CSE 40647/60647: Data Science	**Spring 2018**
Homework 4	
Handed Out: April 3, 2018	*Due: April 19, 2018 11:59 pm*

- This assignment is due at **11:59 PM** on the due date. Contact TA if you have technical difficulties in submitting it on **Sakai**. We shall NOT accept any late submission!
- Homework must be submitted in ZIP format (including .pdf, .py and datafile you use). Name your ZIP file as **YourNetid-HW*x*.zip**. Handwritten answers must be scanned into PDF.
  – YourNetid-HWx.zip
  —- YourNetid-HWx.pdf
  —- YourNetid-HWx-Qy.py
  —- ... (and any supplementary materials)
  —- Please provide your python version in your PDF file. Regardless of tools you use for your python programming, please submit the code in .py format so that TA can run it via command line.
- Please use **Piazza** if you have any question about the homework.

1 Apriori and FP-Growth (50 points)

A database has 10 transactions. Let $min\_sup = 2$. Items are a, b, c, d, and e.

Trans. ID	**Itemset**
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

1. Use Python to implement Apriori to find all frequent patterns (i.e., frequent itemsets) and their counts from the transaction database. Please submit your code as **YourNetid-HW*4*-Q*1*.py**.
2. Draw the first FP-tree that the FP-Growth algorithm creates when given this transaction database. By saying the "first", this FP-tree is not a conditional FP-tree. Write down the reason that FP-Growth is often more efficient than Apriori on the PDF.

You don't have to implement FP-Growth though you should be able to. You are not asked to use it to find the frequent patterns in this homework.

2 Pattern Evaluation Measures (10 points)

The definitions of two measures, *lift* and *cosine*, look rather similar as shown below,

$$\text{lift}(A, B) = \frac{s(A \cup B)}{s(A) \times s(B)}, \tag{1}$$

and

$$\text{cosine}(A, B) = \frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}}, \tag{2}$$

where $s(X)$ is the *relative* support of itemset X. Which measure is *null-invariant*, and which is not, and why? Can you prove it?

3 Closed Patterns (20 points)

A database has 4 transactions as shown below. Let $min\_sup = 2$. Items are A, B, C, D, E, F, and G.

Trans. ID	Itemset
1	{A, C, F, G}
2	{A, B, C, F}
3	{A, B, C, D, F}
4	{B, D, E}

Which patterns from the following are **closed patterns**? Please briefly describe your idea for each pattern on why it is closed or not.

- Pattern 1: {D}
- Pattern 2: {A, B, C, F}
- Pattern 3: {B, F}
- Pattern 4: {B, D}
- Pattern 5: {A, C, F}

Seq. ID	Sequence
1	(AB)C(FG)G
2	(AD)CB(ABF)
3	AB(FG)

4 Sequential Patterns (20 points)

A sequence database has 3 sequences as shown below. Items in the same parenthesis means they were got together in one event. Let $min\_sup = 2$. Items are A, B, C, D, F, and G. Which patterns from the following are **sequential patterns**? Please briefly describe your idea for each pattern on why it is a good sequential pattern or not.

- Pattern 1: ACF
- Pattern 2: (FG)B
- Pattern 3: (FG)
- Pattern 4: B(FG)
- Pattern 5: GF