Data Mining Homework 1

1. 這次作業目的是讓大家以 association rule，分析一家超市的交易紀錄。可以使用 Weka 或任何你熟知的語言實作。
2. Dataset 介紹
   (1) 交易紀錄：7500 筆
   (2) 商品種類：200 種
3. 資料已整理成如下圖所示:

	Banana	Bag of Organic Bananas	Organic Strawberries	Organic Baby Spinach	Large Lemon	Organic Avocado	Organic Hass Avocado	Strawberries	Limes	Organic Raspberries	...
0	False	False	True	True	False	False	False	False	False	False	...
1	False	False	False	False	False	False	False	False	False	False	...
2	True	False	False	False	False	False	False	False	False	False	...
3	False	False	False	False	False	False	False	True	False	False	...
4	False	False	True	True	False	False	False	False	False	False	...
...	...	...	...	...	...	...	...	...	...	...	...

4. 該資料可以直接套入 Weka 的 Apriori algorithm，但會找到如下的 Rule （False -> False）：

```
Organic Reduced Fat 2 percent Milk=False Organic Roma Tomato=False 7277 ==> Organic Baby Broccoli=False 7178
Organic Ginger Root=False Organic Reduced Fat 2 percent Milk=False 7250 ==> Organic Baby Broccoli=False 7151
```

   請進行簡單的資料前處理，讓 Apriori algorithm 順利找出我們感興趣的 Rule：

```
Organic Hass Avocado=True 550 ==> Bag of Organic Bananas=True 201
```

5. 接著請在報告中回答以下問題：
   Q1. 使用 Apriori algorithm，並自己設定 confidence 跟 support 試看看能 mine 出哪些 rule。挑選其中兩條 rule 並說明該 rule 所代表的意義。(在報告中，你需要附上所使用的 confidence 跟 support 的數值，及在這個設定下你選出的兩條 rule，並加上對該條 rule 的說明)
   Q2. 在 confidence 設為最低的情況下（即為 0%），將 support 設定高於多少會剛好 mine 不出任何 rule？而這個門檻的 support 值代表什麼意義？（精確到小數第二位即可）
   Q3. 在以 confidence 為 metric 且 support 設定為 0.01 的情況下，前 10 條最佳 rule 中，是否發現某個商品常常出現在這些 rule 中？(請列出前 10 條 rule 及常出現的商品為何並解釋此現象)
   Q4. 試著調整 support 及更改 metric，mine 出跟 Q3 結果較不一樣的 rule。(請解釋你為何挑選這個 metric，與先前找到的 rules 有甚麼差異)
   Q5. 簡單描述實作本次作業的過程。(參數是如何設定的，做了甚麼前處理...等)

6. 作業所需要繳交的項目：
   (1) 1~2 頁的報告，中英文都可，請使用 word 檔或是 PDF 檔。
   (2) 若是用 Weka 完成作業的同學，請附上自己資料前處理後的檔案。
   (3) 若是用其他 Tool（包含自行實作的）的同學請附上 code，如果有對資料進行前處理，也須一併附上，並在報告中說明使用甚麼 Tool。
   (4) 最後將所有項目包裝成壓縮檔上傳至 moodle，檔名請取為學號+_DM_HW1，ex：M12345678_DM_HW1