OUTLINE FOR FINAL PROJECT

TEAM MEMBERS

Name	GitHub	UNI
Tejas Dharamsi	https://github.com/Dharamsitejas	td2520
Abhay S Pawar	https://github.com/abhayspawar	asp2197
Janak A Jain	https://github.com/janakajain	jaj2186
Vijayraghavan Balaji	https://github.com/vijaybalaji30	vb2428

We will be working on these four things to improve our current models and recommendations.

DATA

We are currently using Book Crossing dataset to build our models. This data is extremely sparse (>99.9% sparsity). Hence, we are looking to use other datasets like Amazon, Goodread, etc. by themselves or by making a larger dataset after combining all these together.

FEATURE ENGINEERING

1. We still haven't used all the information that is there in the Book crossing dataset.
    a. **User features**: Dataset contains age of the user which can be used. We can get genre of the book using the ISBN and have features like his average rating for each genre, number of books he has read in each genre, etc.

b. **Book Features**: We will try using word2vec embedding of the book name and synopsis. We will also try word occurrence features for some frequent words. We can also use their ISBN to get their ratings, author, etc. and use these as features

BETTER MODELS

We will be exploring various modeling methodologies in this part. We will explore the models not only from training time and performance perspective, but also from efficiency in production environment perspective. We will largely explore the methodologies like FMs, approximate nearest neighbors, LSH, etc. discussed in class using standard implementations and implement a couple of these from scratch ourselves. We will compare the results of these. We are also planning on using some deep learning models like Neural Factorization Machines which can model non-linear interactions between features unlike FMs.

USING IMPLICIT FEEDBACK

This dataset also contains information about if a user read a book (and didn't rate it). We will try to incorporate this implicit feedback into our models as well. Easiest way is to assign a mean rating for these read books. We will explore other methodologies.

BETTER TOOLS

Till now we have been using python and related libraries for our work. For the final project, we will explore Scala and Spark.