Project Proposal

Team Name: Maverick

Team Members
Kavya Balihallimatta: 013819923
Kruti Thukral: 012586041
Sahana Ramesh: 013832065

Project Idea: ***Book Recommendation system using Goodreads book dataset***
The main idea of our project is to design a personalized book recommendation system. We can use explicit ranking and other available features and metadata to derive additional information. Pre-processing would be an important step as there are 4 datasets which need to be combined. Some of the algorithms that can be evaluated and compared are as below:

- Word2Vec based recommendation
    - Most of the time there is a pattern in the reading behavior of the readers. If we have access to the shelf history of the reader, we can maintain the sequence of books. We use a neural network to transform these book sequences into vectors/mathematical objects. It will then be possible to conduct mathematical operations on the books in the sequence to make book recommendations.
- Content-based recommendation
    - We would use the content-based recommendation algorithm to predict the top books the reader might be interested in. The content-based recommendation takes the following metadata into consideration, for predicting the top items a reader might be interested in.
        - Authors
        - Popular_shelves
        - Languages
        - Country codes
    - We will use TF-IDF on the popular shelves to figure out which shelves best describe the book. The shelves can then be mapped to features in the item profile. Similarly, for the user profile, we will use TF-IDF for the shelves as per the book history of the reader.
    - An item profile can be created using the metadata for the books and a user profile can be created using the metadata for the books the user has read before. Similarity between the two profiles can be used to make recommendations.

- Combining collaborative filtering and sentiment classification of review text for improved book recommendation

- In this hybrid model, collaborative filtering will carry out the first level filtering and the sentiment classifier will perform the second level of filtering. The final recommendation list is a more accurate and focused set.

Data Set: We would be making use of Goodreads book datasets as it is a rich dataset with many features. This dataset spans across multiple genres however for our purpose we would be looking at one particular genre "comics and graphics".
Sources:

- https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home (main link)
- https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home#h.p_evDuwuTozQVZ (comics genre)

The following are the datasets and their features which will be combined to get an integrated dataset. The statistics provided are for the "comics and graphics" genre.

- **User-Book Interaction dataset (7.3 million entries, 2.51 GB)**
  - User_id - A unique identifier for a user
  - Book_id - A unique identifier for a book. This can be used to look up the corresponding book metadata in the book dataset
  - Review_id - A unique identifier for a review. This can be used to look up the corresponding review information in the review dataset
  - Is_read - A flag indicating whether the user has read the book or not
  - Rating - Numerical rating by the user for the particular book
  - Date_added - Date on which user added the book to the shelf
- **Review dataset (0.55 million entries, 731 MB)**
  - Review_id - A unique identifier for a review
  - N_votes - Count of upvotes for the particular review
  - Review_text - Descriptive review of the book
- **Book dataset(0.08 million entries, 353 MB)**
  - Book_id - A unique identifier for a book
  - Country_code - Country code for the publisher of the book
  - Language_code - Language of the book
  - Description - Elaborate summary of the book
  - Format - the Format of the book such as link or paperback
  - Publication_year - Year in which book was published
  - Title - The Title of the book
  - Author_id - A unique identifier for an author. This can be used to look up the corresponding author information in the author dataset
- **Author dataset (0.82 million entries, 105 MB)**
  - Author_id - A unique identifier for an author
  - Name - Name of the author