UCLA CS 145 Homework #5

DUE DATE: Wednesday, 03/06/2019 11:59 PM

Note

- You are expected to submit both a report and code. The submission format is specified on CCLE under HW5 description.
- Copying and sharing of homework are NOT allowed. But you can discuss general challenges and ideas with others. *Suspicious cases will be reported to The Office of the Dean of Students.*

1 Frequent Pattern Mining for Set Data

Given a transaction database shown in Table 1, answer the following questions. Note that the parameter `min_support` is set as 2.

(a) Find all the frequent patterns using Apriori Algorithm. Details of the procedure are expected.

(b) Construct and draw the FP-tree of the transaction database.

(c) For the item d, show its conditional pattern base (projected database) and conditional FP-tree.

(d) Find frequent patterns based on d's conditional FP-tree.

Table 1: The transaction database for the question 1.

TID	Items
1	b, c, j
2	a, b, d
3	a, c
4	b, d
5	a, b, c, e
6	b, c, k
7	a, c
8	a, b, e, i
9	b, d
10	a, b, c, d

2 Apriori for Movie Prediction

In `apriori.py`, fill in the missing lines, and run your completed code on the following three datasets:

(a) `toy.txt`: A toy test case for your debugging;

(b) `user_movies.txt`: Each line is the movies that a user has watched;

(c) `movie_tags.txt`: Each line is the tags of a movie.

Do NOT change any of the following parameters which have been set in the code: `map_name`, `min_support`, `min_conf`. Do NOT modify the `print_items_rules()` function. Please directly copy the entire output of the following command to your report in plain text format.

```
python2.7 apriori.py
```

If we partition the dataset into several sub-datasets, then find local frequent patterns corresponding to each dataset,

1. Prove that each global frequent pattern is a frequent pattern of at least one local dataset.
2. Why is a local frequent pattern **not** guaranteed to be a global frequent pattern?

3 Correlation Analysis

Table 2 shows how many transactions containing beer and/or nuts among 10000 transactions. Answer the following questions based on Table 2.

(a) Calculate `confidence`, `lift`, `chi-square test` and `all_confidence` between buying beer and buying nuts.

(b) What are your conclusions of the relationship between buying beer and buying nuts, based on the above measures?

Table 2: Contingency table for question 2.

	Beer	No Beer	Totel
Nuts	150	700	850
No Nuts	350	8800	9150
Total	500	9500	10000

4 Sequential Pattern Mining (GSP Algorithm)

(a) For a sequence $s = \langle ab\,(cd)\,(ef)\rangle$, how many events or elements does it contain? What is the length of s? How many non-empty subsequences does s contain?

(b) Suppose we have $L_3 = \{\langle (ac)e\rangle, \langle b(cd)\rangle, \langle bce\rangle, \langle a(cd)\rangle, \langle (ab)d\rangle, \langle (ab)c\rangle\}$ as the frequent 3-sequences, write down all the candidate 4-sequences C_4 with the details of the join and pruning steps.