UCLA CS 145 Homework #1

DUE DATE: Sunday, April 18 , 23:59 PM

Note

- Late submissions will generally **NOT** be accepted. Each student has an one-day extension for **ONE** of the three homework assignments if the student contacts the instructor and TA **BEFORE** the submission deadline to arrange the only late submission.
- Discussions on homework assignments are encouraged, but any form of cheating and plagiarism will **NOT** be tolerated. Every student must submit his/her own solutions on Gradescope by the deadline. *Suspicious cases will be reported to The Office of the Dean of Students.*

Table 1: The transaction database.

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

1 Frequent Pattern Mining with Apriori Algorithm

Given a transaction database shown in Table 1, answer the following questions. Note that the parameter `min_support` is set as 3. For each question, the details of your work are expected.

(a) (10%) Apply the *Apriori* algorithm to find all frequent itemsets.

(b) (5%) How many times does *Apriori algorithm* scan the database?

(c) (10%) Show the max frequent patterns and the closed frequent patterns.

(d) (10%) Now consider each item is associated with a profit as shown in Table 2. Denote max(S.profit) the highest profit of an item within the itemset S. Apply the Apriori algorithm again to find all frequent itemsets satisfying the constraint max(S.profit) ≥ 40. (Please notice that you **cannot** simply apply the constraints to the patterns mined from question (a). Instead, you must apply the constraints during the mining process for optimization.)

Table 2: The profit of each item.

Item	profit
a	0
b	5
c	40
d	15
e	15
f	-10
g	-15
h	50
i	35
j	10
k	70

2 Frequent Pattern Mining with FP-Growth Algorithm

Given a transaction database shown in Table 1, answer the following questions. Note that the parameter `min_support` is set as 3. For each question, the details of your work are expected.

(a) (15%) Construct and draw the FP-tree of the transaction database. If the support counts of multiple itemsets are identical, they should be listed in *alphabetical order*.

(b) (5%) How many times does the construction of FP-tree scan the database?

(c) (10%) Use the FP-tree and the projected database to mine frequent patterns with c but without a, d, e, f, h.

3 Sequential Pattern Mining

Given some information shown in the description, answer the following questions about sequential pattern mining.

(a) (5%) For a sequence $s = \langle (ab)(cd)efg \rangle$, how many events (elements) does it contain? What is the length of s? How many non-empty subsequences does s contain?

(b) (10%) Suppose we have the frequent 2-sequences $L_2 = \{\langle ac \rangle, \langle (ab) \rangle, \langle bc \rangle, \langle ab \rangle, \langle bd \rangle\}$, write down all the candidate 3-sequences C_3 (after joining and pruning).

(c) (10%) Given the sequential database $\{\langle a(bc)(ac)d(cf) \rangle, \langle (ad)c(bc)(ag) \rangle, \langle (eg)(ab)(df)cb \rangle, \langle e(af)cbc \rangle\}$, write down $\langle b \rangle$-projected database.

(d) (10%) Continue from part (c), find all the length-2 sequential patterns with prefix $\langle b \rangle$ with $min\_support = 2$, following the **PrefixSpan Algorithm**.