SOFTWARE ENGINEERING THEORY PROJECT

ASSIGNMENT III

Members:

Palak Singhal *16CO129*

Sharanya Kamath *16CO140*

Topic: Book Recommendation for Library Management System

Implementation:

In pattern recognition, the ***k*-nearest neighbors algorithm** (***k*-NN**) is a non-parametric method used for classification and regression.

- In our project we will use this algorithm to find clusters of similar users based on common book ratings, and make predictions using the average rating of top-k nearest neighbours.
- We will only consider popular books from the dataset. This is to ensure statistical significance.
- For finding out these particular books, we will consider both books data and ratings data. We will combine these two and those books with the highest total rating score will be deemed as popular.
- For improving computing speed, and avoid running out of storage space, we will limit our users data to those in the India.
- The user will search a book of his/her choice and the algorithm operating on the above described dataset will display options to the user based on the preferences of other users who searched similar books.
- Fuzzy matching will be added to make it more user friendly. Fuzzy matching allows the user to input misspelled book name and still get recommendation instead of the exact book title.
- K nearest neighbors takes extremely large amount of time when run on a massive dataset. To overcome this locality sensitive hashing technique will be used.
- After doing all these implementation techniques, the final aim would be to compare the accuracy and computational time in each case.

Languages:

Python, Numpy

Frameworks:

Pandas, Scikit-learn

Dataset details

- Book-Crossings is a book ratings dataset compiled by Cai-Nicolas Ziegler.
- This dataset has 11 lakh ratings of over 2 lakh books by 90,000 clients.
- The rating scale is from 1 to 10. The dataset contains 3 tables: book rating, book information, and user information.
- Book rating data set: It consists of a list of book ratings that different users have provided. It includes 11.5 lakh records and 3 fields: book Rating, user ID and ISBN. We observed that the distribution of book ratings is highly uneven, and majority of the ratings are zero.
- Book data set: consists of book details. It has over 2 lakh records and 8 fields: book title, book author, ISBN, publisher, etc.
- User data set: It provides the user demographic information. It has over 2 lakh records and 3 fields: user id, age and location. We observed that most active users are among the age group of 20–30s.

Work Completed:

- We observed the dataset by plotting graphs of its various attributes. The observations are noted in the dataset details section.
- We shortlisted the popular books from the dataset based on the total rating count.
- We used NearestNeighbor function of scikit learn and predicted books which the user might like based on the present search.
- We implemented knn algorithm in a separate python file using a simple dataset of flower species.

Work to be Completed:

- Adding knn implementation for Book-Crossings dataset.
- Making modifications in the knn algorithm to make it more efficient
- Apply locality sensitive hashing technique to knn
- Comparison of all the above mentioned techniques based on time and efficiency of the model.