Advanced Database System - Project 3

Group 2

Name	UNI
Li-Chieh Liu	ll3123
Qing Lan	ql2282

Files List

requirement.txt - Dependency package
main.py - Our main program
INTEGRATED-DATASET.csv
example-run.txt

How to run our program

do `pip install -r requirement.txt`

Next, you have to specify your command line parameter as follow:
do `python main.py INTEGRATED-DATASET.csv <min_sup> <min_conf>`

NYC Open Data data set you used

We use **DOHMH New York City Restaurant Inspection Results**.

Table

Resuarant Location	Type	Grade	Grade Date
Bronx	Hambuger	B	03/01/2017

Why use this data set

The original table is like that, we find it interesting to find the relashionship of resuarant location, the inspection date and the grade. For example: Restuarants in Manhattan are generally dirtier than Brooklyn.

How do you map the original NYC Open Data data set(s) into your INTEGRATED-DATASET file

A row of this data set is like the first row, so this data set is not the ordinary True-False table. We want make it into the second row, which means that we must make numeric data to some interval to make association rule mining works.

Table

Resuarant Location	Type	Grade	Grade Date
Bronx	Hambuger	B	03/01/2017
Bronx	American	B	Spring

So
1. We make a coarse catergorize for the Food Style, for example: Chinese and Korean are all considered as Asian, Hamberger and BBQ are considered as NorthAmerican. 2. We convert Grade Date into interval, that is by Season.

Since we do the two above mapping, we make a very fine data coarser, so that more data will satisfy our min_sup, and therefore find more interesting assocation rule.

The internal design of our a-priori algorithm

We simply use the original version, i.e. Section 2.1 and 2.1.1 of the Agrawal and Srikant paper in VLDB 1994. The algorithms has two step: 1. Generate all large item sets, the union of Lk for k=2...k, that all have support greater than our minimum support. - To be more specific, we first get large 1 item set L1, and a for loop that using Lk-1 to generate Ck. Ck is the candidate of large k item set, the candidate who survive through the min_sup can be part of Lk, and the list goes on. - The algorithm to generate Ck using Lk-1 is also the same as the paper, which is simply by comparing all entry except the last entry, and do some pruning to eliminate Ck those have impossible subset. 2. Generate Association Rule for those union of Lk. Make every bianry partition for every item_set that is qualified, and try to compute the confidence for this pair.

An example with Support 0.2 and Confidence 0.7

- Around 80% of Restaurant in Manhattan has the Grade A
- More than 80% of North American food-style Restaurant have grade A

You can also see the entire result in our example-run.txt file for this example