2021 Data Mining Homework 1

1. 本次作業目的是讓大家以 Association rule，分析一家電商的交易紀錄。可以使用 Weka 或任何你熟知的程式語言實作演算法來分析。

2. Dataset介紹

(1) 交易紀錄：10,000 筆

(2) 商品種類：1,000 種

3. 資料已整理成如下圖所示:

	WHITE HANGING HEART T-LIGHT HOLDER	REGENCY CAKESTAND 3 TIER	JUMBO BAG RED RETROSPOT	PARTY BUNTING	LUNCH BAG RED RETROSPOT	ASSORTED COLOUR BIRD ORNAMENT	SET OF 3 CAKE TINS PANTRY DESIGN	PACK OF 72 RETROSPOT CAKE CASES	LUNCH BAG BLACK SKULL.	NATURAL SLATE HEART CHALKBOARD	...	ENAMEL COLANDER CREAM
0	False	False	False	False	False	False	False	False	False	False	...	False
1	False	False	False	False	False	False	False	False	False	False	...	False
2	False	False	False	False	False	False	False	False	False	False	...	False
3	False	False	False	False	False	False	False	False	False	False	...	False
4	True	False	False	False	False	False	False	False	False	False	...	False
...	...	...	...	...	...	...	...	...	...	...	...	...
9995	False	False	False	False	False	False	False	False	False	False	...	False
9996	False	False	False	False	False	False	False	False	False	False	...	False
9997	False	False	False	False	False	False	False	False	False	False	...	False
9998	False	False	False	False	False	True	False	False	False	False	...	False
9999	False	False	False	False	False	False	False	False	False	False	...	False

10000 rows × 1000 columns

每 row 為一筆購物車的交易紀錄

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

```
Best rules found:

 1.  DOLLY GIRL BEAKER=False 9926 ==> SPACEBOY BEAKER=False 9900
 2. SPACEBOY BEAKER=False 9928 ==>  DOLLY GIRL BEAKER=False 9900
```

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

```
ALARM CLOCK BAKELIKE RED =True 505 ==> ALARM CLOCK BAKELIKE GREEN=True 303
```

5. 接著請在報告中回答以下問題：

Q1: 根據上課所介紹的 Rule Evaluation Metrics: Confidence 及 Lift，假設我們定義 Confidence 為在 item A 出現於購物籃的情況下，item B 出現的機率:

Conf(A -> B) = Pr(B|A)
Confidence 的其中一項缺點為，它忽略了 B 本身的出現機率 Pr(B)，試以理論或例子證明為何這在衡量 Association rule 的表現上，會是一項缺點? 並同時解釋為何 Lift 不會有這樣的問題。

Q2: 當一項衡量指標符合 measure(A,B) = measure(B,A)，我們會稱其有 symmetrical 的特性，試問 Confidence 及 Lift 各為 symmetric 或 asymmetric? 請提供證明過程或用反證法舉出某 metric 不符合 symmetric 定義之反例。

Q3: 在以 confidence 為 metric 且 min_support 設定為 0.02 的情況下，以 Apriori algorithm 分析作業提供的資料集，前 10 條最佳 rule 中，是否發現「某類型商品」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)
Hint: 同商品的不同顏色請視為同類型商品。

Q4: 在 confidence 設為最低的情況下（即為0%），將 min_support 設定高於多少，剛好令 Apriori algorithm 找不出任何 rule？而這個門檻的 min_support 值代表什麼意義？（精確到小數第二位即可）

Q5: 試著找出一組 item quadruple (W, X, Y, Z)，此組合在資料集出現的 support 數量至少為 100 筆，符合此定義的組合中， confidence 最高的 rule 為何?
Hint: 需比至小數點後第四位。

Q6: 試著調整 support 及更改 metric ，找出跟 Q3 結果較不一樣的 rules。請解釋你為何挑選這個 metric，與先前找到的 rules 有什麼差異?並列出前十名的 rules。

Q7: 簡單描述實作本次作業的過程。(用什麼工具執行演算法，參數如何設定，做了哪些前處理...等)

6. 作業所需要繳交的項目：

(1) 1頁以上的報告，中英文皆可，請使用 word 檔或是 PDF 檔。

(2) 部分證明題可以手寫後拍照的方式作答，請將作答過程照片檔貼上報告的 word 檔或是 PDF 檔一併附上，並於報告中標好各題題號，如未合併於報告檔案中，視同該題未作答。

(3) 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。

(4) 若是用其他 Tool（包含自行實作的）的同學請附上程式碼，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。

(5) 最後將所有項目包裝成壓縮檔上傳至moodle 作業區，檔名請取為:
學號+_DM_HW1，ex：M12345678_DM_HW1

(6) 作業抄襲視同未繳交。