Amanda Wang, Jennifer Jasperse, Arely Alcantara (NetIDs: afw3, jbj3, arely2)
Team: Push Mode People

Progress Report

Since writing our proposal, we have been busy at work building out our book recommender system using the GoodReads datasets. We have also been trying to document our code as we go using Jupyter notebooks.

Based on the feedback that we received for our proposal, we realize that building a hybrid model is too much given the scope and time for this project, so we will not be implementing it.

Below we will list our progress, remaining tasks, and challenges faced.

Tasks in progress:

- We started out by cleaning out the books and reviews dataset - focusing on English entries only. This took a bit longer than the anticipated 2 hours because each of us designed our own data-cleaning methods and due to the volume of processing required. We have completed preliminary cleaning on subsets of the data that are easier to work with and the full datasets - 15M reviews and 2.36M books.
- We have researched and implemented a very simple content-based recommender where we take a book and generate the top 7 related books to recommend - we're hoping to flesh this out and incorporate some user data.
- We were able to research, build, and evaluate a preliminary user-based collaborative filtering recommender system that uses a KNN model. This already took approximately 20 hours.

Tasks Remaining:

- Research and implement a web search query (Task 1)
- Evaluate Task 1
- Test, debug, and revise code for Task 1
- Evaluate Task 2
- Test, debug, and revise code for Task 2
- Demo/final presentation

Challenges:

- A major challenge faced while designing the collaborative filtering model was the limitation on the size of data we could work with. Because we are using Jupyter notebooks, there is a size limit for the user-similarity matrix. As a workaround, we are currently just using datasets of ratings for specific genres, specifically poetry, which has fewer ratings and users [154,555 reviews and 36,636 users]. We will try to capture the user similarity information for the larger datasets using sparse matrices instead.
- Someone also mentioned in the peer reviews whether we are using the entire dataset or just a subset so we are still debating on the best approach - we were also wondering if the dataset must be in the repo to facilitate setup and testing for the reviewers later on as we have had technical difficulties in pushing the full datasets to our repository. The full datasets are also much more difficult to train efficiently.

Question:

- Is the technical documentation supposed to be separate from the code or is it ok if we use Jupyter notebooks for the code + documentation blended together?