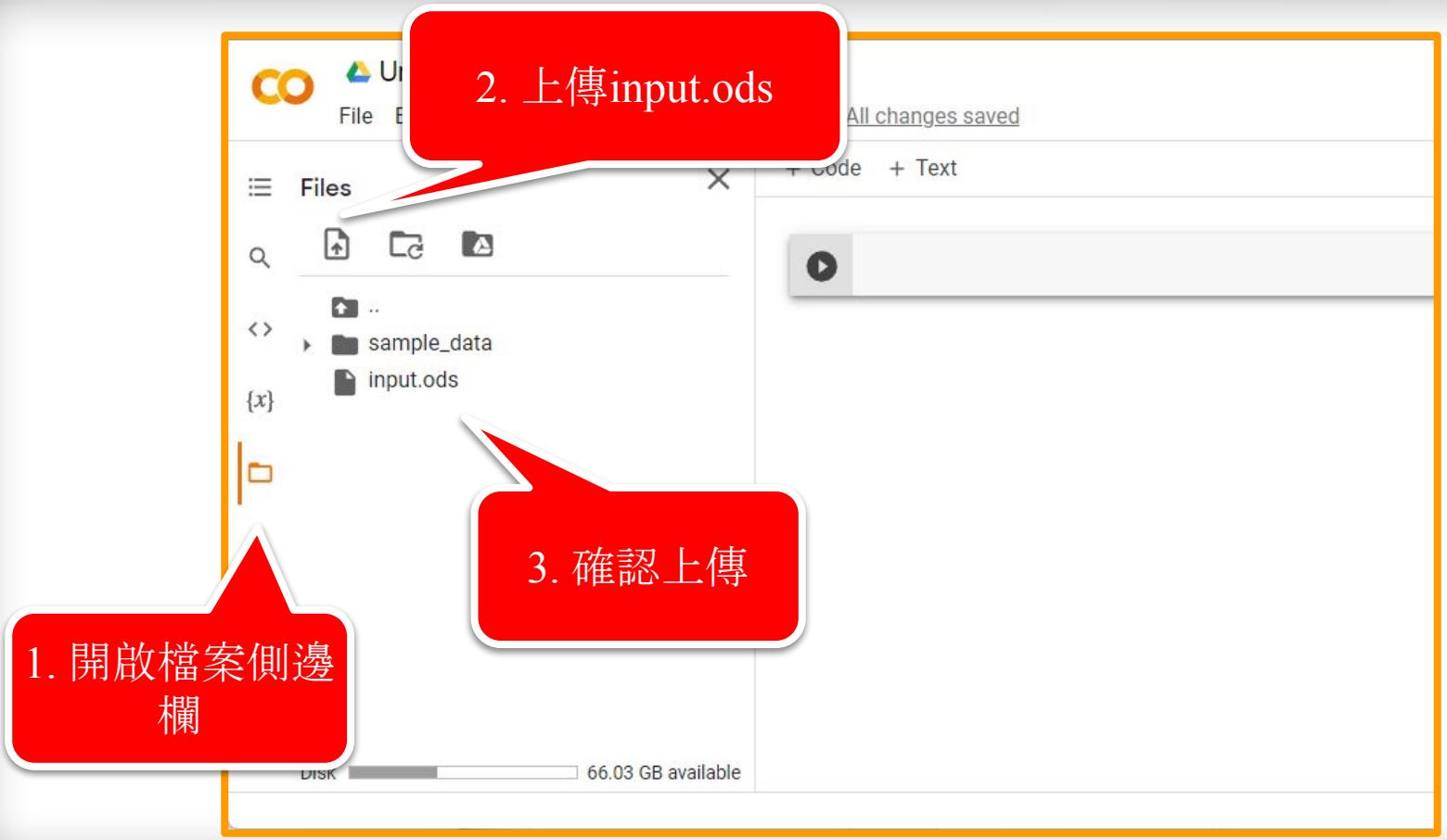
開啟「Collaboratory」

http://1.pull.pull.info/22/nsysu

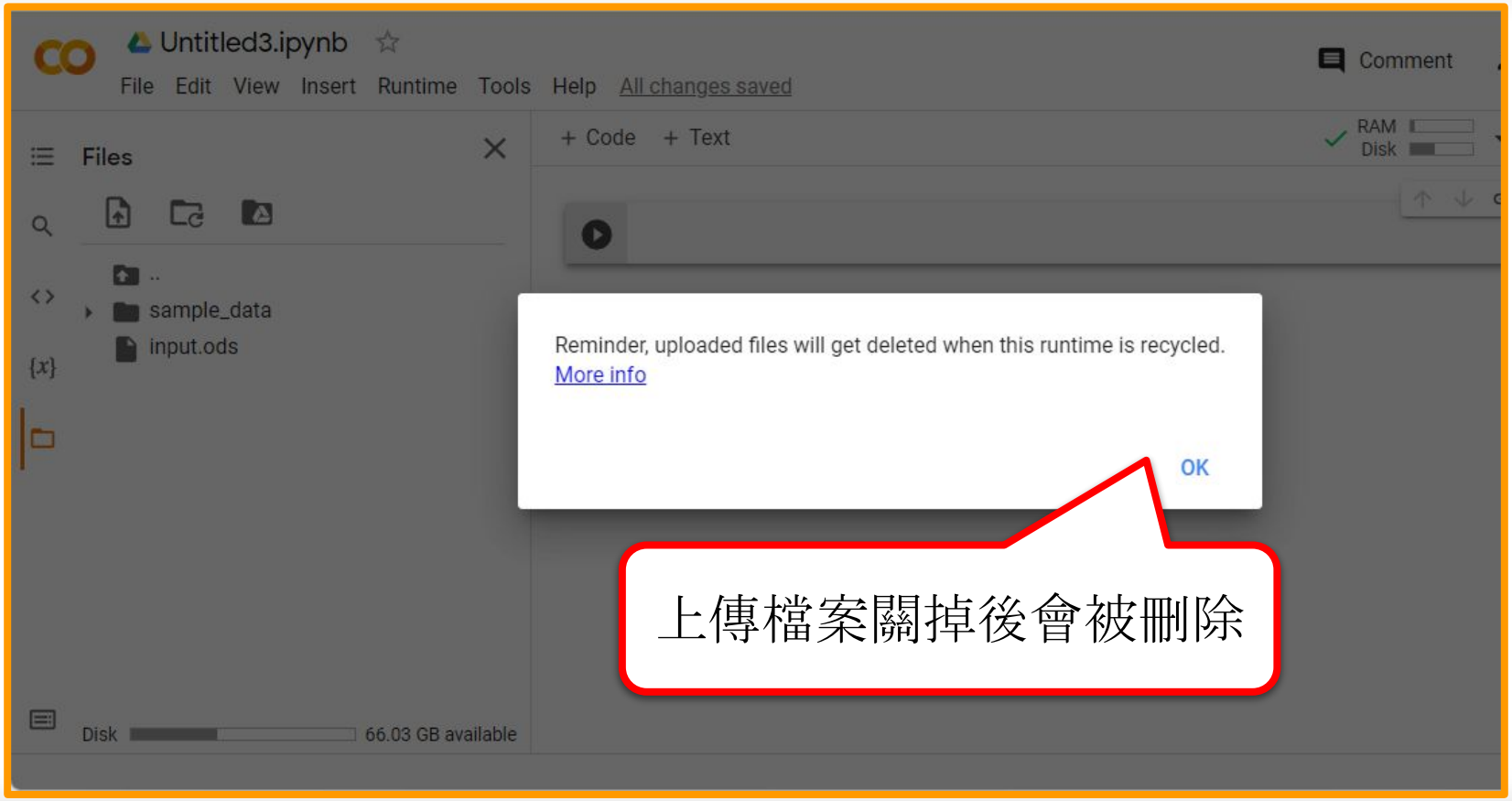


Collaboratory
開新筆記本

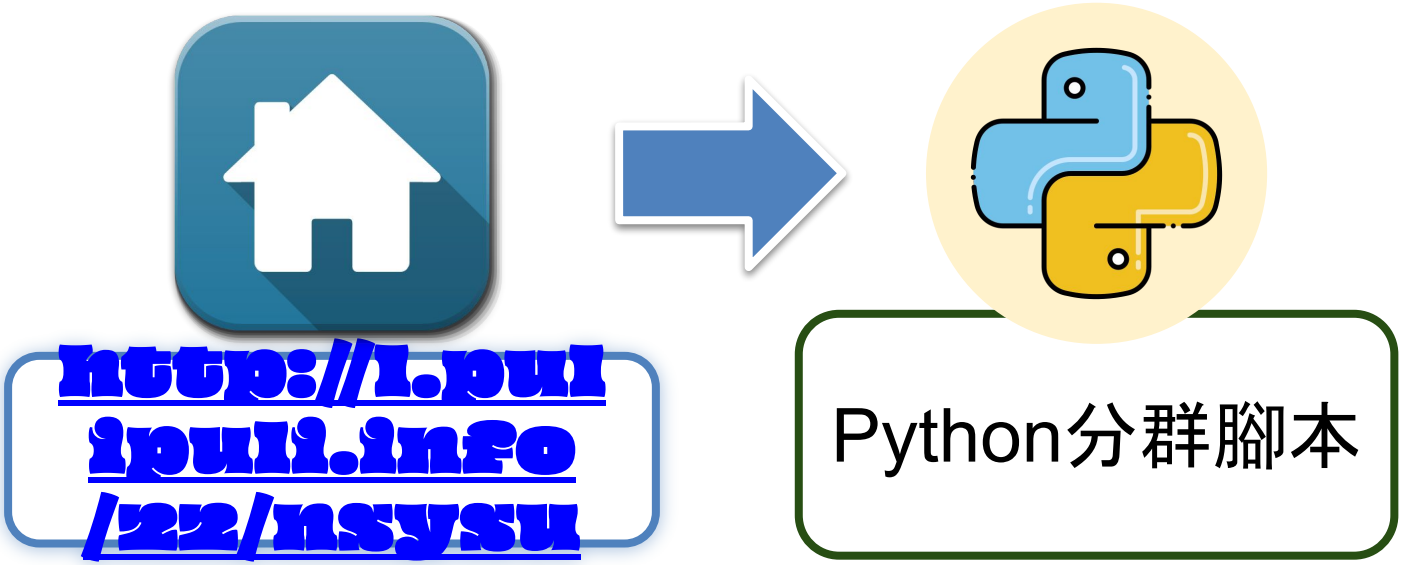
上傳input.ods檔案 (1/2)



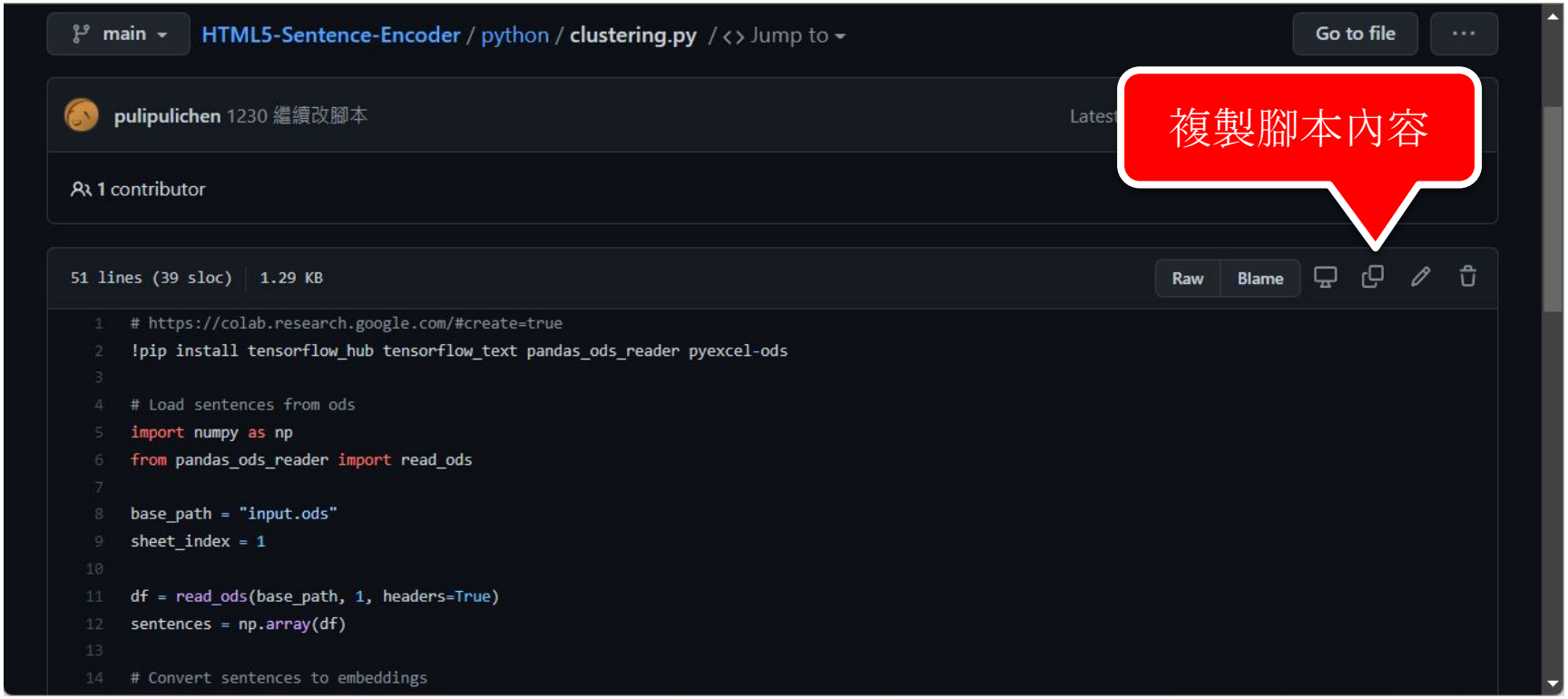
上傳input.ods檔案 (2/2)



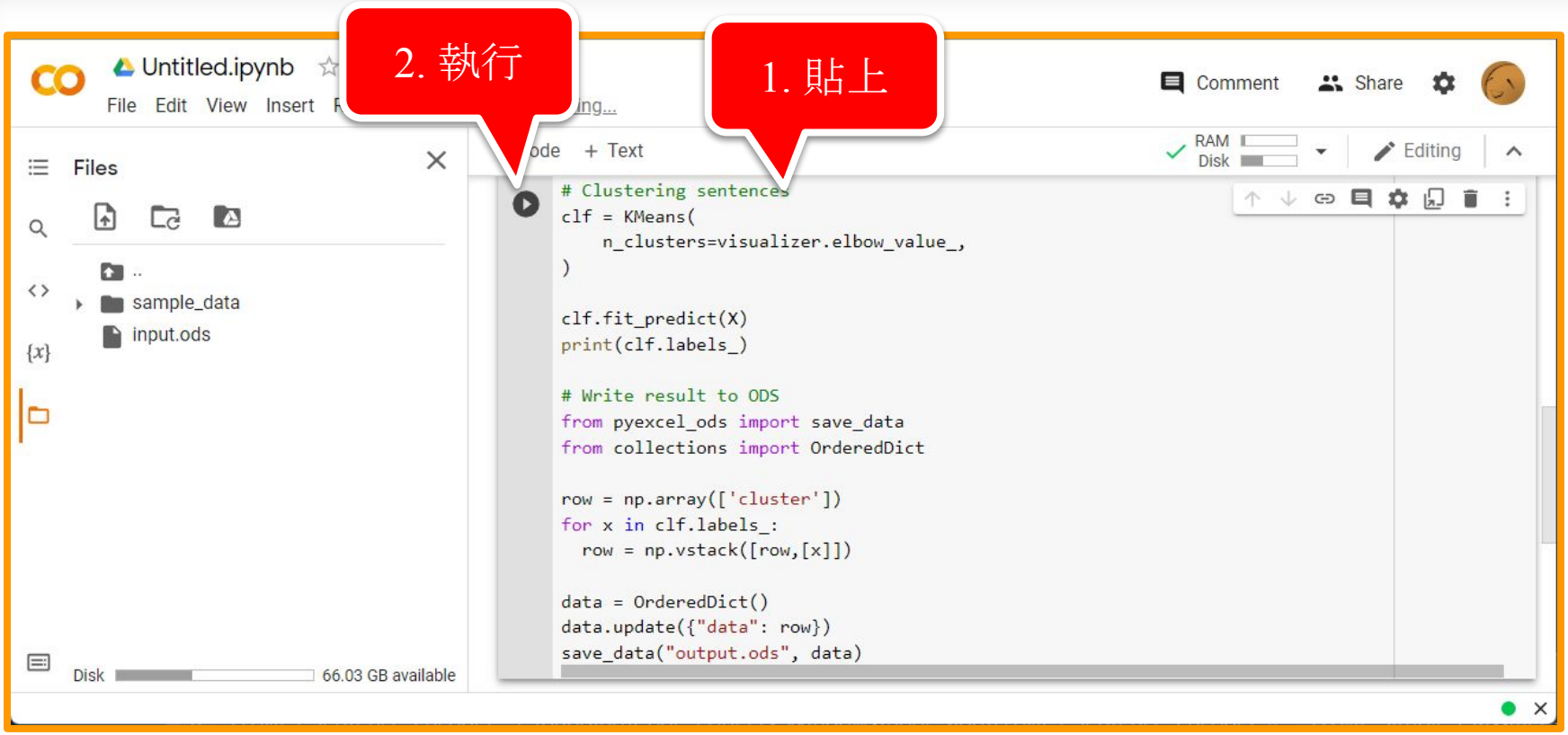
開啟分群腳本



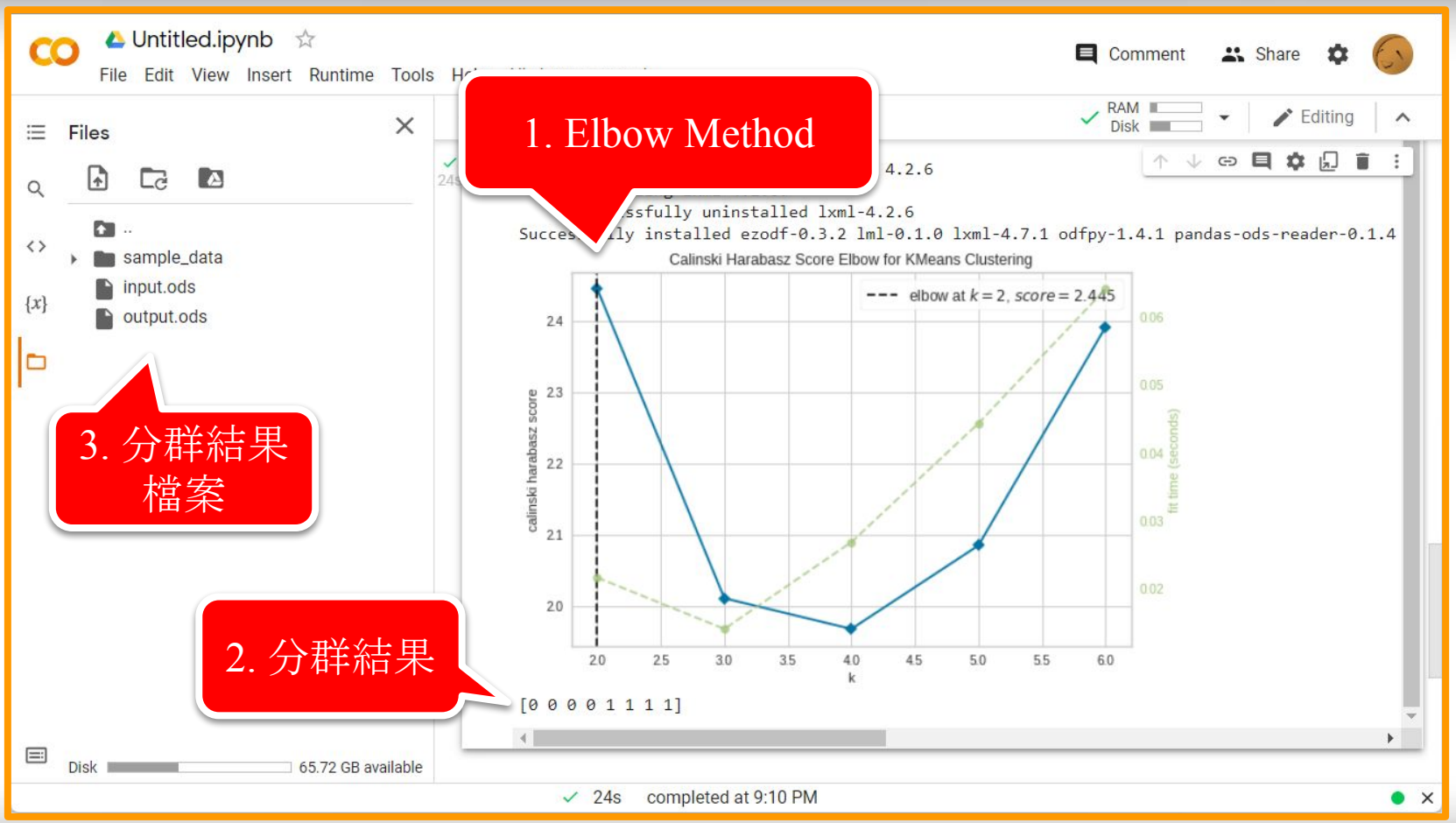
複製腳本內容


main
HTML5-Sentence-Encoder / python / clustering.py
Jump to
Go to file
pulipulichen 1230 繼續改腳本
1 contributor
複製腳本內容
51 lines (39 sloc) 1.29 KB
Raw
Blame
https://colab.research.google.com/#create=true
!pip install tensorflow_hub tensorflow_text pandas_ods_reader pyexcel-ods
# Load sentences from ods
import numpy as np
from pandas_ods_reader import read_ods
base_path = "input.ods"
sheet_index = 1
df = read_ods(base_path, 1, headers=True)
sentences = np.array(df)
# Convert sentences to embeddings

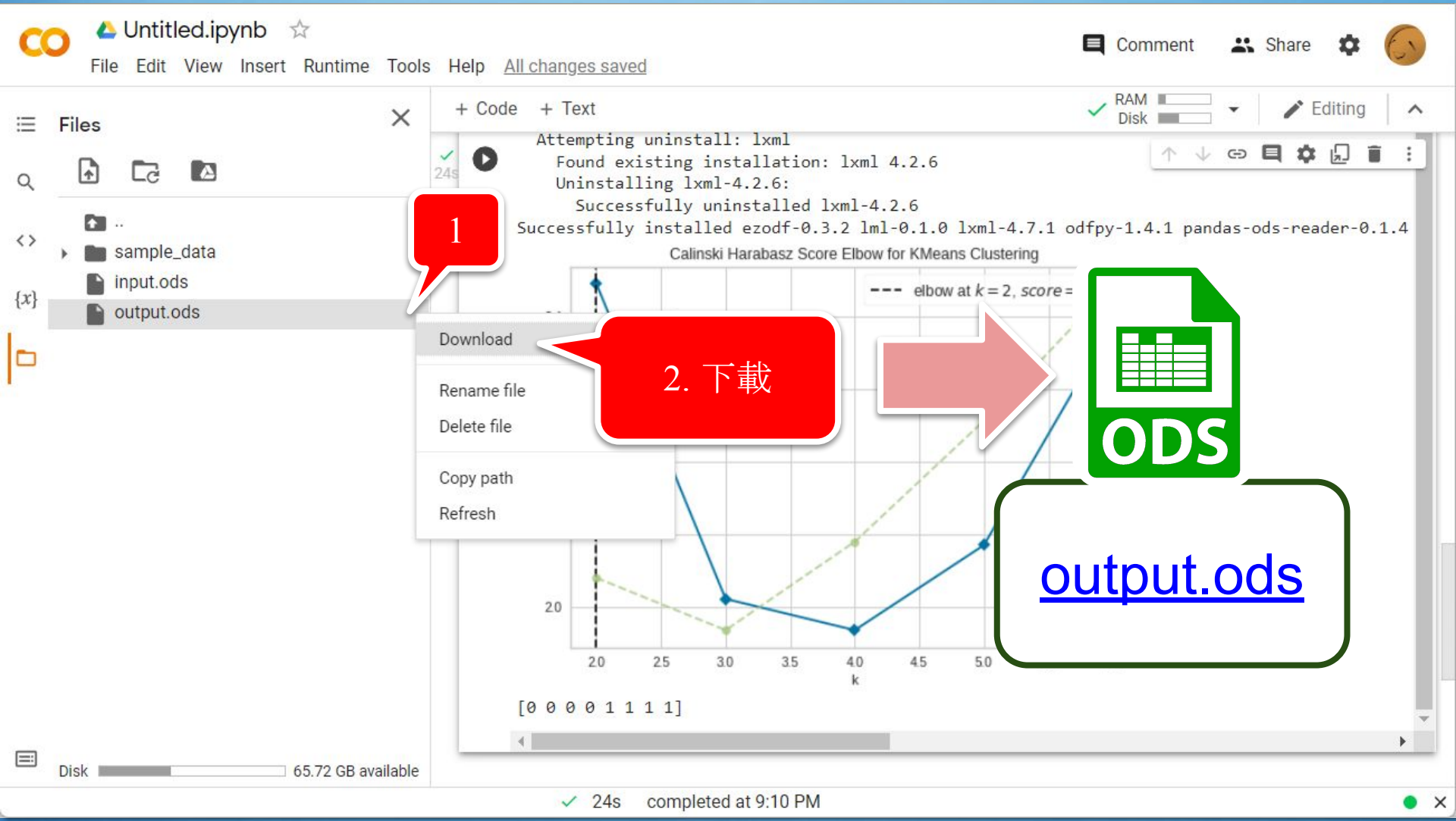
到「Collaboratory」貼上並執行



觀察執行結果

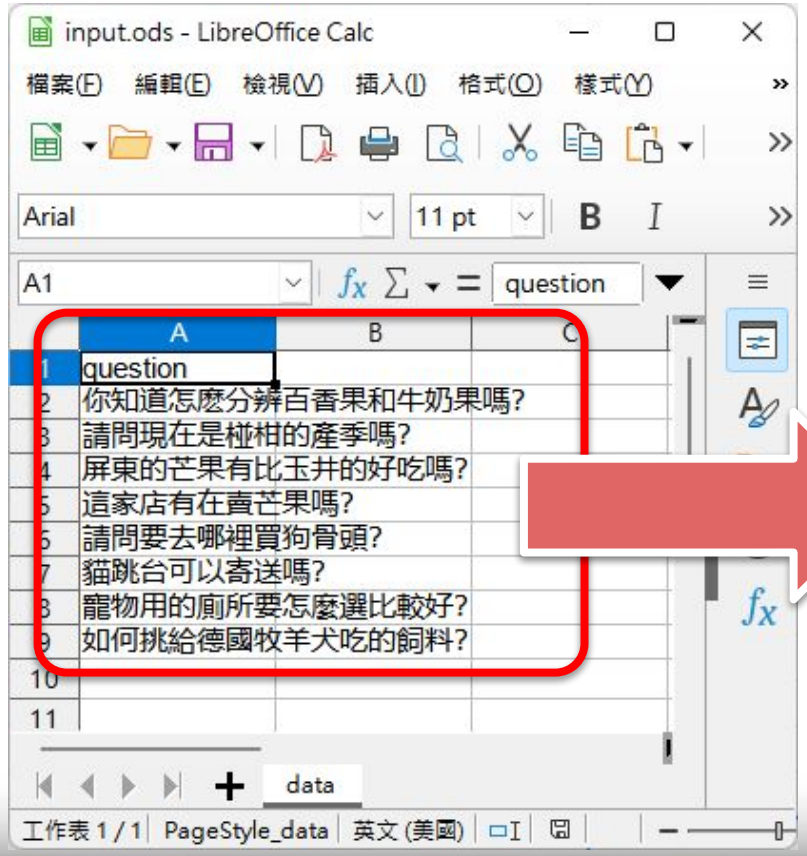

Untitled.ipynb
File Edit View Insert Runtime Tools
Comment
Share
RAM
Disk
Editing
Files
sample_data
input.ods
output.ods
1. Elbow Method
3. 分群結果檔案
2. 分群結果
4.2.6
ssfully uninstalled lxml-4.2.6
ly installed ezodf-0.3.2 lml-0.1.0 lxml-4.7.1 odfpy-1.4.1 pandas-ods-reader-0.1.4
Calinski Harabasz Score Elbow for KMeans Clustering
elbow at k = 2, score = 2.445
calinski harabasz score
fit time (seconds)
k
[0 0 0 0 1 1 1 1]
Disk
65.72 GB available
24s completed at 9:10 PM

下載output.ods

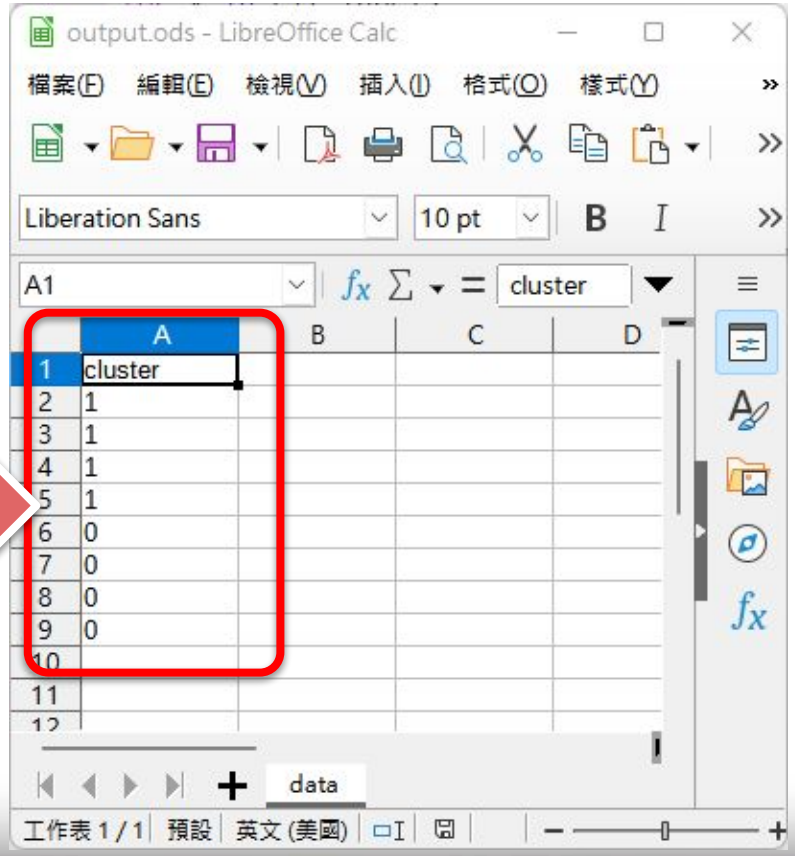


觀察分群結果

input.ods

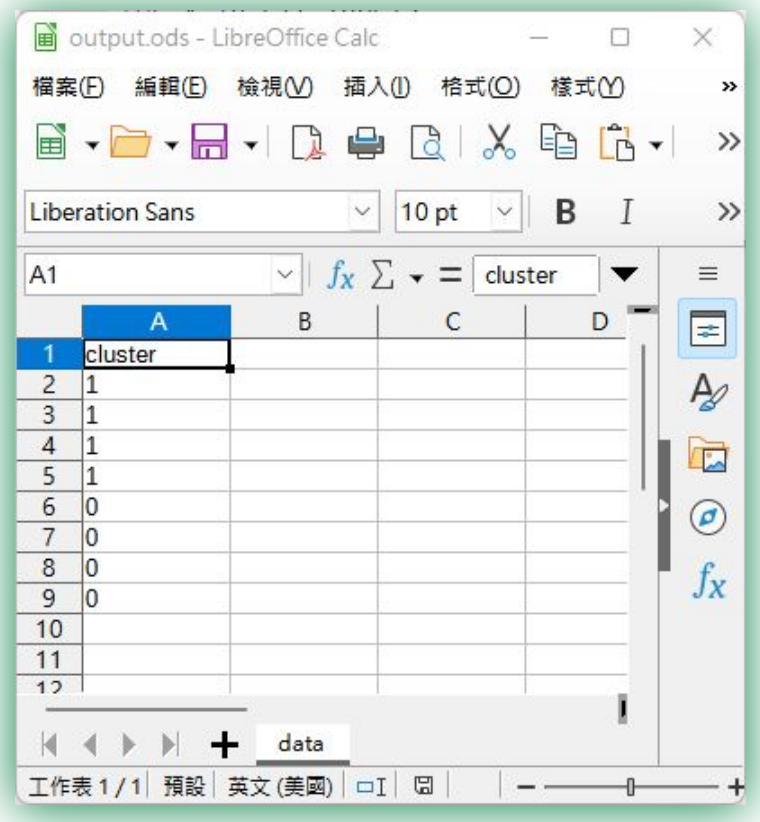


output.ods



用Python實作分群

換你囉



Python分群腳本解說

安裝需要的Python套件

```
# https://colab.research.google.com/#create=true
!pip install tensorflow_hub tensorflow_text pandas_ods_reader pyexcel-ods

'''
Loa
'''
imp
fro

base_path = "input.ods"
df = read_ods(base_path, sheet_index, headers=True)
sentences = np.array(df)

'''
Convert sentences to embeddings
'''
import tensorflow_hub as hub
import tensorflow_text

embed = hub.load("https://tfhub.dev/google/universal-sentence-encoder-multilingual/3")
```

!pip install <package>

從input.ods讀取資料


```
'''
Load sentences from ods
'''
import numpy as np
from pandas_ods_reader import read_ods

base_path = "input.ods"
df = read_ods(base_path, shee  ndex, headers=True)
sentences = np.array(df)

'''
Convert sentences to embeddings
'''
import tensorflow_hub as hub
import tensorflow_text

embed = hub.load("https://tfhub.dev/google/universal-
embedding = embed(sentences)

X = np.array(embedding.numpy().tolist())
```

把文本轉換成語義向量, 再轉換為陣列

```
14  '''
15  Convert sentences to embeddings
16  '''
17  import tensorflow_hub as hub
18  import tensorflow_text
19
20  embed = hub.load("https://tfhub.dev/google/universal-sentence-encoder-multilingual/3")
21  embedding = embed(sentences)
22
23  X = np.array(embedding.numpy().tolist())
24
25  '''
26  Determine optimal k of clustering
27  '''
28  from yellowbrick.cluster import KElbowVisualizer
29  from sklearn.cluster import KMeans
30
31  model = KMeans()
32  visualizer = KElbowVisualizer(model
33    , k=(2,7)  # k is range of number of clusters.
34    , metric='calinski_harabasz'
```

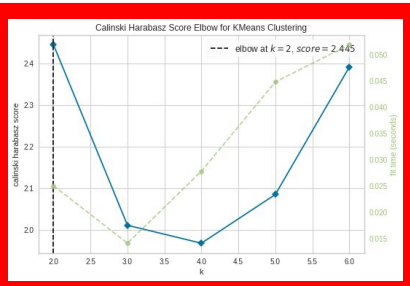
使用USE模型

把張量轉換為陣列

分析最佳分群數量k

從2~7之間決定
最佳的分群數量k

根據CH指標




```
'''
Determine optimal k of clustering
'''
from yellowbrick.cluster import KElbowVisualizer
from sklearn.cluster import KMeans

model = KMeans()
visualizer = KElbowVisualizer(model
    , k=(2,7)  # k is range of number of clusters.
    , metric='calinski_harabasz'
    , timings= True)
visualizer.fit(X) # Fit data to visualizer
visualizer.show() # Finalize and render figure

'''
Clustering sentences
'''
clf = KMeans(
    n_clusters=visualizer.elbow_value_, # Set optimal k
)
clf.fit_predict(X)
```

以最佳分群數量k進行分群

以剛剛找出的
最佳k值進行分群

分群結果
[0 0 0 0 1 1 1 1]

```
39  '''
40  Clustering sentences
41  '''
42  clf = KMeans(
        n_clusters=visualizer.elbow_value_, # Set optimal k
44  )
45  clf.fit_predict(X)
46  print(clf.labels_)  # Print the result of clusters
48  '''
49  Write result to ODS
50  '''
51  from pyexcel_ods import save_data
52  from collections import OrderedDict
53
54  # Convert 1D array to 2D array
55  row = np.array(['cluster'])
56  for x in clf.labels_:
57    row = np.vstack([row,[x]])
58
59  data = OrderedDict()
```

將分群結果寫入output.ods

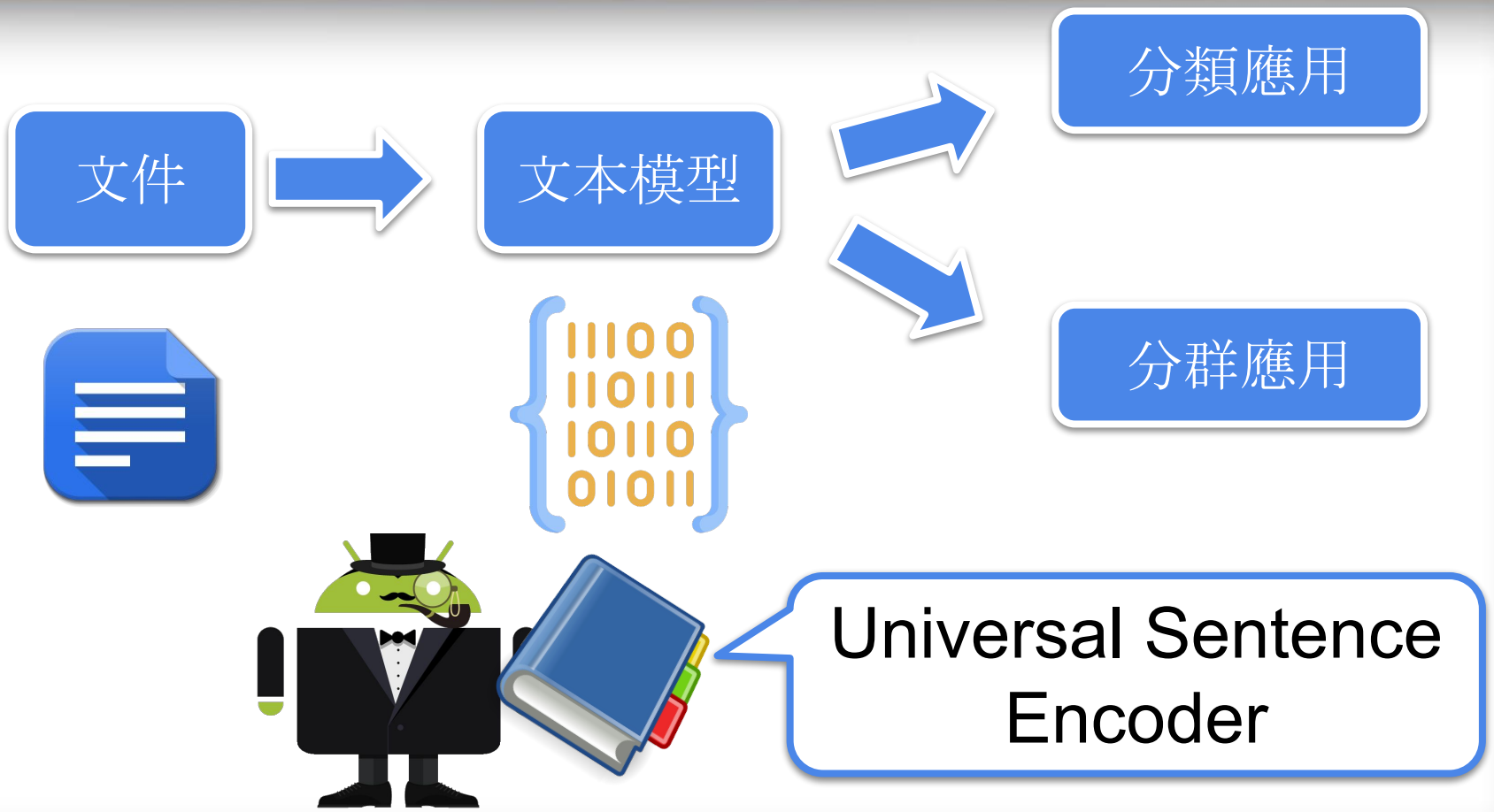
```
41  '''
42  clf = KMeans(
43      n_clusters=visualizer.elbow_value_, # Set optimal k
44  )
45  clf.fit_predict(X)
46  print(clf.labels_)  # Print the result of clusters
47
48  '''
49  Write result to ODS
50  '''
51  from pyexcel_ods import save_data
52  from collections import OrderedDict
53
54  # Convert 1D array to 2D array
55  row = np.array(['cluster'])
56  for x in clf.labels_:
57    row = np.vstack([row,[x]])
58
59  data = OrderedDict()
60  data.update({"data": row})
61  save_data("output.ods", data)
```

輸出檔案的名稱

Part 6.

結語

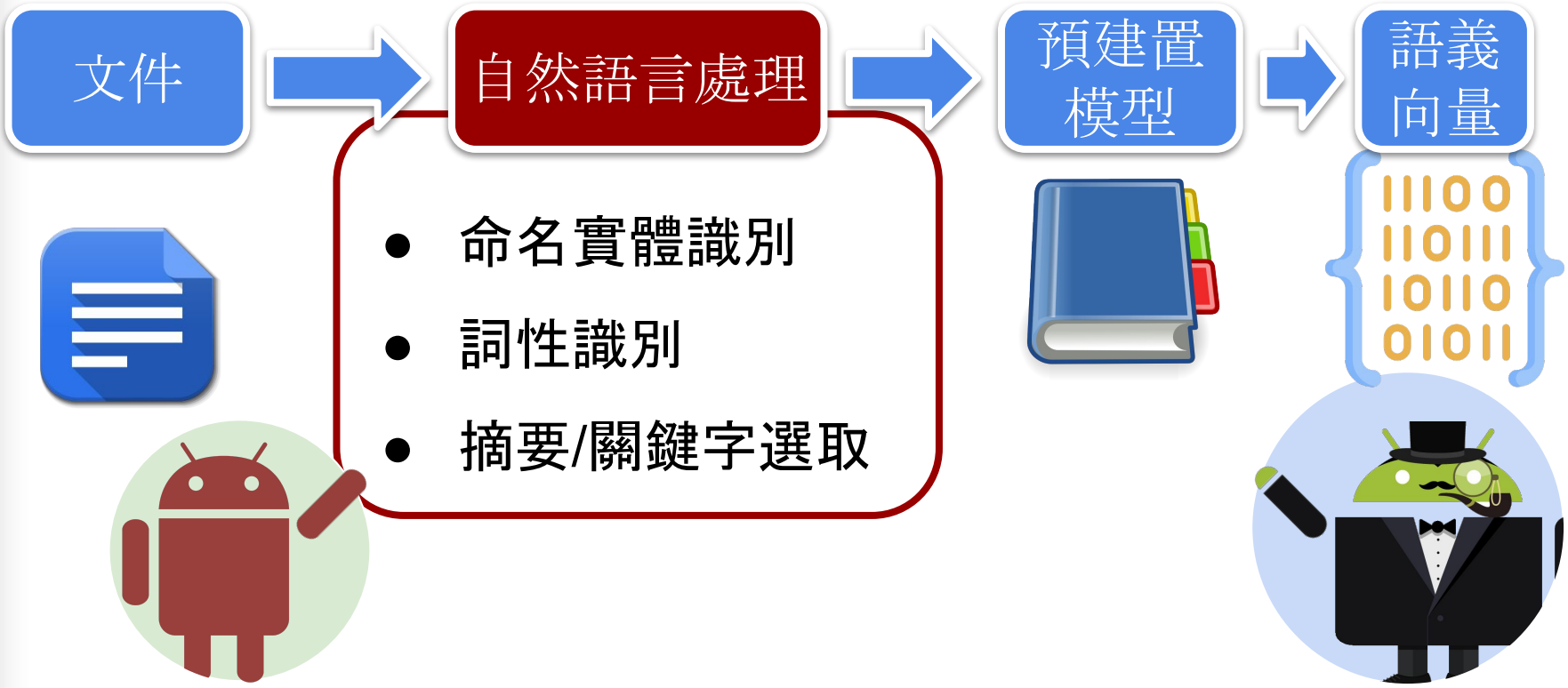
從語義向量到應用



其他的語義向量預建置模型

- **GPT-2**:使用40GB網路資料建置包含了1.5B個參數的Transformer架構
- **BiGRU**:結合attention model與hierarchical attention model對長短不同的句子做不同的處理
- **BERT**:採用雙向編碼、Transformer架構與非監督式編碼建構的模型
- **XLNet**:結合了Transformer-XL跟BERT, 採用autoregressive架構建置的模型
- **Sentence-T5**:使用encoder-decoder方法建置包含了最多11B個參數的模型

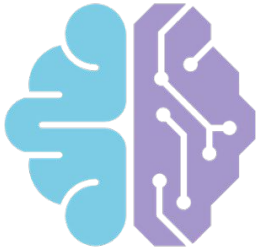
結合進階的自然語言處理



國立政治大學
人工智慧與數位教育中心
Artificial Intelligence and E-learning Center

校內活動課程
- 講座及課程持續辦理
- 學分課程持續開設
- 長期規劃AI學分學程

校外活動課程
- AI工作坊規劃
- 線上課程持續開設
- 前瞻人才培育規劃



AI 知識學習入口網
- 整合AI技術應用工具
- 應用工具使用教學
- 技術應用交流平台

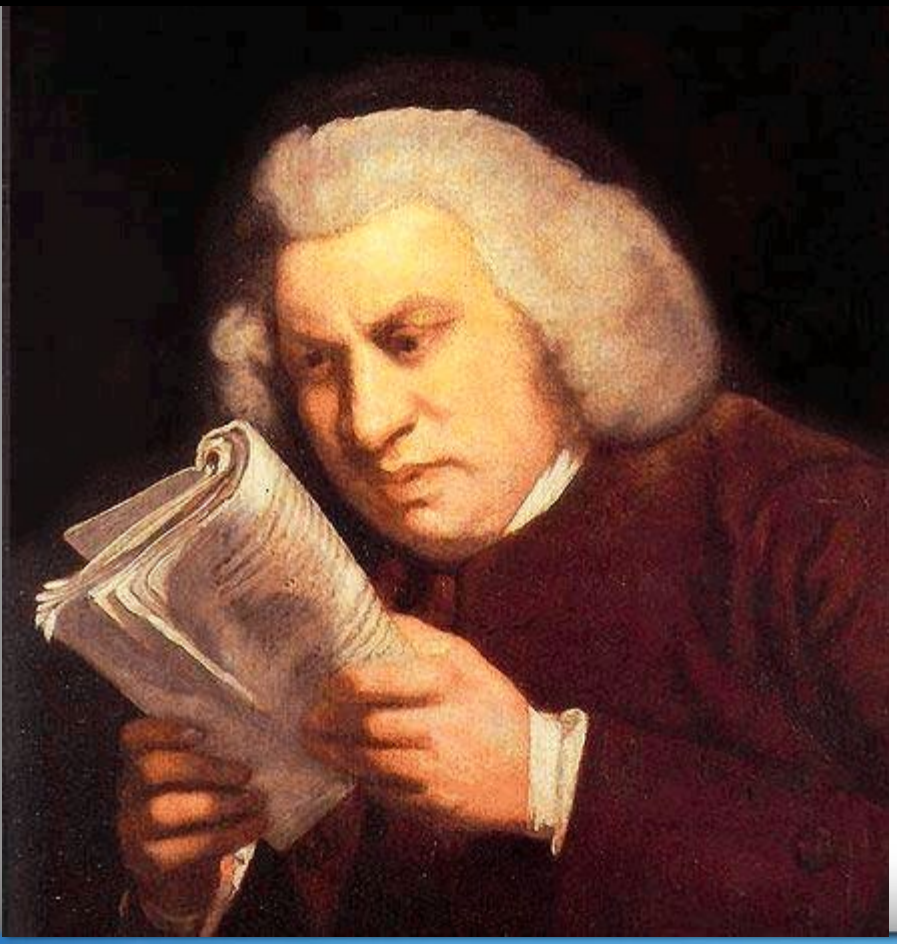
中心之友
- 電子學習歷程
- 永續回饋制度
- 課程地圖規劃

AI知識學習入口網

YouTube頻道

AI專欄

我到底看了什麼？



QA時間

拿起手機開啟網址

slido.com

http://www.shujuren.org/article/827.html

布丁布丁吃什麼？

http://blog.pulipuli.info