**Assignment 1 Report**

Hongjian Cui, U08398995

**1 Read prediction**

**1.1 Approach**

The main method I use for read prediction is collaborative filtering to get the preference of users and books. Choosing Logistic Regression model to train this task. Because the test set has 50% read books, I balance the result by adjust the larger proportion by ranking its probability of logistic model.

**1.2 Feature Design**

**Book avg rating**: Predict book’s average rating in training set subtract all books’ average rating.

**Book popularity**: Number of readers of predict book / Maximum number of readers of all books.

**User activity**: Number of readers of predict book / Maximum number of readers of all books.

**User avg Jaccard similarity**: Average Jaccard similarity between predict reader and readers who have read predict book. ***(The same as feature: user avg Pearson similarity)***

**Book avg Jaccard similarity**: Average Jaccard similarity between predict book and books the predict reader has read. ***(The same as feature: book avg Pearson similarity)***

**2 Category prediction**

**2.1 Approach**

The main method for category prediction is computing the TF-IDF matrix and choosing the number of word features. Choosing Logistic Regression Model to train this task.

**2.2 Feature Design**