COMP 4331: Spring 2020
Assignment 1

Deadline: 23:59 17th of March, 2020

1 Submission Guidelines

- Assignments (including all attachments) should be emailed to comp4331spring2020@gmail.com before the deadline. All submissions after the deadline will not be accepted and will incur a mark of 0.

- Your final submission, titled A1_stuid.zip (where stuid is your itsc student id i.e. atzhou) should be zipped file composed of two files:

  1. **A1_stuid_report.pdf** or **A1_stuid_report.docx**: Your report detailing your environment and experimental outcomes. This file **must** be either a pdf or docx file.

  2. **A1_stuid_code.zip**: A zipped file containing all source code used to complete this assignment. All code should be compilable and well-commented. Please either provide a separate file for each technique as mentioned in section 2.

- By submitting any work, you are acknowledging that you are the sole contributor of the work unless specified and properly referenced. **All plagiarism will not be tolerated and incur a mark of 0.**

- Your work will be graded on correctness, efficiency and clarity of communication.

- To inquire about any questions you may have regarding the assignment, please email:

  atzhou@connect.ust.hk or jfangak@connect.ust.hk

2 Frequent Itemset Mining via Programming

The dataset 'a1dataset.txt' is located on Canvas, you may find it in "Files → Assignment 1".

2.1 Mine Frequent Itemsets

You are required to write a Python 2/3 program to mine the frequent itemsets over the dataset via the following methodologies ($minsup = 400$), respectively:

- Using Apriori (Compile the Apriori method in the A1_stuid_code_Apr.py). (**20 marks**)
- Using Hash Tree (Compile the Hash Tree method in the A1_stuid_code_ht.py). (**10 marks**)
- Using FP-Growth (Compile the FP-Growth method in the A1_stuid_code_fp.py). (**10 marks**)

You are required to report the **running time** of each method. Also, the **environment** you use (System, CPU, RAM, etc) should be provided. Please give the **reason** why their performances vary in terms of efficiency. (**40 marks**)

2.2 Closed Frequent Itemsets

Based on the frequent itemsets you mined previously, please write a Python 2/3 program to mine the **closed frequent itemsets** and **maximal frequent itemsets**. Please name your program as "A1_stuid_code_close.py" and post your results in A1_stuid_report.pdf/docx. (**20 marks**)

2.3 Data Description

The dataset is donated by Tom Brijs and contains the (anonymized) retail market basket data from an anonymous Belgian retail store. There are a total amount of 88,163 receipts being collected. For simplicity, each number uniquely identifies an item.

3 References

Brijs T., Swinnen G., Vanhoof K., and Wets G. (1999), The use of association rules for product assortment decisions: a case study, in: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego (USA), August 15-18, pp. 254-260. ISBN: 1-58113-143-7.