Yu-Chun Chen

Professor McAuley

Report for Assignment 1

Task 1

In this task, we are given the information of user id and book id pair, along with the rating for the book from the user in training set. At first, I planned to utilize both jaccard similarity given user-book pairs and pearson similarity given ratings. However, these two similarities are highly correlated, and would thus result in multicollinearity. Then, I decided to optimize jaccard similarity algorithm with a validation set of size 20000, obtained from the last 10000 pairs from the training set and generating negative rows from them. The algorithm I implemented is as follows: for each user-book pair, such as (u,b), in the test set, I found all the other pending books for this user in the test set, naming it list PB. Then, for each book in this list PB, I computed the jaccard similarity of users in the training set who have read b and the union set of users in the training set who have read either one of the books that u has read. After all jaccard similarity values in the list PB for user u is calculated, 1 is assigned to every user-book pairs that have higher-than-median jaccard values within PB, and 0 is assigned to every other books within PB for user u. In this way, we can produce a balanced prediction in which half of the predictions are positive and other half negative, as this piece of information was previously given.

Task 2

In this task, we are given the information of user id, review text, rating, and review text. I decided to extract features from user id and review text. User ids are one-hot encoded, and review texts are transformed into lower case, punctuation and stopwords stripped before creating a TF-IDF matrix with tfidfvectorizor. Then, Pipeline and FeatureUnion are used to combine one-hot encoded user ids and tfidf matrix from review text with logistic regression to train data. Although bigram approach was attempted during training, accuracy has rather decreased, and thus I continued with only unigram analysis. After using a validation set of the last 10000 data inputs from the training set to tune parameters, C=15 is found to be the most accurate parameter for logistic regression model in this case.