Progress Report

Haoyue Bai, Yuqun Wu

1) Progress made so far

We decided to use package scrapy to scrape the book, author information from a book website. Till now, we manage to write a backend crawler for two book websites: Amazon Books, and Goodreads. We first wrote the Amazon Books, but later we found that Goodreads has a better review and recommendation system, so we also modified our code to fit in Goodreads.

Our crawler can extract book name, book's author, similar books, and book's review from the website, and store them into the local MongoDB database. For each book, we would also use its similar books and its author's other books to keep scraping more books.

2) Remaining tasks

We need to implement a backend using flask to access the database and expose API to the frontend. It needs to be integrated with the crawler part.

We still need to design a simple user interface and implement the frontend part. We plan to implement the front end part using React and connect the frontend to backend.

Still, we are planning to implement a recommendation system to suggest users new books which meet their interests.

3) Any challenges/issues being faced.

1. Sometimes we would be blocked because of high frequency visits, so we might need to do some operation to prevent this. We might choose to connect to VPN once blocked, or limit the minimum visiting interval.
2. Currently, we need to manually stop the scraping, but we want our crawler to stop it when it meets some maximum book number.
3. We are currently thinking about what book website we might choose. The advantage of Amazon is that it provides the price of the book, but goodreads has a higher reputation of comments and recommendation among book readers.