COMP40370 Practical 2

DATA WAREHOUSES AND ASSOCIATION RULES (Part B)

Prof. Tahar KECHADI

Academic year 2023-2024

Assignment Files

- `./Practical-02-B.pdf` Assignment questions (this file).
- `./Online_Retail.xlsx` Data file for Part B: Q1 and Q2

Expected output files

- `./Prcatical-02.ipynb` Python notebook solutions.
- `./Prcatical-02.html` Python notebook in HTML format.

Requirements: Python 3.9+, pandas 1.3+, SQLAlchemy 1.4+, mlxtend 0.20

Part B: Association Rules

The main aim of this practical is to learn how to apply some popular association rules algorithms (Apriori and FP-growth) on some datasets. Python `mlxtend` library along with previously introduced libraries are required to complete this part of the practical. The dataset you need to use to complete the following questions is in "`Online_Retail.xlsx`".

Q1: Transaction Data Cleaning

1) Discard all rows with null values in `Description` and `CustomerID`. Remove all records with `InvoiceNo` starting with 'C'.
2) Remove records with Description 'POSTAGE'. Discard records with `InvoiceNo` having only one item purchased.
3) There are some customers who have different invoices issued on the same day. Merge those different invoices under one `InvoiceNo`. Remove items, which are sold less than 1000 in total.
4) Select records only related to '`United Kingdom`'. How many records do you have?
5) Create a dataframe (transactions) with `InvoiceNo` as an index and all items as columns. One row should show the quantity of each item purchased for every transaction (`InvoiceNo`) and **Zero** for unpurchased items. Convert quantity to 1 (**hot encoding**) to represent an item purchased.

Q2: Frequent Items and Association Rules

1) Use the Apriori algorithm to generate frequent itemsets with a minimum support equals to 0.02 (2%). In your answer, comment on the frequent itemsets.
2) Use the FP-Growth algorithm to generate frequent itemsets with a minimum support equals to 0.02 (2%). How these results compare to the Apriori's results?
3) Using these frequent itemsets, find all association rules with a minimum confidence equals to 0.5 (50%). Draw a scatter plot of rules showing support vs confidence.

4) Discuss the rules when the support is larger than 0.028 (2.8%) and confidence is larger than 0.5 (50%).
5) Draw the map for the most important association rules using `mlextend pivot()` and `seaborn heatmap()` functions.

The final deadline for the submission is Wednesday, 1st of November at 23:59. All submissions must be done in Brightspace.