Sarah Sahn (sahn10), 4CR
Final Project

Code Narrative

First draft of code: *goodreads fantasy books.ipynb*
Final draft of code with documentation: *"diverse" fantasy books on Goodreads.ipynb*

I began working on the programming part of my project with the idea that I wanted to explore tags (or "shelves") on Goodreads to look at the intersection of fantasy books and books tagged as a "diverse" in some way. After meeting with Elizabeth to discuss some ideas, I began by getting a developer key for the Goodreads API and exploring the Python wrapper for the API. I found from playing around with the API that I could fairly easily pull data given a specific, but there didn't seem to be an easy way to look at shelf data and get book titles from that using the API. The API methods seem to be designed more for developers to allow Goodreads users to interact with the site through third-party apps, rather than to pull information about a lot of books.

Another challenge to this project was the relative opacity of the Goodreads site when it comes to identifying user-generated tags or shelves. On an individual book's page, you can click through to see popular shelves, but the pages for shelves only show a limited amount of data, defined by popularity and release date. However, the API has a method that allows the user to identify "popular shelves" for a given title, and returns a list of the 100 most popular shelves for that book. Given a list of books, then, I could get a data set of popular shelves for each and explore that. (I played with the API in the file *goodreads tests.py*.)

The Goodreads API uses Goodreads ID numbers to identify titles (or possibly also ISBNs, but the documentation didn't make that very clear). I started by using the list of popular fantasy books, and poked around in the source code to develop an XPath query to get the ID numbers out. I came up with one query that worked in XPath Helper in Chrome, but returned an error when I tried to run it in Python because it cast too wide a net. After reworking it, I got a clean dataset. However, since the list of popular fantasy books is relatively short (only about 1200 books), I decided instead to scrape the ID numbers from the results I got when I searched "fantasy" by genre, which turned up about 50k results. The same XPath query worked. A little googling also helped me find a way to pull the data from the webpages instead of downloading them to scrape from.

Up until this point I had been working on my code in PyCharm, but scraping 50k data points takes time, I discovered, so I decided to move the project into a Jupyter Notebook so I could isolate the loop to scrape the data in a single cell. (I originally wrote it the loop I used to scrape in the file *fantasy books.py*. I started to use PyCharm for this but I preferred using the browser interface. All of my code after this is written (and narrated, to some extent!) in the notebook *goodreads fantasy books.ipynb*. With all the book IDs from my search, I then created a for loop in a new cell to pull each book's title and its popular shelves, and populate those items in a dictionary. (This process also took some time to run. I did the dishes while I waited.)

With all of this data, I was then faced with the question of what I wanted to do with it. The remainder of my code (starting from cell 4) is more exploratory. First I needed to get all the lists of popular shelves into one list, and then get a list from there of the unique shelves. That created a bit of a problem because it turned out the API didn't pull the shelf tags as strings, which I discovered when the Counter just counted everything once and didn't detect duplicates, so I had to loop through and make the shelves into strings, then run the Counter on them. In my second draft of my code, I rearranged the cells in my notebook so I could replace the shelf lists in the dictionary with lists of strings.

At this point I was still left with the question of what to do with all that shelf data. What keywords could I use to find "diverse" shelves and how would I identify them? I decided, for the sake of simplicity, to find all the shelves that had "diverse" or "diversity" in them to start. Another option be to create a list with words related to diversity to pull more shelves, since the string "divers" only returned 6 shelves.

With that list of "diversity" shelves, I looped through my entire dictionary and collected a list of titles that were on those shelves, eliminated duplicates with a Counter, and used that list of titles to get the shelves that intersect with "diversity." Given my original 50k+ titles, the list of "diverse titles" was relatively short, but still returned plenty of results to work with, especially given that each title returned a list of 100 shelves.

This is the point where I came to a stop, but I have a few ideas for future directions with this code and research. First, with the list of unique shelves that I got, I could visualize the intersections. I could also seek intersections of a smaller subset of shelves—for example, do readers tag books with female protagonists with other diversity labels? What about the intersection of diversity and young adult? Another direction to take would be to scrape data from Goodreads discussion groups or curated lists. Ultimately, this could lend itself to an analysis of social reading behavior and could aid in discovery of diverse books.

This project resulted in something very different from any of the projects I read about for my annotated bibliography, but the process of exploring the digital humanities as a field through my reading, and exploring the data from Goodreads, had me thinking through how the values of the digital humanities, my own values as a humanities scholar, and the process of writing code and analyzing data intersect. I don't have any real answers, but I do have a lot of questions about how a project like this might fit into my future as an academic or as a librarian, and how I might facilitate further change in how academic institutions understand and value scholarship of all kinds.