

作業3 - 商品搜尋與過濾機制

數據分析課程 - 作業 3

abbottabbott399@gmail.com 威霆

m11007314@gapps.ntust.edu.tw 竣崴

簡介－資訊檢索(Information Retrieval)

資訊檢索(information retrieval)的任務，是從大量文本中，找出符合使用者搜尋詞(query)的技術。



Google 搜尋

好手氣



iphone 14

搜尋

熱門 > 雨傘 | 口罩 | 旅遊 | 衛生紙 | 行李箱 | iphone













樂公益

保險館

超市快配

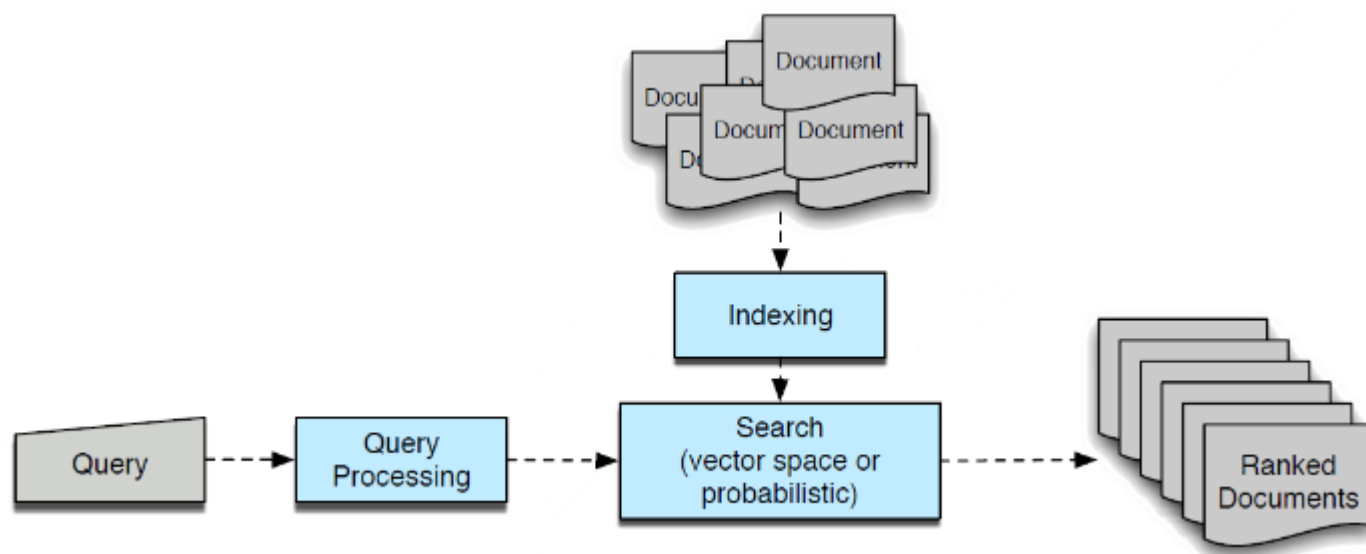
省錢達人

綜合排序 銷量排行 新上市 價格 ↑ 最低價 ~ 最高價 確認 ☐ 折價券 ☐ 分期0利率 ☐ 商品有量 ☐ 快速到貨 ☐ 超商取貨 ☐ 大家電安心配 展開

 原機貨 防塵升級 隱形防 【Apple 蘋果】iPhone 14 Pro 256G(6.1吋)(超值殼貼組) \$38,700 可訂購時通知我	 原機貨 防塵升級 隱形防 【Apple 蘋果】iPhone 14 128G(6.1吋)(超值殼貼組) \$28,200 速 登記	 滿2件折1000 【Apple 蘋果】iPhone 14 Plus 256G(6.7吋) \$35,400 速 登記 贈品	 A15 仿生晶片 【Apple 蘋果】iPhone 14 128G(6.1吋) \$27,900 速 登記	 滿1件折401 【Apple 蘋果】iPhone 14 256G(6.1吋) \$30,999 (售價已折) 速 登記	 滿1件折712 【Apple 蘋果】iPhone 14 256G(6.1吋)(犀牛盾耐震殼組) \$31,488 (售價已折) 速 登記
 超銳晶面面板/玻璃機背/組 【Apple 蘋果】iPhone 14 128G(6.1吋) \$27,900 速 登記	 滿1件折300 【Apple 蘋果】iPhone 14 Plus 256G(6.7吋)(超值殼貼組) \$35,400 (售價已折) 速 登記	 【Apple 蘋果】iPhone 14 128G(6.1吋) \$27,900 速 登記 贈品	 滿2件折1000 【Apple 蘋果】iPhone 14 Plus 128G(6.7吋) \$31,900 速 登記	 原機貨 防塵升級 隱形防 【Apple 蘋果】iPhone 14 Pro Max 256G(6.7吋)(超值殼貼組) \$42,700 可訂購時通知我	 藍芽5.1 玩樂藍芽耳機 【Apple 蘋果】iPhone 14 Pro 256G(6.1吋)(真無線藍芽耳機... \$39,200 售完補貨中

簡介－資訊檢索(Information Retrieval)

資訊檢索流程：



簡介 – TF-IDF

query：蘋果



簡介 – TF-IDF

$$w_{x,y} = \text{tf}_{x,y} \times \log \left(\frac{N}{\text{df}_x} \right)$$

詞頻（term frequency，tf）

逆向文件頻率（inverse document frequency，idf）

TF-IDF

Term x within document y

$\text{tf}_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

query：蘋果

Doc1：蘋果的秘密

Doc2：橘子的秘密

Doc3：香蕉的秘密

term	蘋	果	橘	子	香	蕉	的	秘	密
query	1*3	1*3	0	0	0	0	0	0	0
Doc1	1*3	1*3	0	0	0	0	1*0	1*0	1*0
Doc2	0	0	1*3	1*3	0	0	1*0	1*0	1*0
Doc3	0	0	0	0	1*3	1*3	1*0	1*0	1*0



作業說明

目的: 資訊檢索、TF-IDF 練習

題目: 完成八成品的程式碼，



八成品程式碼 – 建立 tf-idf 矩陣

`sklearn.feature_extraction.text.TfidfVectorizer`

Examples

```
>>> from sklearn.feature_extraction.text import TfidfVectorizer
>>> corpus = [
...     'This is the first document.',
...     'This document is the second document.',
...     'And this is the third one.',
...     'Is this the first document?',
... ]
>>> vectorizer = TfidfVectorizer()
>>> X = vectorizer.fit_transform(corpus)
>>> vectorizer.get_feature_names_out()
array(['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third',
       'this'], ...)
>>> print(X.shape)
(4, 9)
```

八成品程式碼 – 將 query 字串轉成 tfidf 向量

`transform(raw_documents)`

[\[source\]](#)

Transform documents to document-term matrix.

Uses the vocabulary and document frequencies (df) learned by fit (or fit_transform).

Parameters:: **raw_documents : iterable**

An iterable which generates either str, unicode or file objects.

Returns:: **X : sparse matrix of (n_samples, n_features)**

Tf-idf-weighted document-term matrix.

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
vectorizer.transform(['raw_documents'])
```

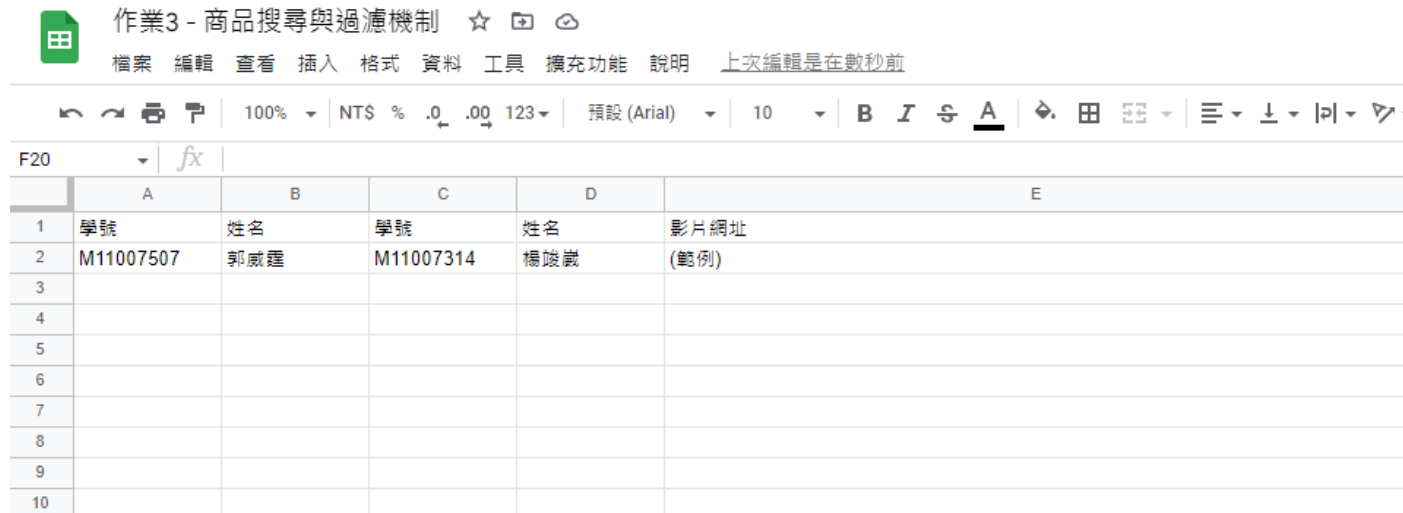

繳交規定

繳交日期：10/26 ~ 11/09

繳交內容：完成八成品的程式碼(moodle繳交)、5分鐘的講解影片(繳交至下方連結)

<https://docs.google.com/spreadsheets/d/1LFWRyRptxch4ocuXICvFzQJ5Pa7PNtni2IMZhuW2pg4/edit?usp=sharing>

檔案名稱：學號_姓名



作業3 - 商品搜尋與過濾機制 ☆ 圖 圖

檔案 編輯 查看 插入 格式 資料 工具 擴充功能 說明 上次編輯是在數秒前

100% NT\$ % .0 .00 123 預設 (Arial) 10 B I U A

	A	B	C	D	E
1	學號	姓名	學號	姓名	影片網址
2	M11007507	郭威霆	M11007314	楊竣崴	(範例)
3					
4					
5					
6					
7					
8					
9					
10					

Thank you

abbottabbott399@gmail.com 威霆

m11007314@gapps.ntust.edu.tw 竣崴