1111_P764600_1 Data Mining Project 1

Last updated: 2022/10/21
TA email: nckudm@gmail.com

Scoring Mechanism

Programming (60+20 points)

Output the largest frequent itemsets using the following algorithms:

1. Apriori (60 points)
2. FP-Tree (optional; bonus 20 points)

Report (40 points)

1. Generate your own dataset with the IBM Generator (see `IBM_Quest_Synthetic_Data_Generator_使用教學.pdf` on Moodle)
   or
2. Use the 2022 released testdata
   to test your algorithms.

Find and answer (40 points)

What do you observe in the below 4 scenarios?
(For both support and confidence, High and Low are arbitrary choices. You may set them according to your preference. As for the 2022 released testdata, it is suggested to use numbers above 0.1 for both `min_conf` and `min_sup`).

- High support, high confidence
- High support, low confidence
- Low support, low confidence
- Low support, high confidence
- Any topics you are interested in
  Please write down your discoveries in the `.pdf` report file.

Programming Language

You could choose any programming language you are familiar with for this project.

- Python3:
  - Please make sure your python version is >= 3.7.
  - You are only allowed to use built-in libraries, `numpy` and `pandas`.
- Other languages:
  - Please do not use any packages that have implemented Apriori for you. If you are not sure with what packages to use, plase confirm by sending an email to the TA mail specified at first rows.

IBM Data Format

Testdata

Download the `2022 IBM testdata檔案` on MOODLE and unzip the directory; change the `.data` file to `.txt` and you should be able to open the file.
The file you open will look like the below screenshot.
Each line contains `customer id`, `transaction id`, `item` (in this file, `customer id` equals to `transaction id`). That is, `transaction 1` contains items [12612,39159,82636,83116].
Please process the data as you prefer.

Generate your own data

If you wish to generate your own data, run the `IBM Quest Synthetic Data Generator` in Moodle.
Mac OS, Linux
Follow this Github repo instruction to generate data. To be specific, open terminal and type the following, then you should be able to get `gendata.data` and `gendata.pat` in the current directory.

```
git clone https://github.com/nrthyrk/quest
cd quest
make
./gen lit -fname gendata #Change options based on your preference
```

Windows
Download the `IBM-Quest-Data-Generator` in MOODLE and follow the `使用教學.pdf`.

Submission

- Deadline: **Nov. 4, 2022, 6:00 pm**
- Submission Format:
  Your should submit a `.zip` file with the name `{student_id}_DM_Project1`. It should be unzipped into a directory with the same name, and the directory structure should be similar to this:

```
{student_id}_DM_Project1
├── inputs (directory for input files)
├── main.py
├── report.pdf (write your `answer` to the `find and answer`)
... (your other modules)
└── outputs (directory for output files)
    ├── ibm-2021-apriori.csv (frequent itemsets; for format see the below)
```

- Code template
  This template could be used by

1. writing your algorithms into class methods or functions,
2. calling them in `main.py`,
3. typing the command-line arguments `python3 main.py --min_sup 0.1 --min_conf 0.1 --dataset ibm-2021.txt` (arguments could be changed) under the directory.
   You do not need to use this code template, but please make sure your code can work by passing this command.

- File format regulations on the `.csv` file in `outputs` directory:
  1. Output frequent item sets in column `freqset` and their supports `support`. Please do not output the index column.
  2. Items within the same frequent itemset is separated by single-space; a frequent itemset is enclosed by a bracket {}.
  3. Reference output format.