Project 2 - Recommendation Engine

Problem Statement:

So, have you ever wondered which book to read next? Well, I often have and to me, book recommendations are a fascinating issue. And that is exactly what we are going to do today
So, our dataset comprises of 4 files->

a. Ratings.csv
b. Books.csv
c. Book_tags.csv
d. tags.csv

So, as the name suggests **ratings.csv** contains all users' ratings of the books. There are a total of 980k ratings, for 10,000 books, from 53,424 users. While **books.csv** contains more information on the books such as author, publication year, book_id & so on. Then, we have the '**book_tags.csv'** file. So, this file comprises of all tag_ids users have assigned to the books and corresponding tag_counts. So, the tag_id's basically denote the categories into which the books fall into. And the counts denote the number of books belonging to each category. And we have the '**tags.csv'** file. This file contains all the tag_names corresponding to the tag_ids. i.e, it gives out the labels corresponding to different tag_id's

Tasks to be performed:

1) In the first phase, we'd do a bit of data cleaning.

a. So, we'll start off by removing the duplicate ratings. i.e., there are cases where a user has rated the same book more than one time. So, we'll go ahead & remove all these instances.
b. After which, we'll go ahead & remove those users who have rated fewer than 3 books

2) In the second phase we'll do some data exploration

a. We'll start off by extracting a sample set of 2% records from the entire dataset.
b. Then, we will make a bar-plot for the distribution of ratings. i.e we'd want to analyze the count of different ratings.
c. After which, we'll make a plot to understand how many times each book has been rated.
d. Then, we'll make a plot for the percentage distribution of different 'genres'.
e. Going ahead, we'll find the top 10 books with highest ratings.
f. And finally, we'll find out the 10 most popular books

3) In the 3rd phase, we'll finally do some recommending!!!!

a. So, we'll start off by building the 'user-based collaborative filtering' model.
b. Then, we'll recommend 6 new books for two different readers