Data Mining Project 1

學號: P76111262 姓名: 林晨鈞

* Prpgramming

1. Apriori:

請執行main.py，Apriori已完成

1. FP-Tree:

請執行fp\_main.py，因為本身設計的演算法有問題，所以執行較大資料集會執行過久，我依照老師上課講義產生出一模一樣的測資(如右) 一張含有 文字 的圖片

自動產生的描述

我存成fp\_data.txt，欲執行可以複製以下指令:

python fp\_main.py --min\_sup 0.2 --min\_sup 0.2 --dataset fp\_data.txt

還請助教斟酌給分!

1. Kaggle(Bonus):

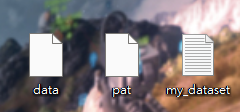
請執行kaggle\_main.py

因為此kaggle資料集還尚須去掉含None的label欄位…等資料前處理，以符合我的程式的input，所以新建一個kaggle\_main.py，欲執行請用此.py檔，輸出檔案為BreadBasket\_DMS-kaggle.csv

* 分析Report

用IBM Quest Data Generator產生自己的dataset，並取名為”my\_dataset.txt”，以下report用此資料集進行實驗，此資料集Transation ID數為4114





1. What do you observe in the below 4 scenarios?

* 高的support值代表需要出現較多次的Itemset才能成為Frequent Itemset，所以Frequent Itemset的數量會較少。反之亦然。
* 高的confidence值代表關聯性很高，形成association rule的條件變得嚴苛。反之亦然。
* High support, High confidence:

因為high support、high confidence的原因，所以frequent itemsets較少且關聯性較高者才能形成association rules，此種組合下，雖然可輕易看出關聯性高的項目，但同時也會失去很多具有不俗的關聯性的項目，因此參考價值較為不好。

* Min\_sup:0.025 / Min\_conf:0.9 Result: 11個association rules



一張含有 桌 的圖片

自動產生的描述

* High support, Low confidence:

因為high support、low confidence的原因，所以frequent itemsets較少，但關聯性不高者也能形成association rules，此種組合下，大多數的itemset會在選擇frequent itemset時就被排除掉，此組合有一定的參考價值。

* Min\_sup:0.025 / Min\_conf:0.1 Result: 16個association rules



一張含有 桌 的圖片

自動產生的描述

* Low support, Low confidence:

因為low support、low confidence的原因，所以frequent itemsets很多且關聯性低者也形成association rules，此種組合下，多數的itemset之間皆高機率形成association rules，非常不好看出區別，因此參考價值十分差。

* Min\_sup:0.01 / Min\_conf:0.5 Result: 17557個association rules



一張含有 桌 的圖片

自動產生的描述

* Low support, High confidence:

因為low support、high confidence的原因，所以frequent itemsets較多但要形成association rules的條件變得嚴苛，此種組合下，形成的association rules較完整且數量也夠多，因此有一定的參考價值。

* Min\_sup:0.01 / Min\_conf:0.99 Result: 440個association rules



一張含有 桌 的圖片

自動產生的描述

* Bonus:

kaggle資料集簡介:

The dataset consists of 21293 observations from a bakery.

The data file contains four variables, Date, Time, Transaction ID and Item.

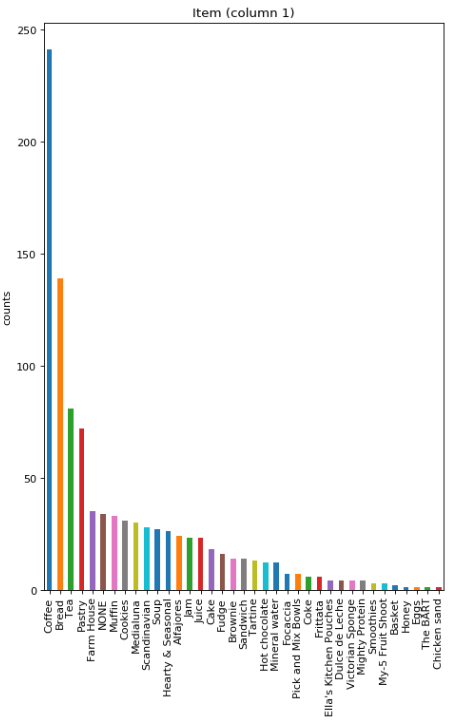
Transaction ID ranges from 1 through 9684.

However, there are some skipped numbers in Transaction IDs.

Source Website : https://www.kaggle.com/sulmansarwar/transactions-from-a-bakery?select=BreadBasket\_DMS.csv

(以下為BreadBasket\_DMS.csv格式，此資料共9684筆交易)

一張含有 桌 的圖片

自動產生的描述  


因為此kaggle資料集還尚須去掉含None的label欄位…等資料前處理，以符合我的程式的input，所以新建一個kaggle\_main.py，欲執行請用此.py檔，輸出檔案為BreadBasket\_DMS-kaggle.csv

* High support, High confidence:
* Min\_sup:0.009 / Min\_conf:0.6 Result: 1個association rules
* 此min\_sup和min\_conf的組合會產生一條association rules，可以發現Toast和Coffee兩個商品間有非常高的關聯性



一張含有 桌 的圖片

自動產生的描述

* High support, Low confidence:
* Min\_sup:0.009 / Min\_conf:0.4 Result: 16個association rules
* 此min\_sup和min\_conf的組合，可以發現Coffee這個商品，最常出現在consequent，大多數人在買完其他商品都會買Coffee



一張含有 桌 的圖片

自動產生的描述

* Low support, Low confidence:
* Min\_sup:0.003 / Min\_conf:0.4 Result: 44個association rules
* 此min\_sup和min\_conf的組合，可以發現Coffee這個商品，最常出現在consequent，大多數人在買完其他商品都會買Coffee，但是antecedent相對較雜亂，比較不好看出更細的細節。



一張含有 桌 的圖片

自動產生的描述

* Low support, High confidence:
* Min\_sup:0.003 / Min\_conf:0.6 Result:10個association rules
* 此min\_sup和min\_conf的組合，可以發現Coffee這個商品，最常出現在consequent，大多數人在買完其他商品都會買Coffee，antecedent相對較Low sup/Low conf組合，篩選出關聯度較高的association rules，以利於分析。



一張含有 桌 的圖片

自動產生的描述

* **結論:**

從上面實驗的例子，我最大的發現就是，在這間麵包店裡面，最夯的產品就是咖啡，當一個客人到這間麵包店裡面買任何商品時，高機率都會再配上咖啡。