Homework 5 (Due March 21st)

1. Your input datasets are transactional data with the following format:

25 52 164 240 274 328 368 448 538 561 630 687 730 775 825 834

39 120 124 205 401 581 704 814 825 834

35 249 674 712 733 759 854 950

39 422 449 704 825 857 895 937 954 964

15 229 262 283 294 352 381 708 738 766 853 883 966 978

26 104 143 320 569 620 798

7 185 214 350 529 658 682 782 809 849 883 947 970 979

227 390

71 192 208 272 279 280 300 333 496 529 530 597 618 674 675 720 855 914 932

Each line records a transaction (a set of items being purchased together).

Your goal is to compute

1. The co-occurrences of each pair of items, Count (A, B)=# of transactions contains both A and B, and the conditional probability Prob(B|A)=Count(A,B)/Count(A).
2. The co-occurrences of each triple of items, Count (A,B,C) =# of transactions contains both A and B, and the conditional probability Prob(A|B,C)=Count(A,B,C)/Count(B,C)
3. You need implement both the “stripes” approach and the “pair” approach om Hadoop (See slides on MapReduce Algorithm Design and Chapter 3 in Jimmy Lin's textbook). Both stripes and pairs need to be generalized to b.
4. For b, you can assume that the count of pairs are already computed in step and you need to reuse their results (read from the result file produced by step a).

The sample datasets can be found at: <http://fimi.ua.ac.be/data/>

2. Revising your algorithms (in 1) to count the co-occurrences of pairs of words Count(A,B) and the conditional probability Prob(B|A)=Count(A,B)/Count(A).

Your inputs are text files and considering using the following criteria for a unit (corresponding a transaction in 1):

a. each line is a unit

b. each sentence is a unit

c. each paragraph is a unit (Bonus)

Run your algorithms on the Enrone email datasets:

http://www.cs.cmu.edu/~enron/enron\_mail\_20110402.tgz

**Submission**: The source code of six algorithms (stripes and pairs for problem 1 (a& b) and problem2); a simple description on how to run your programs; and a screen shot on running each algorithm. Finally, all your programs shall be tested and run on Amazon EC2, and you need run your program on 4 mappers and 2 reducers.