Project Proposal

Dataset & Business Use Case

Dataset - Goodreads dataset (https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home)

The dataset contains - (1) meta-data of the books, (2) user-book interactions (users' public shelves) and (3) users' detailed book reviews. These datasets can be merged together by matching book/user/review ids. The entire dataset contains -

- 2,360,655 books (1,521,962 works, 400,390 book series, 829,529 authors);
- 876,145 users;
- 228,648,342 user-book interactions in users' shelves (include 112,131,203 reads and 104,551,549 ratings).

For the purpose of this project, I propose to filter the dataset by genre of book pertaining to 'Poetry', as the entire original dataset is too huge.

Business use case

Goodreads derives its revenues by promoting book campaigns, where it works with major publishers to promote titles. The company's data services mainly center on an open API for utilizing its book data on third-party sites. The company today probably has a better idea of success indicators for books than anyone. Data from Goodreads could be utilized to understand how book characteristics, writers and other factors perform among certain reader sets. It would be interesting to apply the machine learning concepts to perform this task. But, graph networks draw more insights from the data that are hidden to regular machine learning techniques.

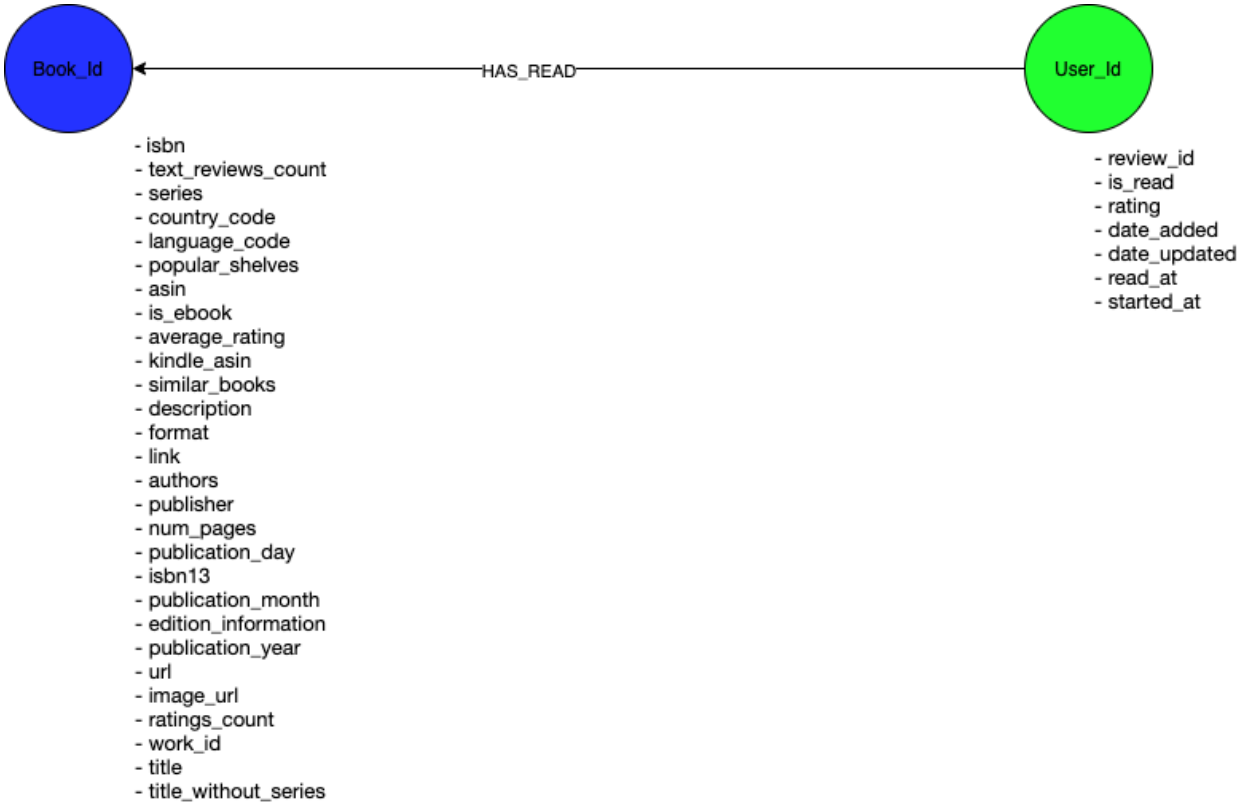
So, the business use case here is -
For the analytics team of the book-based social network platform Goodreads: Use graph analytics and build implicit book-user graphs:

- to visualize and analyze networks of interactions between users, books, and authors via their reading preferences [graph visualizations]
- to identify and recommend books to users by engineering network features [recommender systems using graph machine learning]
- to identify influential users and user groups (communities) on the Goodreads social network [graph algorithms]

Essentially, I will work on designing a network based recommender system for the Goodreads platform. Furthermore, I will build a graph database and present results of the above use cases via visualization to the stakeholders.

Having a good recommender system can be of great value to the company to retain its users as better users provide better reviews which the publishing companies are looking for.

Graph Data Model



Graph Projections

The two mono-partite graphs are as below.

For the first, User_Id - two users are said to be connected if they have read (shared) the same books. Number of shares becomes the weight here.

Second, Book_Id - two books are said to be connected if they have the same users and the number of users that read the book becomes the weight. Say if user 1 and user user 2 have read that book, the weight becomes 2 here.