CIT550 Project Milestone 1 - Project proposal
TeamName: Team A+

---

1. Group Information

Name:Hongri Jia
Email:hongri@seas.upenn.edu
Github Username：henry208j

Name:Yazhuo Wang
Email:yzalicew@seas.upenn.edu
Github Username：alice-yz-wang

Name:Jiameng Chen
Email:cjiameng@seas.upenn.edu
Github Username: chjm23

Name:Yi Cao
Email:yc3136@seas.upenn.edu
Github Username：yc3136

2. Description of application idea

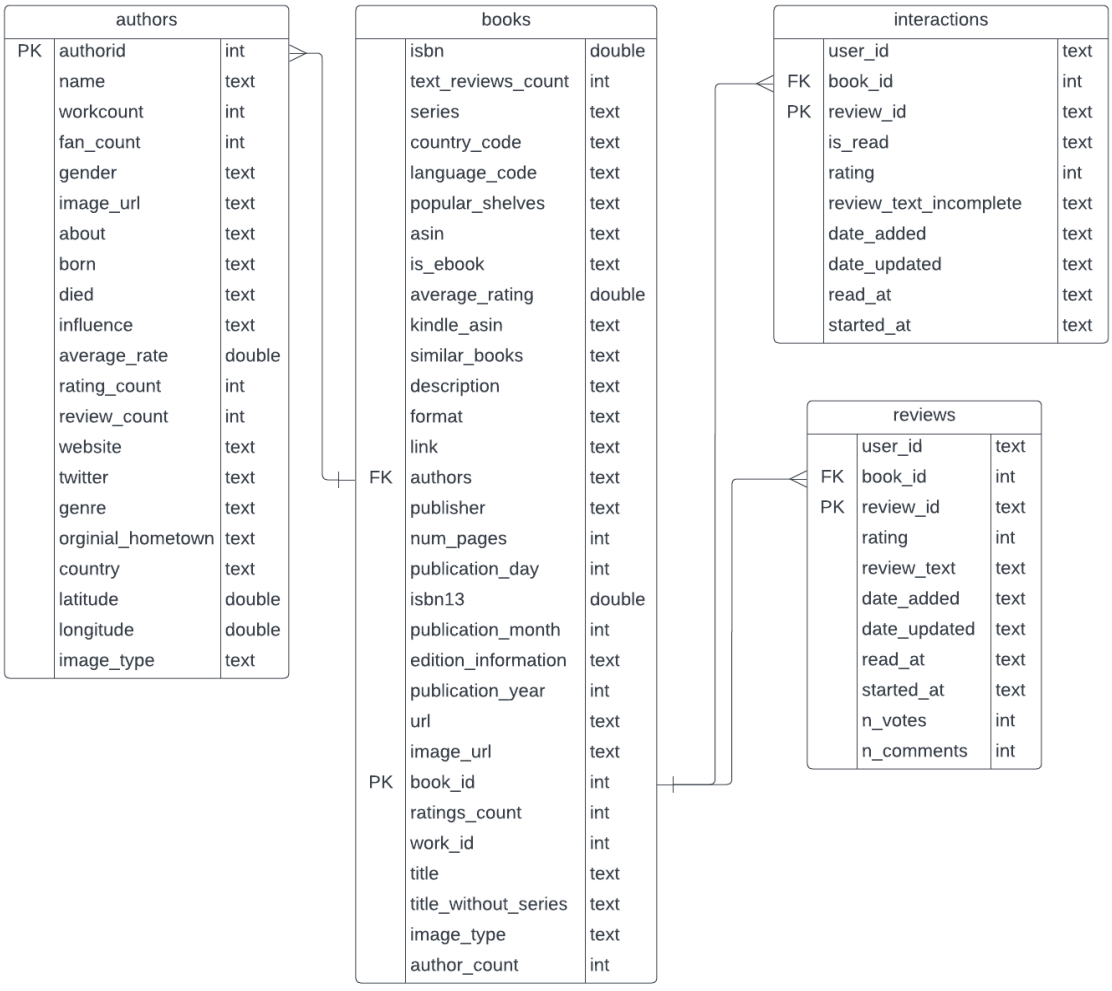
Upon emergence in the 1930s, comics have been a major part of pop culture. For our generation, no childhood memory is complete without the Captain, Batman and all the superheroes. The goal of our project is to build a book catalog to document, analyze and give people their customized recommendation of comic memory.

Given how nowadays, we tend to live in a bubble where all the information we receive is what we tend to like and believe, we specifically added the function "surprise me" to encourage our readers to break the bubble and embrace the uncertainty. Inspired by a bookstore that covers books entirely to surprise their readers, we plan to tweek the query of recommendation to make the surprise novel yet agreeable.

We plan to design three pages, homePage, booksPage, and authorsPage:

- The homePage is the main page, which will include an introduction of our project and the "surprise me" function.
- The booksPage lists all comic books. Web users could sort books by title, review counts, average rating, and etc., and filter by language, publication year, average rating, and etc. When a book is selected, detailed information about that book will be listed in a new page, which will include book description, reviews, similar books, and etc.
- The authorsPage lists all authors who write comic books. Web users could sort authors by name, counts of books written, average rate, and etc, and filter by gender, hometown, and etc. When an author is selected, detailed information about that author will be listed in a new page, which will include books written by the author, author's social media links, and etc.

3. Dataset



Datasets Relationship Diagram

- **Data Source 1:**
  https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home#h.p_evDuwuTozQVZ
  (Book Metadata, User interaction, Review Data)

- **Data Source 2:**
  https://www.kaggle.com/datasets/choobani/goodread-authors
  (Author information)

Datasets 1: Book Metadata (16,288 entries x 31 columns)

- **Description:**
  This dataset includes the meta-data of 89,411 comic books. After data processing and cleaning we have 16,288 entries. Each entry of book metadata includes book name, isbn, popular_shelves, avg rating, book description, and etc.

- **Sample entry:**


```
{'isbn': '',
'text_reviews_count': '7',
'series': ['189911'],
'country_code': 'US',
'language_code': 'eng',
'popular_shelves': [{'count': '58', 'name': 'to-read'},
                    {'count': '15', 'name': 'fantasy'}],
'asin': 'B00071IKUY',
'is_ebook': 'false',
'average_rating': '4.03',
'kindle_asin': '',
'similar_books': ['19997', '828466', '1569323', '425389'],
'description': 'Omnibus book club ed……',
'format': 'Hardcover',
'link': 'https://www.goodreads.com/book/show/7327624-the-
unsch……',
'authors': [{'author_id': '10333', 'role': ''}],
'publisher': 'Nelson Doubleday, Inc.',
'num_pages': '600',
'publication_day': '',
'isbn13': '',
'publication_month': '',
'edition_information': 'Book Club Edition',
```

'publication_year': '1987',
'url': 'https://www.goodreads.com/book/show/7327624-the-unsc……',
'image_url':'https://images.gr-assets.com/books/13…...',
'book_id': '7327624',
'ratings_count': '140',
'work_id': '8948723',
'title': 'The Unschooled Wizard (Sun Wolf and Starhawk, #1-2)',
'title_without_series': 'The Unschooled Wizard (Sun Wolf a……'}
(UCSD Book Graph - Books, n.d.)

Datasets 2: Author information (1,920 entries x 21 columns)

- **Description:**
  This dataset includes the author information of 209,500 authors on Goodreads. After data processing and filtering of comic books authors we have 1,920 entries in total. Each entry includes author name, number of work, image, their review and ratings and some of their personal background information, etc.

- **Sample entry:**

{"authorid":{"0":8409092},
"name":{"0":"Jason Wallace"},
"Workcount":{"0":2},
"Fan_count":{"0":13},
"Gender":{"0":"male"},
"image_url":{"0":"https:\/\/images.gr-assets.com\/authors……"},
"about":{"0":"Jason Wallace is related to Tolkien and a ……"},
"Born":{"0":null},
"Died":{"0":null},
"Influence":{"0":null},
"Average_rate":{"0":3.74},
"Rating_count":{"0":1028},
"Review_count":{"0":175},
"Website":{"0":null},
"Twitter":{"0":null}
,"genre":{"0":null},
"original_hometown":{"0":"Cheltenham"},
"Country":{"0":
"United Kingdom"},
"Latitude":{"0":51.90006},
"longitude":{"0":-2.07972}}

Datasets 3: Review Data (108,323 entries x 21 columns)

- **Description:**
  This dataset includes 542,338 detailed reviews. After data cleaning we have 108,323 entries. Each entry includes the book id, review text and timestamp to help us better evaluate reader's preference of comic books on natural language level.

- **Sample entry:**


```
{"user_id": "dc3763cdb9b2cae805882878eebb6a32",
"book_id": "18471619",
"review_id": "66b2ba840f9bd36d6d27f46136fe4772",
"rating": 3,
"review_text": "Sherlock Holmes and the Vampires ……",
"date_added": "Thu Dec 05 10:44:25 -0800 2013",
"date_updated": "Thu Dec 05 10:45:15 -0800 2013",
"read_at": "Tue Nov 05 00:00:00 -0800 2013",
"started_at": "",
"n_votes": 0,
"n_comments": 0}
```


(UCSD Book Graph - Book Reviews, n.d.)

Datasets 4: User Interaction Data (1,426,160 entries x 21 columns)

- **Description:**
  This dataset includes 7,347,630 entries of user-book interactions. After data cleaning we have 1,426,160 entries. Each entry includes the book id, user_id, rating, and timestamp.

- **Sample entry:**


```
{'user_id': '8842281e1d1347389f2ab93d60773d4d',
'book_id': '25735618',
'review_id': 'ea74f2b6645b7d16f3ede2aca10226f0',
'is_read': True,
'rating': 0,
'date_added': 'Fri Aug 25 13:55:10 -0700 2017',
'date_updated': 'Tue Oct 17 23:53:44 -0700 2017',
'read_at': '',
'started_at': 'Tue Oct 17 09:23:10 -0700 2017'}
```


(UCSD Book Graph - User-Book Interactions , n.d.)

4. **Queries**:

1. Find popular book written by author:
   Given an author name, list all his/her books, order by numbers of read

   ```
   SELECT books.title, COUNT(interactions.user_id) AS peopleRead
   FROM authors
   JOIN books ON authors.authorid = books.authors
   LEFT JOIN interactions ON interactions.book_id = books.book_id
   WHERE authors.name LIKE '%___%' AND interactions.is_read = 'True'
   GROUP BY books.title
   ORDER BY COUNT(interactions.user_id) desc
   ```

2. Given a book name, list all the reviews of the book order by votes

   ```
   SELECT title, reviews.user_id, reviews.review_text
   FROM reviews
   JOIN books ON reviews.book_id=books.book_id
   WHERE books.title =  ${book_name}
   ORDER BY n_votes
   ```

3. Find books of authors who are from a certain place
   Given a city_name, list all books of authors whose hometown is city_name

   ```
   SELECT books.title, books.authors
   FROM books
   JOIN authors ON books.author_id = authors.author_id
   WHERE authors.original_hometown =  ${city_name}
   ORDER BY books.Average_rating
   ```

4. Given a book_title, list all books that are similar

    - Approach 1: select book in similar_book list
      SELECT books.title, books.author, books.average_rating
      FROM books
      WHERE books.book_id IN (
          SELECT books.simlar_books
          FROM books
          WHERE books.title LIKE '%___%'
)
      ORDER BY books.average_rating DESC

    - Approach 2: Use collaborative filtering
      Find book similarity score base on user interaction
      Then order by similarity score and take top 10

5. Surprise me

    Given a book_title, give a list of books based on the user's past book query or popular shelves without revealing information about title or author, only showing blurred covers.

    - Approach 1: find similar books' similar books that are not originally similar

    - Approach 2: Use collaborative filtering
      Find book similarity score base on user interaction
      Then order by similarity score and take books that are 60% - 70% similar

Works Cited

UCSD Book Graph - Book Reviews. (n.d.). From https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/reviews

UCSD Book Graph - Books. (n.d.). Retrieved from https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/books

UCSD Book Graph - User-Book Interactions . (n.d.). From https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/shelves