University of Windsor

60-425 Big Data Analytics and Database Design

Project I (15%)
Mining Frequent Itemsets

Deadline: Tuesday February 13th 2018 at 12:00pm

Important Note: This project can be done in a **group of two** or **individually**. If you want to do the project in a group of two, you have to send the name of your teammate to the graduate assistant Hetal Rajpura via email (**rajpurah@uwindsor.ca**) no later than Friday February 2nd at 12:00pm. After this date, if we don't receive your team, we assume you perform the project individually.

Description
The main objective of this project is to find frequent itemsets by implementing two efficient algorithms: **A-Priori** and **PCY**. The goal is to find frequent **pairs** of elements. You do not need to find triples and larger itemsets.

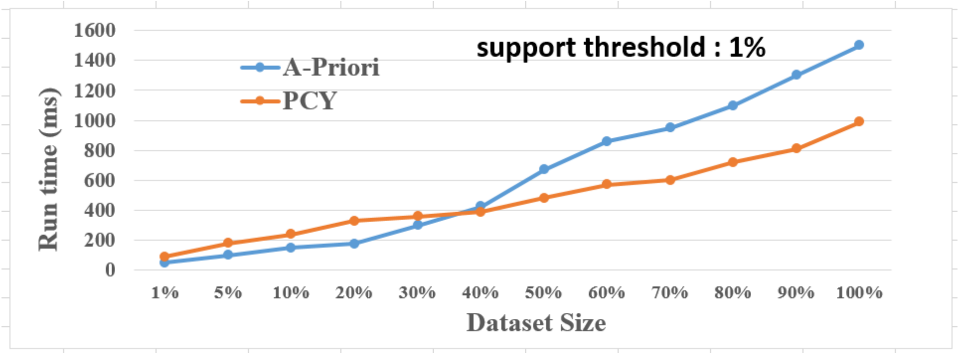
Resources
See lectures 3, 4, 5, and 6 on Blackboard. See chapter 6 of the textbook.

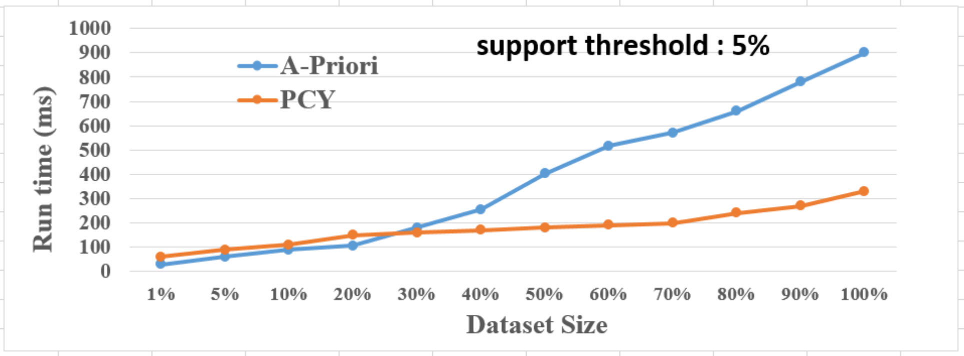
Programming Language
You can choose your favorite programming language (C, C++, Java, C#, and Python etc.)

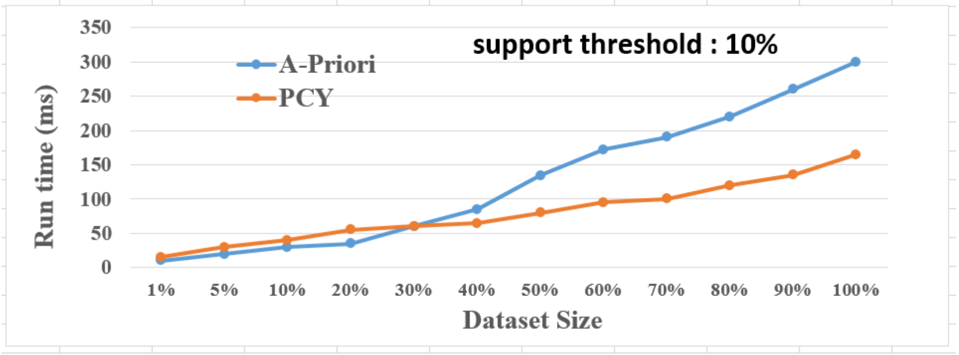
Dataset
The retail dataset contains anonymized retail market basket data (88K baskets) from an anonymous retail store. The preprocessing step to map text labels into integers has already been done. Use Sublime Text, TextPad or Notepad++ or other software to open the file. Do not use Notepad.

Dataset link: *http://mkargar.myweb.cs.uwindsor.ca/retail.txt*

Experiments
Perform the **scalability study** for finding frequent pairs of elements by dividing the dataset into different chunks and measure the time performance. Provide the line chart. Provide results for the following support thresholds: 1%, 5%, 10%. For example, if your chuck is 10% of the dataset, you have around 8,800 baskets. Therefore, if your support threshold is 5%, you should count the pairs that appear in at least 440 baskets. See three samples below for three different support thresholds. Note, the sample charts contain hypothetical numbers!







Optional (Bonus Points)

- Implement Multistage (3 Passes) version of PCY, using one extra hashtable (0.25% extra). (add the results to the line chart)
- Implement Multihash version of PCY, using one extra hashtable (0.25% extra). (add the results to the line chart)

Submission

You have to submit your **code**, along with the **experiments** via email to the graduate assistant Hetal Rajpura (**rajpurah@uwindsor.ca**) before the deadline. Indicate the specification of the machine that you run the experiments on, including the operating system, CPU, and RAM.