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

$$\begin{aligned} read\ book\ \cong\ & \theta_1 + \theta_2 \times [book\ avg\ rating] + \theta_3 \times [book\ popularity] \\ & + \theta_4 \times [user\ activity] + \theta_5 \times [user\ avg\ Jaccard\ similarity] \\ & + \theta_6 \times [book\ avg\ Jaccard\ similarity] \\ & + \theta_7 \times [user\ avg\ Pearson\ similarity] \\ & + \theta_8 \times [book\ avg\ Pearson\ similarity] \end{aligned}$$

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 user avg Pearson similarity)

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

$$category \cong \alpha + \sum_{w \in text} tfidf(w, d, D) \cdot \theta_w$$