60-475 Selected Topics
Big Data Analytics

Project I (15%)
Mining Frequent Itemsets

Deadline: Wednesday February 15th 2017 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 instructor via email no later than Friday January 27th at 12:00pm. After this date, everyone else is assumed to 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 2, 3, 4, and 5 on Blackboard. See chapter 6 of the textbook.

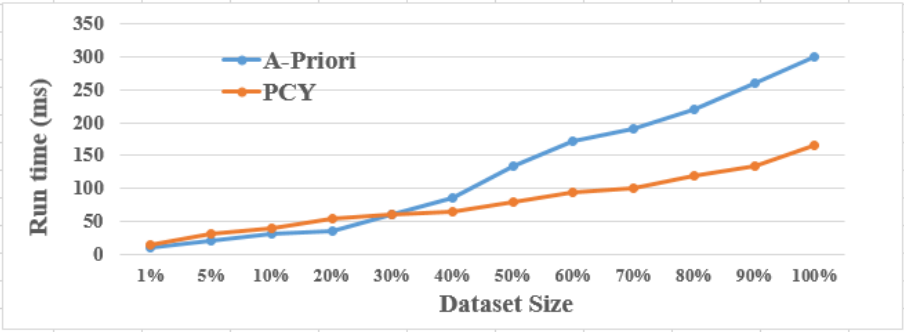
Programming Language
You can choose your favorite programming language (C, C++, Java, C#, and Python etc.)

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. Try different thresholds, e.g. 1%, 5%, 10%, etc. See a sample below. Note, the sample contains hypothetical numbers!

	A-Priori	PCY
1%	10	15
5%	20	30
10%	30	40
20%	35	55
30%	60	60
40%	85	65
50%	134	80
60%	172	95
70%	190	100
80%	220	120
90%	260	135
100%	300	165



Optional (Bonus Points)

- Implement Multistage (3 Passes) version of PCY, using one extra hashtable (1.5% extra). (add the results to the line chart)
- Implement Multihash version of PCY, using one extra hashtable (1.5% extra). (add the results to the line chart)

Submission

You have to submit your **code**, along with the **experiments** via email to the instructor before the deadline. Indicate the specification of the machine that you run the experiments on, including the operating system, CPU, and RAM.