Data Mining Project 1

Scoring Mechanism

Programming (60 points)

1. FP-Tree (40 points)
   TAs will sample 10 answers generated by your algorithm. Each correct answer gain 4 points.

2. Apriori (20 points)
   TAs will sample 10 answers generated by your algorithm. Each correct answer gain 2 points.

Report (40 points)

Generate your own dataset with the IBM Generator (see `IBM_Quest_Synthetic_Data_Generator_使用教學.pdf` on Moodle) to test your algorithms.

1. 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)
   - High support, high confidence
   - High support, low confidence
   - Low support, low confidence
   - Low support, high confidence
   - Any topics you are interested in

Bonus (20 points)

1. Experiment with other dataset(s) selected from Kaggle/UCI.
   - Apply your algorithm to another dataset from Kaggle or UCI.
   - Do some experiments (eg. observe the 4 scenarios as requested for other datasets) and discover some other cool stuffs.
   - Make sure to specify the name of self-selected dataset(s), and include your discoveries in the report.

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 can only use the built-in-library in the programming implementation.
- Other programming languages:
  - Please schedule a time with TAs (nckudm@gmail.com) to come to IKM Lab(65903, CSIE New Building) to demo your project.

Submission

- Deadline: **Oct 25, 2022 23:59**.
- Late submission within 2 days (before Oct 27, 2022 23:59) will get a 20% discount. Submissions delayed for more than 2 days **will not be accpeted.**
- Please make sure that your project contains `main.py` file, `inputs` directory and `outputs` directory.
- 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:

```
hw1
├── inputs (directory for input files)
│   ├── kaggle.txt
│   └── ibm-2021.txt
├── main.py
... (maybe you have other module)
└── outputs (directory for output files)
    ├── kaggle-aprior.csv (result of aprior algorithm applied on kaggle.txt)
    ├── kaggle-fp_growth.csv (result of fp_growth algorithm applied on kaggle.txt)
    ├── ibm-2021-apriori.csv (result of aprior algorithm applied on ibm-2021.txt)
    └── ibm-2021-fp_growth.csv (result of fp_growth algorithm applied on ibm-2021.t
```

- The suggested code template following the above directory structure can be accessed here: hw1-example.zip. Note that there're some imprecisions in terms of data format information, please refer to the explanation in `IBM_Quest_Synthetic_Data_Generator_使用教學.pdf` as standard version.
- TAs will execute your `main.py` by first `cd` into the directory and executing `python3 main.py` with command-line arguments, then the `outputs` directory is expected to be generated along with the result files inside.

The most important thing

Don't Cheat !

If you cheat (copy others' works extensively, including code online) on this project, you will get a 0.