# COMP 4331 Data Mining, Spring 2018

Assignment 1

**Deadline: 23:59pm Mar 5th, 2018**

## 1 Submission Guidelines

- Assignments should be sumbmitted to comp4331spring18@gmail.com as attachments.
- You need to zip the following two files together:
  - A1_itsc_stuid_report.pdf/.docx: Please put all your reports in this file. (Attachments should be original .pdf or .docx, NOT compressed)
  - A1_itsc_stuid_code.zip: The zip file contains all your source codes for the first assignment.
- All attachments, including report and code, should be named in the format of: Ax_itsc_stuid.zip. E.g. for a student with itsc account: sdiaa, student id: 20171234, the 1st assignment can be named as: A1_sdiaa_20171234.zip.
- Submissions after the deadline or not following the rules above are NOT accepted.
- Your grade will be based on the correctness, efficiency and clarity.
- The email for **Q&A**: sdiaa@connect.ust.hk or zyanad@connect.ust.hk
- **Plagiarism will lead to zero mark.**
- Updated date: Feb 21th 2018.

# 2 Frequent Pattern Mining via Programming

There is a transacton database stored in the link: `https://github.com/HKUSTcomp4331/sample-code-data-mining/blob/master/freq_items_dataset.txt`.

## 2.1 Mine Frequent Itemsets

You are required to write Python 2 or Python 3 programs to mine the frequent itemsets in that database with following methodologies (min_sup = 100), respectively:

- Frequent itemset mining using Apriori algorithm (Compile Apriori method in the A1_itsc_stuid_code_Apr.py). (**5 marks**)
- Frequent itemset mining based on FP-Growth (Compile FP-growth method in the A1_itsc_stuid_code_fp.py). (**5 marks**)
- Frequent itemset mining via Recursive Elemination (Compile Relim method in the A1_itsc_stuid_code_relim.py). (**5 marks**)

You are required to report the running time of each method in the assignment document A1_itsc_stuid_report.pdf/docx. Please give the reason why their performances vary in terms of efficiency. (**15 marks**)

## 2.2 Mine Frequent Itemsets

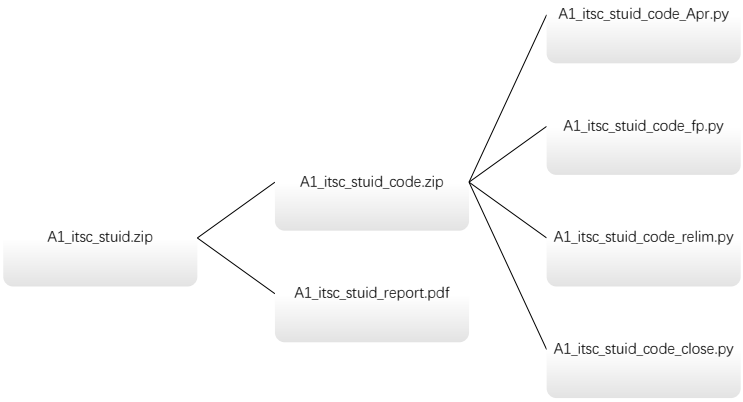
Based on the frequent itemsets you mined previously, please write a Python 2 or Python 3 program to mine the **closed frequent itemsets** and **maximal frequent itemsets**. Please name your program "A1_itsc_stuid_code_close.py" and post your results in A1_itsc_stuid_report.pdf/docx. (**20 marks**)

**Data Description**

The benchmark dataset freq_items_dataset is supported by the IBM Almaden Quest research group, which contains 1,000 items and 100,000 transactions. For simplicity, each number uniquely identifies an item.

# 3 Notes

- Note that all the codes should be compilable and well-commented (provide enough comments for each key line of code), otherwise you may lose some marks if the code is very difficult to understand.
- Please submit your assignment answer as following structures:




- **References:**
  - Mining Frequent Patterns without Candidate Generation, J. Han, H. Pei, and Y. Yin.
  - Keeping Things Simple: Finding Frequent Item Sets by Recursive Elimination, Borgelta C.