CS 412: Introduction to Data Mining	**Fall 2017**
Homework 3	
Handed Out: October 10, 2017	*Due: November* 9, 2017 *11:59 pm*

1 General Instructions

- This homework is due at 11:59 PM on the due date. We will be using Compass (http://compass2g.illinois.edu) for collecting this homework. Please contact TAs if you are having technical difficulties in submitting the homework. We shall NOT accept any late submission!
- The non-programming part of the homework **MUST** be submitted in pdf format. Handwritten answers are not acceptable. For the programming part, You are required to submit both source code and output file. Compress all files (pdf, source code, output) into one compressed file (preferably zip) and submit the compressed file ONLY. Use the following naming convention for your compressed/pdf/source code/output file: Netid-HW3.*, where * refers to the file format.
- 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 Preliminaries

Definition 1 *The* ***occurrence window*** $W_{P,S}$ *of a pattern* P *in a sequence* S *refers to the interval(s) within* S *that contains* P*.*

Definition 2 *The* ***minimum length occurrence window*** *or* ***minimum occurrence window*** $W_{P,S}^{(L-)}$ *of a pattern* P *in a sequence* S *refers to the minimum length interval(s) within* S *that contains* P*. Here, the function* $L()$ *returns length and the superscript* $(L-)$ *denotes minimum length.*

$$W_{P,S}^{(L-)} = \arg\min_{W_{P,S}} L(W_{P,S}) \tag{1}$$

Definition 3 *The* ***outlier based minimum occurrence window*** $W_{P,S}^{(O-)}$ *of a pattern* P *in a sequence* S *refers to the interval(s) within* S *that contains* P *while containing minimum possible outliers (i.e., items/elements not belonging to the pattern* P*). Here, the function* $O()$ *returns the number of outliers and the superscript* $(O-)$ *denotes minimum outlier.*

$$W_{P,S}^{(O-)} = \arg\min_{W_{P,S}} O(W_{P,S}) \tag{2}$$

For example, consider the sequences in Table 1. If we perform frequent itemset mining on these sequences with support threshold $\theta = 0.5$, we get itemsets such as {A,B,C}, {D,E,F}, etc. The minimum occurrence windows (according to **Definition 2**) of itemset {A,B,C} in S1, S2 and S3 are marked with underlines. Contrast these to the outlier based minimum occurrence windows (according to **Definition 3**) which are marked with overlines. The former involves minimizing window length, whereas the latter involves minimizing outlier count (number of outliers within window). The outlier for itemset {A,B,C} is any item/elements apart from A, B, and C, which are marked with red color in the outlier based minimum occurrence windows. While counting number of outliers within an occurence window, if an outlier element appears multiple times, you need to count it multiple times.

S1:	[A,B,B,B,B,B,C,D,E,A,F,B,C]
S2:	[F,D,D,D,E,E,A,F,D,A,B,Z,C]
S3:	[A,D,F,A,D,E,C,B,F,F,B,A,A,E,A,A,C]

Table 1: Example sequences

3 Question 1 (20 points)

For itemset {D,E,F}, report the following statistics.

- Length of **minimum occurrence window** in S1, S2 and S3. Just report the three numbers (no explanation required).
- Number of outliers in **outlier based minimum occurrence window** in S1, S2 and S3. Just report the three numbers (no explanation required).

4 Question 2 (80 points)

You are given a text file "data.txt". The first line of the file contains two constraints θ, ϵ; followed by multiple sequence of integers, one sequence per line. Both θ and ϵ are real numbers in [0, 1] range. The first constraint θ represents support– the fraction of sequence where a pattern needs to appear for the pattern to be frequent. The second constraint ϵ is outlier resilience– number of outliers in outlier based minimum occurrence window divided by the pattern size (number of distinct elements in pattern) in a supporting sequence. We identify a frequent pattern as outlier resilient if it satisfies ϵ-resilience in at least θ fraction of sequences. You need to write a program that identifies all the outlier resilient itemsets from the sequences in "data.txt".

Example: Consider the following input in "data.txt" file, where $\theta = 0.5$, $\epsilon = 0.5$.

0.5, 0.5
1, 2, 2, 2, 3, 4, 1, 5, 6
4, 5, 6
1, 4, 2, 5

Now, if we perform frequent itemset mining on these sequences with $\theta = 0.5$, we will get the following itemsets. Here, S1, S2, S3 refers to the first, second and third sequence in "data.txt".

{1}: S1, S3
{2}: S1, S3
{4}: S1, S2, S3
{5}: S1, S2, S3
{6}: S1, S2
{1, 2}: S1, S3
{1, 4}: S1, S3
{1, 5}: S1, S3

{2, 4}: S1, S3
{2, 5}: S1, S3
{4, 5}: S1, S2, S3
{4, 6}: S1, S2
{5, 6}: S1, S2
{1, 2, 4}: S1, S3
{1, 2, 5}: S1, S3
{1, 4, 5}: S1, S3
{2, 4, 5}: S1, S3
{4, 5, 6}: S1, S2

Some of these itemsets are outlier resilient at $\epsilon = 0.5$, others are not. For example, consider itemset {4, 6}. The itemset appears in two sequences: S1, S2. Out of these two appearances, its appearance at S2 satisfies ϵ-resilience at $\epsilon = 0.5$ [In outlier based minimum occurrence window: number of outliers = 1, pattern size = 2, number of outliers/pattern size = 0.5]; however its appearance at S1 does not [In outlier based minimum occurrence window: number of outliers = 2, pattern size = 2, number of outliers/pattern size = 1.0]. Therefore, the {4, 6} itemset satisfies ϵ-resilience in $1/3 = 0.33$ fraction of sequences and is not outlier resilient (as $\theta = 0.5$). In contrast, consider itemset {4, 5}. It appears in three sequences (S1, S2, S3), and all three appearances satisfies ϵ-resilience at $\epsilon = 0.5$. Overall, the {4, 5} itemset is outlier resilient.

Output Format: You need to print/write the outlier resilient itemsets in a text file, one itemset per line. The items in the itemset should be printed in sorted order. You can print the itemset themselves (which itemset to be printed in which line) in any order. For example, if the output consists of itemset {4, 5} and {4, 6}, the output could be one of the following.

Output 1:

4, 5
4, 6

Output 2:

4, 6
4, 5