Environment Setup

This Web Scraper Application is suggested to run under Python3.5+ environment. (It would be better if it could run with PyCharm)

Required Packages

- BeautifulSoup4
- Pymongo
- Argparse
- Json
- Validators
- Re
- Request
- NetworkX
- D3.js
- Css fontawesome
- ajax

Instructions on User Interface

- **-url [url]**
  - Instruction:
    - -url followed by the starting url users prefer

- Error Messages:
  - When the input url is not a valid url:

```
Dianas-MacBook-Pro-2:pythonProject4 fatpo$ python main.py -url sdkfh
Invalid url! Please enter a new valid url!
```

```
Dianas-MacBook-Pro-2:pythonProject4 fatpo$ python main.py -url https://wiki.illinois.edu/wiki/pages/viewpage.action
?pageId=616243854
Not a GoodReads url! Please enter a GoodReads url!
Dianas-MacBook-Pro-2:pythonProject4 fatpo$
```

  - When the input url is not a GoodReads:

  - When the input url is not a book url:

```
Dianas-MacBook-Pro-2:pythonProject4 fatpo$ python main.py -url https://www.goodreads.com/author/show/6572605.Taylor
_Jenkins_Reid
Not a book url! Please enter a book url!
Dianas-MacBook-Pro-2:pythonProject4 fatpo$
```

- If the input url meets all requirements, this command should be combined with another command(-s) for use.

- **-s [book_number] [author_number]**
  - Instruction:
    - -s followed by the number of book and author to scrape
  - Error Message:
    - When the number of book is too large:

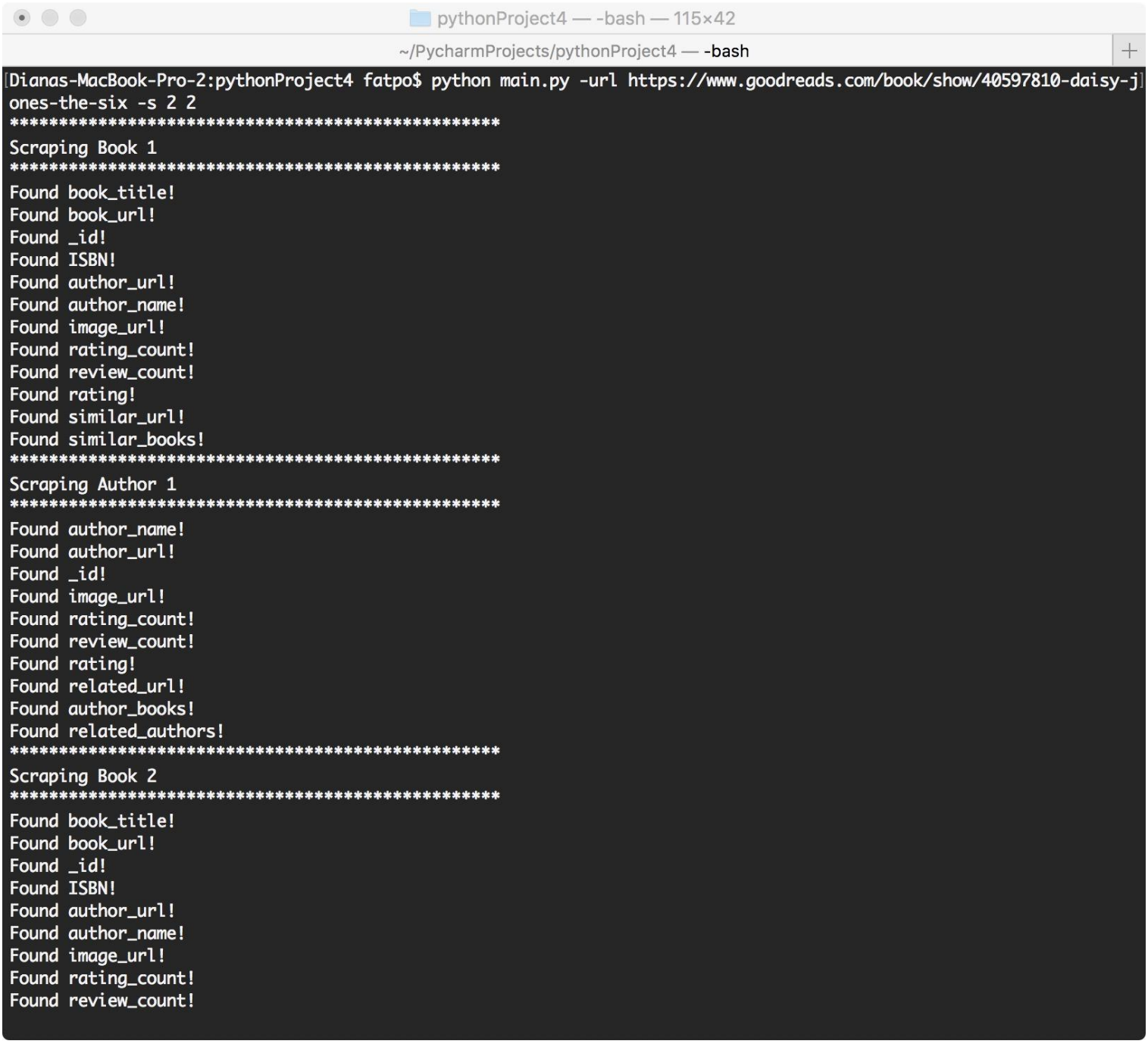
```
Dianas-MacBook-Pro-2:pythonProject4 fatpo$ python main.py -s 300 20
Book number or author number too large! Please enter a valid number!
Dianas-MacBook-Pro-2:pythonProject4 fatpo$
```

    - When the number of author is too large:

```
Dianas-MacBook-Pro-2:pythonProject4 fatpo$ python main.py -s 30 200
Book number or author number too large! Please enter a valid number!
Dianas-MacBook-Pro-2:pythonProject4 fatpo$
```

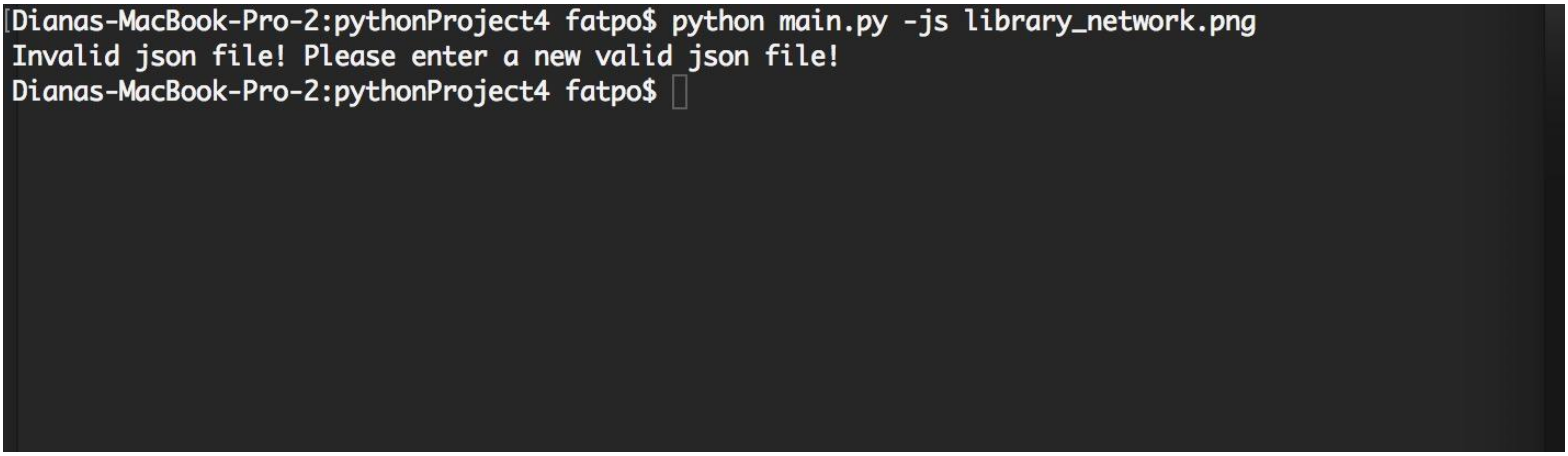
- **-url [url] -s [book_number] [author_number]**
  - Instruction:
    - -url followed by the starting url users prefer, -s followed by the number of book and author to scrape

- Example Run:

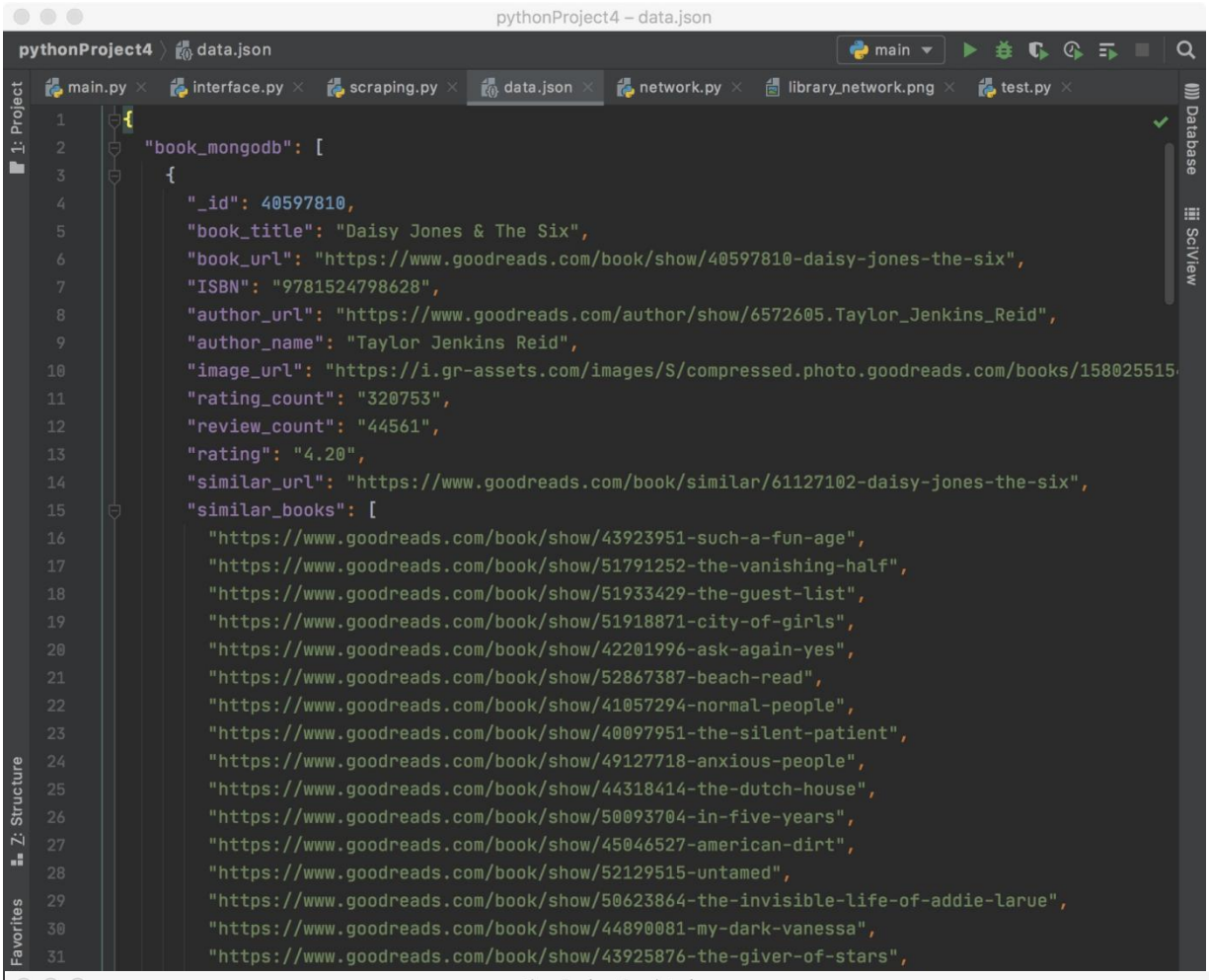


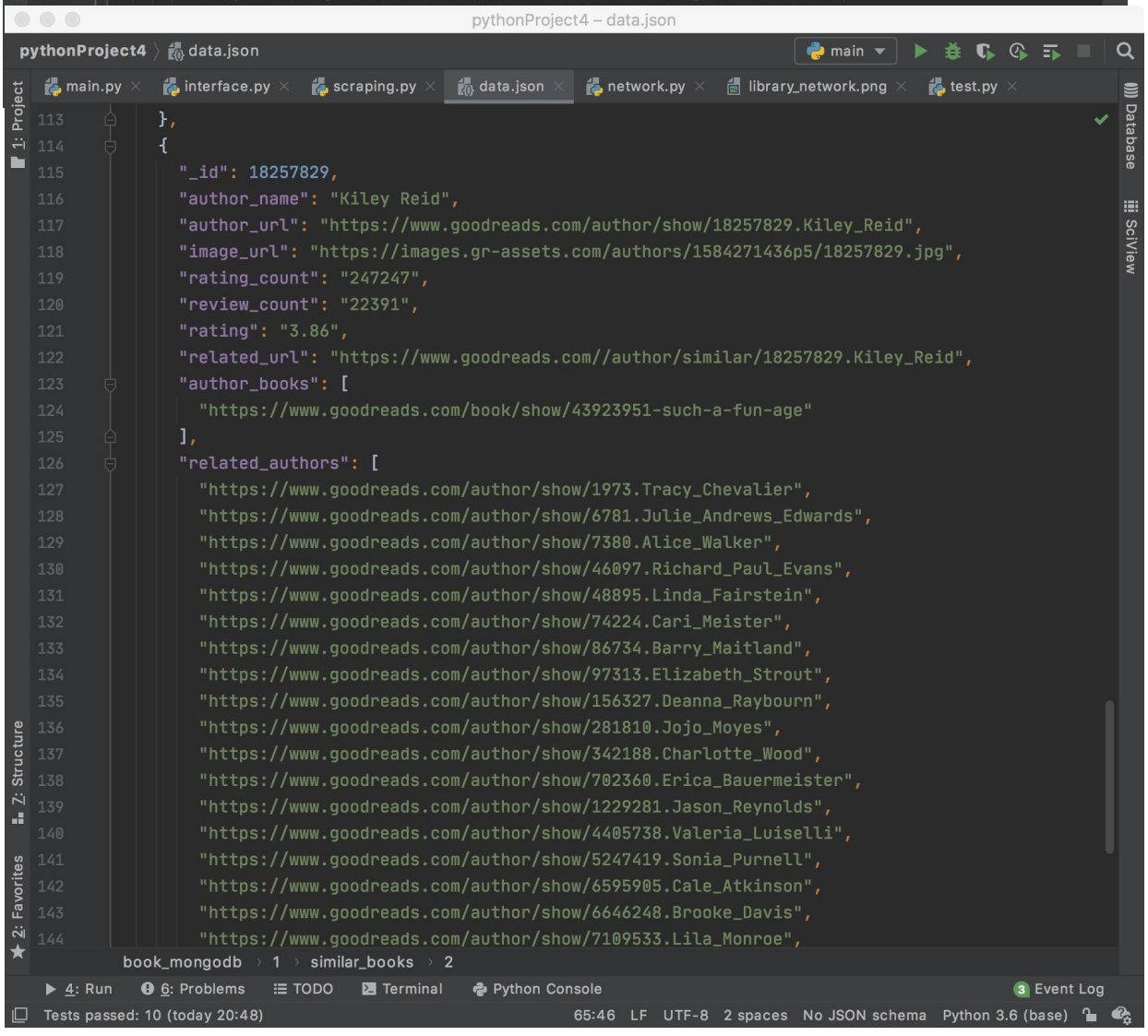
- **-js [json_file_name]**
  - Instruction:
    - -js followed by the json file name that the user want to use to update our database

- Error Message
  - When input file is not a json file



- If the input file is a valid json file(like the picture down below):
  - A valid json file under this situation means a json file with the same structure as that would be exported by this Web Scraper Application

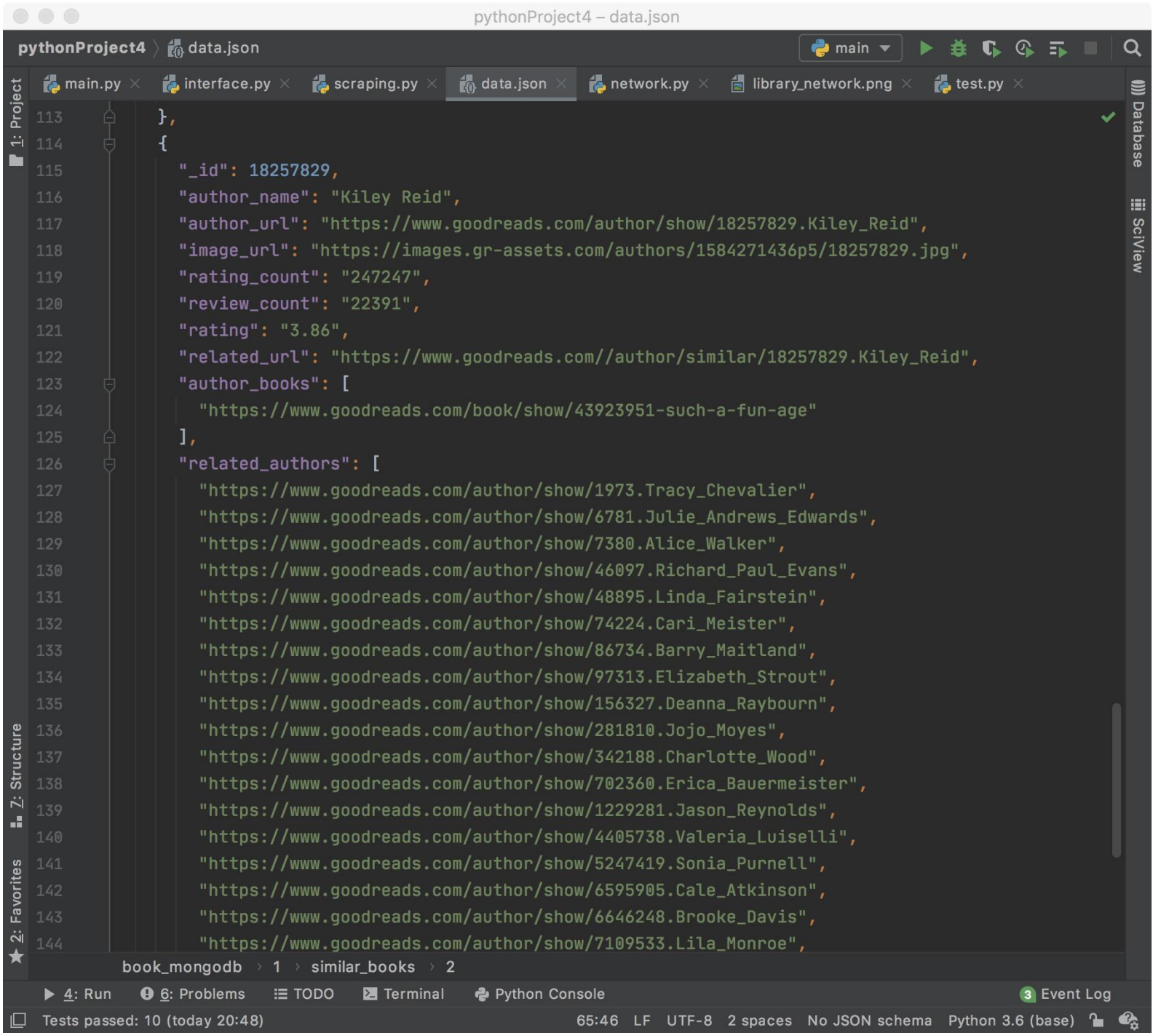

pythonProject4 – data.json
pythonProject4 〉 data.json
main
main.py
interface.py
scraping.py
data.json
network.py
library_network.png
test.py
1: Project
Database
SciView
Z: Structure
Favorites
{
"book_mongodb": [
{
"_id": 40597810,
"book_title": "Daisy Jones & The Six",
"book_url": "https://www.goodreads.com/book/show/40597810-daisy-jones-the-six",
"ISBN": "9781524798628",
"author_url": "https://www.goodreads.com/author/show/6572605.Taylor_Jenkins_Reid",
"author_name": "Taylor Jenkins Reid",
"image_url": "https://i.gr-assets.com/images/S/compressed.photo.goodreads.com/books/158025515
"rating_count": "320753",
"review_count": "44561",
"rating": "4.20",
"similar_url": "https://www.goodreads.com/book/similar/61127102-daisy-jones-the-six",
"similar_books": [
"https://www.goodreads.com/book/show/43923951-such-a-fun-age",
"https://www.goodreads.com/book/show/51791252-the-vanishing-half",
"https://www.goodreads.com/book/show/51933429-the-guest-list",
"https://www.goodreads.com/book/show/51918871-city-of-girls",
"https://www.goodreads.com/book/show/42201996-ask-again-yes",
"https://www.goodreads.com/book/show/52867387-beach-read",
"https://www.goodreads.com/book/show/41057294-normal-people",
"https://www.goodreads.com/book/show/40097951-the-silent-patient",
"https://www.goodreads.com/book/show/49127718-anxious-people",
"https://www.goodreads.com/book/show/44318414-the-dutch-house",
"https://www.goodreads.com/book/show/50093704-in-five-years",
"https://www.goodreads.com/book/show/45046527-american-dirt",
"https://www.goodreads.com/book/show/52129515-untamed",
"https://www.goodreads.com/book/show/50623864-the-invisible-life-of-addie-larue",
"https://www.goodreads.com/book/show/44890081-my-dark-vanessa",
"https://www.goodreads.com/book/show/43925876-the-giver-of-stars",


pythonProject4 – data.json
pythonProject4 〉 data.json
main
main.py
interface.py
scraping.py
data.json
network.py
library_network.png
test.py
1: Project
Database
SciView
Z: Structure
2: Favorites
},
{
"_id": 18257829,
"author_name": "Kiley Reid",
"author_url": "https://www.goodreads.com/author/show/18257829.Kiley_Reid",
"image_url": "https://images.gr-assets.com/authors/1584271436p5/18257829.jpg",
"rating_count": "247247",
"review_count": "22391",
"rating": "3.86",
"related_url": "https://www.goodreads.com//author/similar/18257829.Kiley_Reid",
"author_books": [
"https://www.goodreads.com/book/show/43923951-such-a-fun-age"
],
"related_authors": [
"https://www.goodreads.com/author/show/1973.Tracy_Chevalier",
"https://www.goodreads.com/author/show/6781.Julie_Andrews_Edwards",
"https://www.goodreads.com/author/show/7380.Alice_Walker",
"https://www.goodreads.com/author/show/46097.Richard_Paul_Evans",
"https://www.goodreads.com/author/show/48895.Linda_Fairstein",
"https://www.goodreads.com/author/show/74224.Cari_Meister",
"https://www.goodreads.com/author/show/86734.Barry_Maitland",
"https://www.goodreads.com/author/show/97313.Elizabeth_Strout",
"https://www.goodreads.com/author/show/156327.Deanna_Raybourn",
"https://www.goodreads.com/author/show/281810.Jojo_Moyes",
"https://www.goodreads.com/author/show/342188.Charlotte_Wood",
"https://www.goodreads.com/author/show/702360.Erica_Bauermeister",
"https://www.goodreads.com/author/show/1229281.Jason_Reynolds",
"https://www.goodreads.com/author/show/4405738.Valeria_Luiselli",
"https://www.goodreads.com/author/show/5247419.Sonia_Purnell",
"https://www.goodreads.com/author/show/6595905.Cale_Atkinson",
"https://www.goodreads.com/author/show/6646248.Brooke_Davis",
"https://www.goodreads.com/author/show/7109533.Lila_Monroe",
book_mongodb 〉 1 〉 similar_books 〉 2
4: Run
6: Problems
TODO
Terminal
Python Console
3 Event Log
Tests passed: 10 (today 20:48)
65:46 LF UTF-8 2 spaces No JSON schema Python 3.6 (base)

- **-ex**
  - Instruction:
    - -ex alone would activate the export function of this Web Scraper Application.
    - A js file like this would be generated (picture below):

  - The overall structure would be a dictionary with two keys where each key represents a database collection, and each value is a list of all dictionaries. Inside each list, all the book dictionary and author dictionary are displayed.

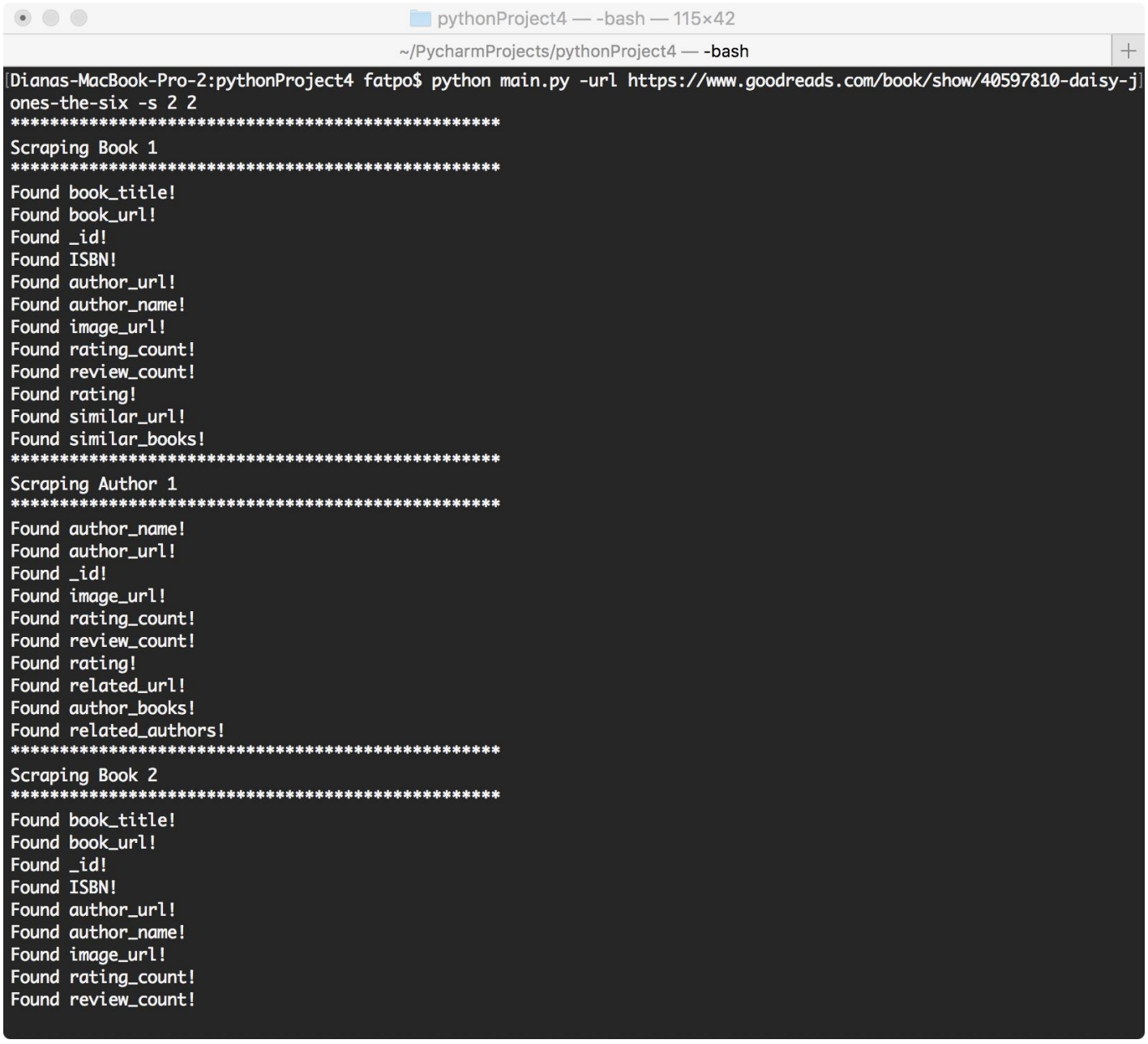
- The json file passed into this program should also follow such a format in order to correctly update our database



Progress Log

- **Search Log**
  - If the information was found during scraping, the message “Found [info] !” would be printed in terminal

  - If the information was not found during scraping, the message “Could not find [info] !” would be printed instead.



- Note that even if some information fragments might be imperfect, we still store this book/author data into our database to save scraping time.

- Scrape Log

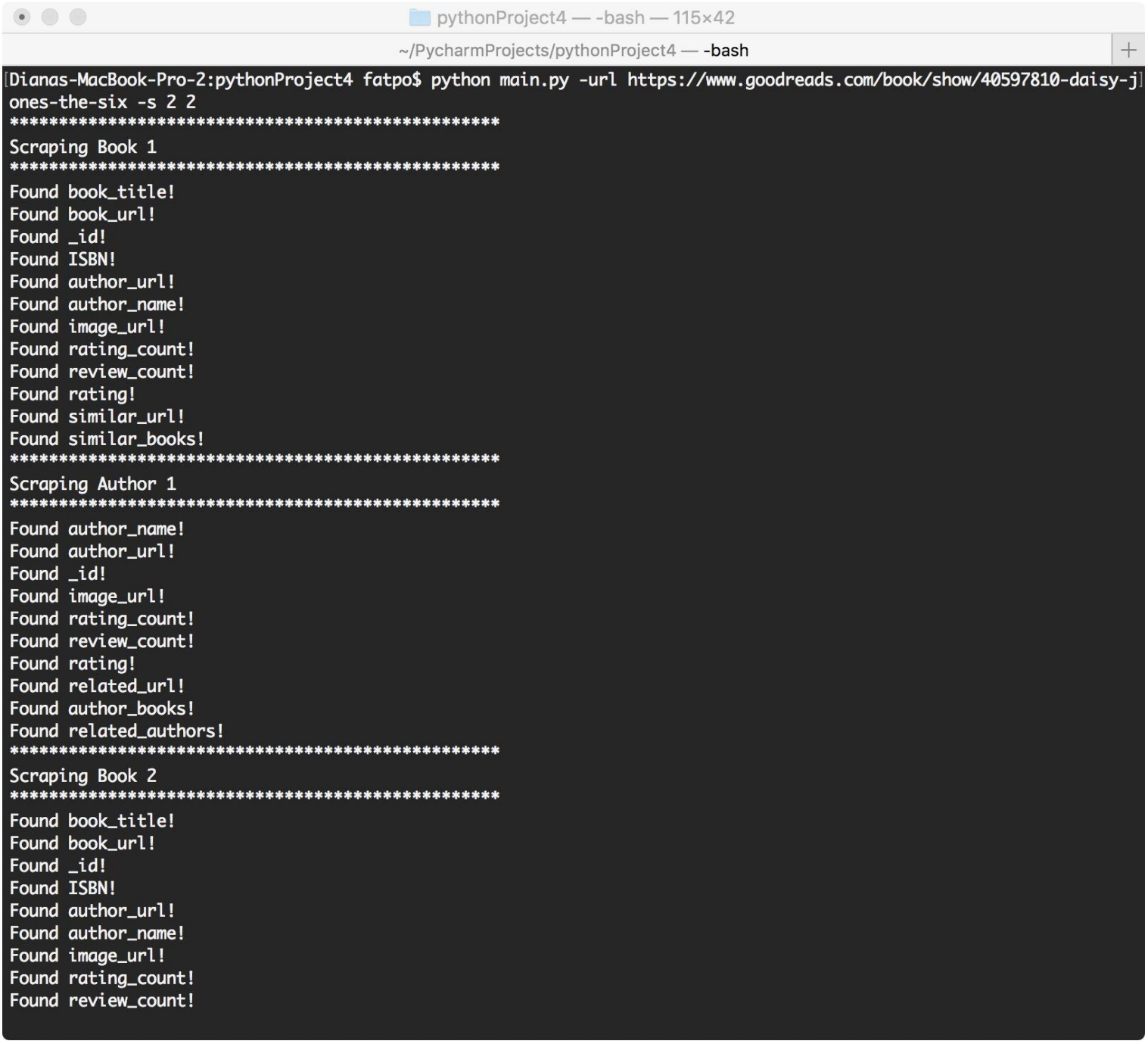
  - Each time while scraping a book or an author, before logging searching messages, a scrape log would be printed.

  - For a book, a message “Scraping Book [book_num]” would be printed.

  - For an author, a message “Scraping Author [author_num]” would be printed.

  - Note that each scrape log is surrounded by * marks to make the terminal view look more clear and easy to read.

- Since based on my scraping order: book-author-similar book-author, the order of the display of my scrape log is book log first and then author log like the picture down below.



Future Tests

- **Interface**
  - **-js**
    - For this part, since I cannot generate enough mal-structured json file for test, this part still need to be tested.
    - Besides, the create and update part of this Web Scraper also need to be tested, which means a valid json file with new data or changed data need to be generated as input for tests.
    - Furthermore, the update method need to be further discussed. For example, can we just delete the dictionary and insert the new updated one instead of update attributions.

HTML Design

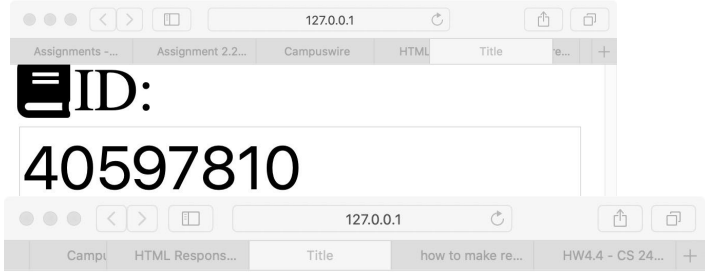
Tests for GET:

1. Check if the GET in HTML could report errors when errors occur.
2. Check if the GET in HTML could return correct output in a table content like below.
3. Check if the GET in HTML could return correct output at the very bottom of the webpage to avoid element moving like below.

ISBN	9781524798628
_id	40597810
author_name	Taylor Jenkins Reid
author_url	https://www.goodreads.com/author/show/6572605.Taylor_Jenkins_Reid
book_title	Daisy Jones & The Six
book_url	https://www.goodreads.com/book/show/40597810-daisy-jones-the-six
id	40597810
image_url	https://i.gr-assets.com/images/S/compressed.photo.goodreads.com/books/1580255154i/40597810._SR1200,6
rating	4.20
rating_count	324936
review_count	45113
similar_books	https://www.goodreads.com/book/show/43923951-such-a-fun-age https://www.goodreads.com/book/show/51791252-the-vanishing-half https://www.goodreads.com/book/show/51933429-the-guest-list https://www.goodreads.com/book/show/51918871-city-of-girls https://www.goodreads.com/book/show/42201996-ask-again-yes https://www.goodreads.com/book/show/52867387-beach-read https://www.goodreads.com/book/show/41057294-normal-people https://www.goodreads.com/book/show/49127718-anxious-people https://www.goodreads.com/book/show/40097951-the-silent-patient https://www.goodreads.com/book/show/50623864-the-invisible-life-of-addie-larue https://www.goodreads.com/book/show/50093704-in-five-years https://www.goodreads.com/book/show/44318414-the-dutch-house https://www.goodreads.com/book/show/52578297-the-midnight-library https://www.goodreads.com/book/show/45046527-american-dirt https://www.goodreads.com/book/show/52129515-untamed https://www.goodreads.com/book/show/44890081-my-dark-vanessa https://www.goodreads.com/book/show/43925876-the-giver-of-stars https://www.goodreads.com/book/show/45294613-dear-edward https://www.goodreads.com/book/show/41150487-red-white-royal-blue https://www.goodreads.com/book/show/42368149-red-white-royal-blue https://www.goodreads.com/book/show/41150483-the-first-mistake https://www.goodreads.com/book/show/38355282-watching-you https://www.goodreads.com/book/show/42201431-the-unhoneymooners https://www.goodreads.com/book/show/52709482-the-unhoneymooners https://www.goodreads.com/book/show/38357345-my-favorite-half-night-stand https://www.goodreads.com/book/show/43885930-life-will-be-the-death-of-me https://www.goodreads.com/book/show/40864790-pumpkinheads https://www.goodreads.com/book/show/38882569-park-avenue-summer https://www.goodreads.com/book/show/40189670-josh-and-hazel-s-guide-to-not-dating https://www.goodreads.com/book/show/40702156-i-owe-you-one https://www.goodreads.com/book/show/42519313-nothing-to-see-here https://www.goodreads.com/book/show/40539018-i-miss-you-when-i-blink
similar_url	https://www.goodreads.com/book/similar/61127102-daisy-jones-the-six

4. Check if the PUT in HTML could report errors when errors occur.
5. Check if the PUT in HTML could notify success right below the request like below.

6. Check if the PUT in HTML could notify failure right below the request like below.
7. Check if the POST in HTML could report errors when errors occur.
8. Check if the POST in HTML could notify success right below the request(similar to PUT).
9. Check if the POST in HTML could notify failure right below



the request(similar to PUT).

10. Check if the DELETE in HTML could report errors when errors occur.
11. Check if the DELETE in HTML could notify success right below the request(similar to PUT).
12. Check if the DELETE in HTML could notify failure right below the request(similar to PUT).
13. Check does all css icon work in HTML like below.

14. Check whether icons also follow responsive design.

15. Check all input boxes follow responsive design.

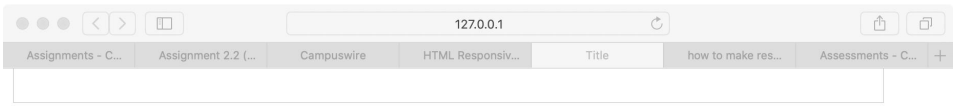
16. Check the responsive design support the maximum requirement.

17. Check the responsive design support the minimum requirement.

18. Check whether rank have a separate route.

19. Check if the GET rank in HTML could report errors when errors occur.

20. Check if the GET rank in HTML could display with customizable k like below.


127.0.0.1
Assignments - C...
Assignment 2.2 (...
Campuswire
HTML Responsiv...
Title
how to make res...
Assessments - C...

DELETE

Daisy Jones & The Six 4.20
A Deadly Education (The Scholomance, #1) 4.05