# 108 DataScience HW3

# Association Analysis

**Dataset 說明:**

這是一個來自零售商的交易數據的資料集

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
| 536365 | 85123A | WHITE HANGING | 6 | 2010/12/1 08:26 | 2.55 | 17850 | United Kingdom |
| 536365 | 71053 | WHITE METAL L | 6 | 2010/12/1 08:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84406B | CREAM CUPID HE | 8 | 2010/12/1 08:26 | 2.75 | 17850 | United Kingdom |
| 536365 | 84029G | KNITTED UNION | 6 | 2010/12/1 08:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84029E | RED WOOLLY HC | 6 | 2010/12/1 08:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 22752 | SET 7 BABUSHKA | 2 | 2010/12/1 08:26 | 7.65 | 17850 | United Kingdom |
| 536365 | 21730 | GLASS STAR FRO | 6 | 2010/12/1 08:26 | 4.25 | 17850 | United Kingdom |
| 536366 | 22633 | HAND WARMER U | 6 | 2010/12/1 08:28 | 1.85 | 17850 | United Kingdom |
| 536366 | 22632 | HAND WARMER F | 6 | 2010/12/1 08:28 | 1.85 | 17850 | United Kingdom |
| 536367 | 84879 | ASSORTED COLO | 32 | 2010/12/1 08:34 | 1.69 | 13047 | United Kingdom |

InvoiceNo 是發票號碼

StockCode 是商品代碼

Description 是商品名稱

Quantity 是購買數量

InvoiceDate 是購買日期

UnitPrice 是商品單價

CustomerID 是客戶編號

Country 是客戶所在地

---

以下是作業前處理的步驟:

1.清洗資料，將空的 InvoiceNo 拿掉，移除被取消的交易(InvoiceNo 開頭有出現 C 即是，EX:C537024)。

2.只分析英國(United Kingdom)的購買紀錄，因為資料集有點大，所以將範圍放到只有英國地區，處理時 Country 欄位請選擇 United Kingdom。

3.把商品中 "POSTAGE" 拿掉，如果 DATA 中出現負值統一以 0 取代，郵資(POSTAGE)不列入本次分析項目中。

4.進行 Association Analysis，篩選的指標請以以下條件為準

support ≥ min threshold(0.01)，confidence ≥ min threshold(0.5)

---

**上傳檔案格式**

| index | Association | Association Rule consequents |
|---|---|---|
| 0 | 3 TIER CA | VINTAGE CHRISTMAS PAPER GIFT BAG |
| 1 | I'M ON HC | ASSORTED CREEPY CRAWLIES |
| 2 | LUNCH B/ | LUNCH BAG CARS BLUE |
| 3 | JUMBO SH | JUMBO BAG RED RETROSPOT |
| 4 | HERB MA | HERB MARKER MINT, HERB MARKER ROSEMARY |
| 5 | 3D HEART | BLUE NEW BAROQUE FLOCK CANDLESTICK, WRAP 50'S CHRISTMAS |
| 6 | WHITE FR | 20 DOLLY PEGS RETROSPOT, REGENCY TEA PLATE ROSES |
| 7 | JUMBO B/ | JUMBO BAG RED RETROSPOT |
| 8 | JUMBO B/ | JUMBO BAG PINK VINTAGE PAISLEY |
| 9 | VINTAGE | 3 HOOK HANGER MAGIC GARDEN |
| 10 | DOTCOM | JUMBO BAG WOODLAND ANIMALS, JUMBO BAG RED RETROSPOT |
| 11 | HAND WA | HAND WARMER OWL DESIGN |
| 12 | RED HAR | RING OF ROSES BIRTHDAY CARD, DOORKNOB CRACKED GLAZE PINK |
| 13 | BLUE STR | MULTICOLOUR POLKADOT PLATE, PINK & WHITE ROSEBUD RING |

依照 prediction.csv 中的 Association Rule 順序去標 label，Association Rule antecedents 以及 consequents 的欄位可能會有不只一個 item，(antecedents 即為 left-hand-side items，consequents 即為 right-hand-side items)如果有出現兩個以上的 item 時，兩個商品會以","的方式隔開(逗號+空格)，同學們從 ans_example.csv 中將 label 欄位填上 0 或 1，0 代表 antecedents 和 consequents 並沒有 Association Rule 的關係，1 代表有關係。

---

**評分方式**

程式 performance 50%

報告 50%:

1. (25%)

演算法流程及程式架構介紹、有進行哪些資料預處理手續、資料分析過程中遇到的困難及嘗試解決的方法、作業中學到的東西、程式碼執行說明(包含執行環境等資訊)

2. (25%)

此作業比較特殊的地方是，利用 Association Analysis 處理完資料後(不限於只分析英國)，還希望同學透過這些 Rules 主觀的去分析自己對於這些 Rules 的看法或是一些有趣的現象。

舉例來說: 我今天分析完法國所有的購買紀錄，發現法國人買 A 商品時也會買 B 商品。

或是調整 association rules 的度量標準(換成 lift 之類或是調整 threshold)，去說明在這度量標準之下 Rule 改變的原因，像是將 confidence 調高，發現說某國人民其實很愛買某商品之類。

這些分析都是主觀的，只要同學有將想法表達出來就行，並沒有正確答案。

---

可以使用任何的套件

Kaggle: https://www.kaggle.com/t/188b8bc1473e4c7fa9df761c829e37aa