MSC-BDT5002, Fall 2021

Knowledge Discovery and Data Mining

Assignment 1

Deadline: Sep. 22nd, 11:59pm, 2021

• Submission Guidelines

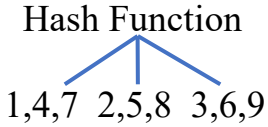
- Assignments should be submitted via Canvas.
- All assignment files should be packed into one .zip file **named in the format of**: Ax_itsc_stuid.zip.
  e.g., for a student with **ITSC account**: zxuav, student id: 20181234, the 1st assignment should be named as: A1_zxuav_20181234.zip.
- You need to zip the following three files together:
  1) A1_itsc_stuid_answer**.pdf**: please put all your answers in this document including output answers for Q1 & Q2.
  2) A1_itsc_stuid_Q1_code: this is a **folder** that should contain all your source code for Q1.
  3) A1_itsc_stuid_Q2_code: this is a **folder** that should contain all your source code for Q2.
- For programming language, in principle, **python** is preferred.
- TA will check your source code carefully, so your code MUST be **runnable**, your result MUST be **reproducible**.
- Keep your code clean and comment on it clearly. Missing the **necessary comments** will be deducted a certain score.
- Your grade will be scored based on correctness and clarity.
- Please check carefully before submitting to avoid multiple submissions.
- Submissions after the deadline or not following the rules above are **NOT** accepted.
- **Plagiarism will lead to zero points.**

(Please read the guidelines carefully)

Q1. Hash Tree (40 marks)

Suppose we have 31 candidate item sets of length 3:
{1 2 3 }, {1 4 5}, {1 2 4}, {1 2 5}, {1 5 9}, {1 3 6},
{2 3 4}, {2 5 9}
{3 4 5}, {3 5 6}, {3 5 9}, {3 8 9}, {3 2 6}
{4 5 7}, {4 1 8}, {4 7 8}, {4 6 7},
{6 1 3}, {6 3 4}, {6 8 9}, {6 2 1}, {6 4 3}, {6 7 9}
{8 2 4}, {8 9 1}, {8 3 6}, {8 3 7}, {8 4 7}, {8 5 1}, {8 3 1}, {8 6 2}

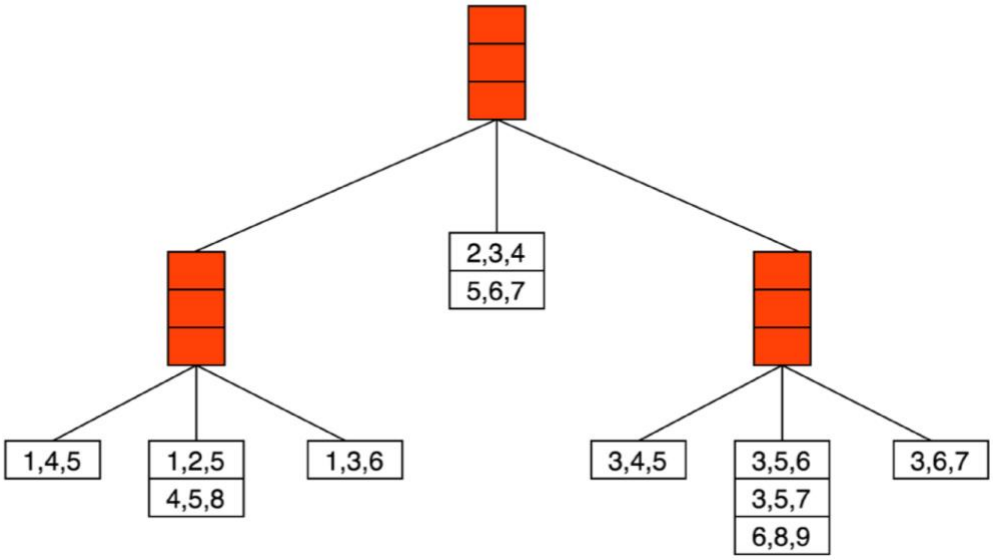
The hash function is shown in the figure below.



(a) Please write a program to generate a hash tree with max leaf size 3, output the nested list (or nested dict) of the hash tree hierarchically and draw the structure of the hash tree (you can write program to draw this hash tree or just manually draw it according to the nested list you output). [**Submission guideline**: please write the nested list/dict and the hash tree together in the **A1_itsc_stuid_answer.pdf** and put **codes files** into the **folder** with name **A1_itsc_stuid_Q1_code**.] (35 marks)

Give an example:
If the nested list is (underline is just to make the structure clearer; you don't need to draw it in your assignment):
[[[1,4,5], [[1,2,5],[4,5,8]], [1,3,6]], [[2,3,4], [5,6,7]], [[3,4,5], [[3,5,6], [3,5,7], [6,8,9]], [3,6,7]]]

Then the corresponding hash tree is:



(b) Given a transaction that contains items {1, 2, 4, 6, 7, 8}, how many comparisons are needed using the hash tree which you generate above? Please circle these candidates in the hash tree. [**Submission guideline**: please draw the circles and answer the comparisons number together in the **A1_itsc_stuid_answer.pdf**.] [**No programming required]**. (5 marks)

Notes
1. You MUST code by yourself to complete the algorithm.
2. The hash tree must be constructed by your algorithm. In other words, if the dataset changes, your algorithm should also output the correct answer.

Q2. FP-Tree (60 marks)

Frequent Pattern Mining is very important for the retail industry to increase profits. Suppose you are the owner of a grocery, there is a sale records of your store.

- **Data Description:**

DataSetA.csv

Input file of frequent pattern mining algorithms. It contains 12526 records and each record records every single transaction in the grocery store. Each line represents a single transaction with names of products. The following table is an example of it.

Lassi,Coffee Powder,Butter,Yougurt,Ghee,Cheese,
Ghee,Coffee Powder,
Lassi,Tea Powder,Butter,Cheese,
Cheese,Tea Powder,Panner,Coffee Powder,Butter,Bread,
Cheese,Yougurt,Coffee Powder,Sugar,Butter,Sweet,
……

- **Question**

(a) Please write a program to implement FP-growth algorithm and find all frequent itemsets with **support >= 2500** in the given dataset. **[submission guideline:** please print the result (all frequent itemsets with **support >= 2500)** in the **A1_itsc_stuid_answer.pdf** and put **codes files** into the **folder** with name **A1_itsc_stuid_Q2_code**.] (45 marks)

Here we give an example of problem (a)'s output:

('Coffee Powder', 'Yougurt'): 2800,

('Coffee Powder', 'Milk'): 2618,

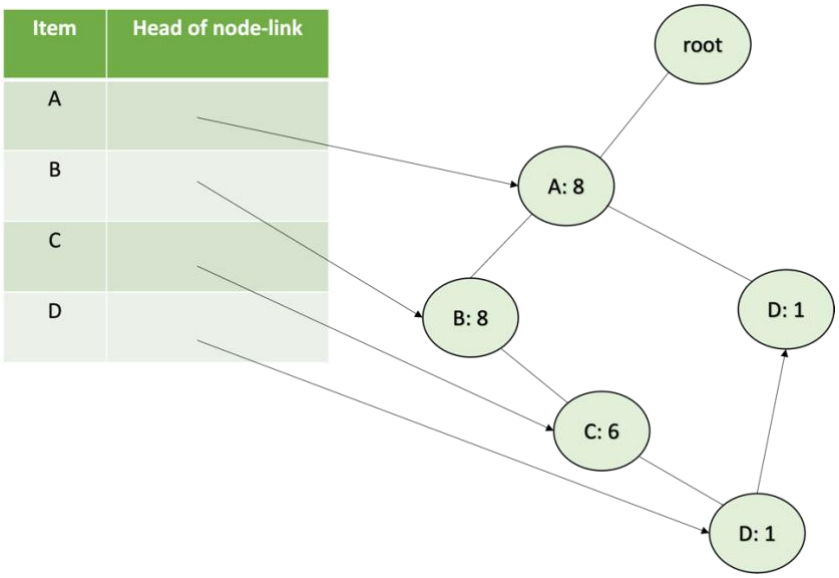
('Coffee Powder', 'Ghee'): 2518,

Each line represents a frequent itemset. Please sort frequent itemsets in descending order of support count.

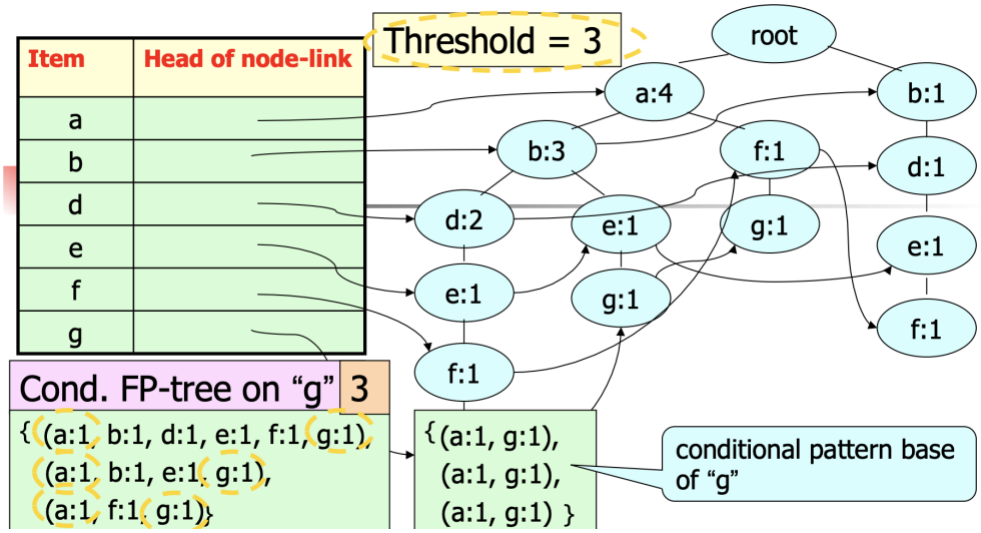
format: (term1, term2 ...): support_count

(b) What is conditional pattern base of "D" for following FP-Tree (support threshold = 2, not a complete FP-Tree, only the part related to "D" remains).

[**Submission guideline**: please print the result (conditional pattern base of "D") in the **A1_itsc_stuid_answer.pdf**.] **(15 marks) [No programming required!]**



Here we give an example of problem (b)'s output:



For the above FP-Tree (support threshold = 3), when required to find conditional pattern base of "g", we expect you print result like:

{(a:1, g:1), (a:1, g:1), (a:1, g:1)}

Notes

1. You MUST code by yourself to complete the algorithm.