CSE 4510/5400 Data Mining
Due 5pm, Mar 16
Submit Server: Class = datamining, Assignment = hw3

1. Written assignment (from textbook) [pdf file or hardcopy in class]:
   (a) Algorithm 6.1 counts (Lines 6-11) the support of each candidate itemset because it might not be frequent. Illustrate with an example to demonsrate itemsets from $F_{k-1} \times F_{k-1}$ (generation and pruning) might not be frequent.
   (b) Ch6, Q2, p404
   (c) Ch6, Q7, p406 (Apriori uses $F_{k-1} \times F_{k-1}$)
   (d) CSE 5400 only: Ch6, Q3, p405

2. Programming assignment:
   (a) Implement Algorithms 6.1 (p337), 6.2 and 6.3 (updated version on the slides)
      i. allow *minsup* and *minconf* as user-specified parameters (p330)
      ii. Algorithm 6.1, line 5: use $F_{k-1} \times F_{k-1}$ and for eack k, print:
         - k
         - the number of candidate itemsets generated before pruning
         - the number of candidate itemsets after pruning (C_k)
         - the candidate itemsets after pruning (C_k)
      iii. Algorithm 6.1, add line 12.5: print the frequent itemsets F_k
      iv. CSE 5400 only:
         - allow user-specified options of using $F_{k-1} \times F_{k-1}$ or $F_{k-1} \times F_1$ for finding candidate itemsets
   (b) Print the rules in a format similar to the rules on p331.
   (c) Datasets are on the course website:
      i. Market basket (Table 6.2)
      ii. Contact lens
   (d) Implementation:
      i. The same implementation should be able to handle the two different data sets
      ii. Use C (GNU gcc), C++ (GNU g++), Java (Oracle Java), LISP (CLISP), or Python. If you don't have a preference, use Java since it's more portable.
      iii. Your program preferably runs on code01.fit.edu (linux).
   (e) Submission:
      i. README.txt:
         - what the different files are
         - how to compile and run your program with the two data sets with different *minsup* and *minconf*, and for CSE 5400 with different methods for finding candidate itemsets (preferably on code01.fit.edu).
      ii. source code files