CS 412: Introduction to Data Mining	**Spring 2018**
Homework 3	
Handed Out: March 7, 2018	*Due: March 28, 2018 11:59 pm*

1 General Instructions

- This assignment is due at 11:59 PM on the due date. We will be using Compass (http://compass2g.illinois.edu) for collecting the non-programming part of this assignment. Contact TAs if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The non-programming part of homework MUST be submitted in pdf format. Handwritten answers are not acceptable. Name your pdf file as YourNetid-HW3.pdf
- For Question 1, you need to explain the logic of your answer/result for every subquestion. A result/answer without any explanation will not receive any points.
- The programming part of this assignment will be hosted on hackerrank (https://www.hackerrank.com/) as a programming contest. To participate in this contest, please open a hackerrank account with your illinois.edu email id. If your username in hackerrank is different from your net id, let us know by filling out your net id and username in the spreadheet (link provided in Piazza). The contest framework will allow you to verify the correctness of your submission based on a set of sample test cases. We may use additional test cases to grade your submission. Please check the assignment page on course website in a couple of days for accessing the contest.
- It is OK to discuss the problems with the TAs and your classmates, however, it is NOT OK to work together or share code. Plagiarism is an academic violation to copy, to include text from other sources, including online sources, without proper citation. To get a better idea of what constitutes plagiarism, consult the CS Honor code (http://cs.illinois.edu/academics/honor-code) on academic integrity violations, including examples, and recommended penalties. There is a zero tolerance policy on academic integrity violations. Any student found to be violating this code will be subject to disciplinary action.
- Please use Piazza if you have questions about the homework. Also feel free to send TAs emails and come to office hours.

2 Question 1 (3 points)

Based on the following database of 5 transactions, use Frequent Pattern Growth algorithm with minimum support = 0.4 to find frequent patterns.

TransactionID	Items
1	b,d,f,g,l
2	f,g,h,l,m,n
3	b,f,h,k,m
4	a,f,h,j,m
5	d,f,g,j,m

1. Generate an ordered list of frequent items based on the raw transaction database. Note that by items we mean length 1 patterns.
2. Generate Header Table and FP-tree based on the frequent item list. Link nodes to the corresponding positions in the Header Table.
3. Generate Conditional Pattern Bases and Conditional FP-trees for items *m,h,b,j* based on the FP-tree, and list the frequent patterns computed based on each of the Conditional FP-trees.
4. Why do we order the items in each transaction by their frequency before constructing the FP-tree?
5. Which of the patterns generated in Part 3 are closed? Which are maximal?
6. List at least 2 association rules with minimum confidence = 0.6 from the frequent patterns computed in earlier parts.

3 Programming Question 2 (5 points)

This question aims to provide you a better understanding of the Apriori algorithm and closed pattern mining. Participate in the programming contest hosted at hackerrank: https://www.hackerrank.com/hw3-cs412-s18

- Please read the problem description carefully.
- The input will always be valid. We are mainly testing your understanding of frequent pattern mining, not your coding skills.

- Please pay special attention to the output format. We will be using the hackerrank based autograder and it is extremely important that your generated output satisfies the requirement.
- We don't have specific constrains for this programming question. The only constrains are the standard environment constraints in hackerrank: https://www.hackerrank.com/environment.
- The grading will be based on how many total test cases you passed. You are provided with two sample test cases to test your code. For the final grading, we will use additional test cases to test your code.
- If you have any questions, post on piazza and we will maintain a FAQ for the homework 3 programming question.