**Proposal of Modeling Approach**

* Content-based filtering:
  + To generate recommendations, each book is assigned a feature vector, where the “features” are book genres, keywords from the book description, etc. For a given book, the component of its feature vector corresponding to each feature is a Boolean value {0,1}, depending on whether the book has that particular feature.
  + The similarity between two books is measured by their cosine similarity, meaning the cosine of the angle between their feature vectors, which ranges from [0,1], with 1 indicating the highest possible similarity.
  + For a user-specified list of books with ratings out of 5 stars, a weighted sum/average of their feature vectors is generated, weighted by the user ratings of each book. The cosine similarity between the resulting vector and each feature vector in the dataset of books is then computed, and the books with the highest similarities are then reported as recommendations.
  + Hit testing can be used to measure the quality of the resulting recommendations.
* Collaborative filtering:
  + To generate recommendations we factor a matrix Y of users' book ratings into two matrices: X, which represents hidden "features" of books and W, which represents users' preferences for those features. In our algorithm we also include a bias vector, b.
  + We compute these matrices by starting with random values and using gradient descent to make the product (X\*W+b) approximate the known ratings.
  + To recommend books for a specific user, we take the learned book features (X) and, using the user's past ratings, employ linear regression to find their unique preferences (W and b). Multiplying X by this user's W and adding b predicts their ratings for all unread books. Then, we recommend the highest rated ones.
  + We can evaluate our recommendations by using hit testing (how often recommended books are chosen) and cosine similarity (how closely predicted ratings match actual ratings).
* K-means clustering:
  + To speed up our recommendation algorithms, we plan to first use K-means to cluster the books based on the “features” we computed to initialize content based filtering.
  + Thus, for new users, we may first use the books they’ve already rated to see which clusters of books they may be interested in, and then only compute their ratings/similarities for said books.
* Finalizing Recommendations:
  + To finalize our recommendations, we take the recommendation lists generated by both the content based and collaborative filtering based algorithm, and combine them based on a user chosen ratio.