Homework 6 (Due March 21st)

1. This problem follows the first question in Homework 5.

Your input datasets are transactional data with the following format:

25 52 164 240 274 368 401 448 538 561 630 687 730 775 825 834

39 52 124 205 401 581 704 814 825 834

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), and can be annotated by its line number, such as T1, T2, T3, ….

Your goal is to discover all the transaction pairs (Ti, Tj) whose common items are more than 2 (using Hadoop/MapReduce). For example, T1 and T2 is a good pair as their common items are 3.

Note that the number of transactions can be very large (>=1M).

**Submission**: The source code of the algorithm; a simple description on how to run your programs; and a screen shot on running each algorithm. Finally, your program shall be tested and run on Amazon EC2.

Hit: Please refer to this paper: http://www.csee.ogi.edu/~zak/cs506-pslc/docsim.pdf