Assignment 4

Discussion Questions (5 points)

1. How might we be able to evaluate the quality of a supervised learning algorithm (i.e. – recommender or classifier)?
    a. What are some potential strategies to improve the quality of results generated?
2. As mentioned previously, Mahout was borne out of the Apache Lucene search engine project before coming its own top-level project. What application would machine-learning techniques have in the area of search, and what unique implications does the problem of search pose for machine-learning?

Assignment (Total 15 points)

The Resources section contains a zip file, dating.zip, with 2 CSVs containing datasets from a dating agency:

The ratings.csv file includes the fields: UserID, ProfileID, Rating

- UserID is the user who provided rating
- ProfileID is user who has been rated
- Ratings are on a 1-10 scale where 10 is best (integer ratings only) scores

The users.csv file contains a list of users with fields: UserID

(5 points)

Using the item-based recommender with the Pearson Correlation similarity metric, generate 2 recommendations for User IDs 1 through 5. Copy and submit any commands used, and the results. (NOTE: You can use Mahout in Local Mode, if you prefer.)

(10 points)

Build a spam filter using a Naive Bayes classifier with categories "spam" and "ham". Use the SpamAssassin corpus at: http://spamassassin.apache.org/publiccorpus/, specifically the 20021010_spam_tar.bz2 and 20021010_easy_ham.tar.bz2 corpora for your training and test data.

HINT: You should start by unpacking the spam and easy_ham directories within a containing directory, which will serve as the input for the seqdirectory tool. You can use a 40% split for the training/testing data split.

Submit all commands used, output summary, and confusion matrix.