

# Final Project Documentation & Demo

----- Xinyi Zhao

## Project code

Link to github: [https://github.com/sarahzhao21/final\\_project.git](https://github.com/sarahzhao21/final_project.git)

There are 8 files for python packages in this repository:

- 1) README.md, containing instructions for running my APP
- 2) final\_proj.py, the program for the web scraping and crawling and creating the .sqlite file of my database.
- 3) cache.json, the json file created when the final\_proj.py has been running and the cache was being used by web scraping.
- 4) best\_seller\_books.sqlite, the database created by final\_proj.py which will be used for the flask and APP design.
- 5) app.py, the program to obtain the information from users and present the information by the requests of users.
- 6) index.html (in the folder 'templates'), the program that combined with app.py and show the interaction form to users
- 7) results.html (in the folder 'templates'), the program that combined with app.py and present the information by table based on the requests of users
- 8) plot.html (in the folder 'templates'), the program that combined with app.py and present the information as bar chart based on the requests of users.

## Data sources

1. The New York Times Best Sellers: <https://www.nytimes.com/books/best-sellers/>

This website provides the lists of the best seller books in different categories in "New York Times".

Format: HTML, JSON

How: I did web crawling and scraping to get the information of best seller books and get the url of another link in "Apple Books" for each book and do the web crawling to "Apple Books". Caching was used during web scraping and crawling.

Summary of data:

- 1) There are 140 records available (11 book categories and 10 to 15 best seller books for each categories)
- 2) 140 records have been retrieved.
- 3) description of records:

fields	Data type	Description
Id	INTEGER	The Id number of the book
Title	TEXT	The title of the book
Category	TEXT	What kind of book it is (fiction, novel ....)
Author	TEXT	The author who wrote the book
Publisher	TEXT	The publisher of the book
Rank	INTEGER	The rank of this book in it's category last week
Weeks_on_the_list	INTEGER	How many weeks the book was on the best-seller list
Description	TEXT	A short review about this book
Apple_url	TEXT	The url that connects to the website of 'Apple Books '

## 2. Apple Books

Each book from the New York Time Best Sellers has an 'Apple Books' url, here is one of the example: <https://books.apple.com/us/book/american-dirt-oprahs-book-club/id1459876569>

This website provides all kinds of detailed information about the books showed on New York Times.

Format: HTML, JSON

How: By accessing the beautiful soup of the Apple Books url for each book, I did the web scraping to obtain the information of 'title', 'genre', 'price', 'rating', 'released date', 'language', 'length', 'seller' and 'size'. Caching was used during web scraping.

Summary of data:

- 1) There are 140 records available (each record refers to one book on New York Time best seller)
- 2) 140 records have been retrieved.
- 3) description of records:

fields	Data type	Description
Id	INTEGER	The Id number of the book
Title	TEXT	The title of the book
Rating	REAL	The rating of the book on "Apple Books"
Price	REAL	The price of the book on "Apple Books"
Genre	TEXT	The genre of the book on "Apple Books"
Released_date	TEXT	The date that the book released on the website
Language	INTEGER	What kind of language the book was written by
Length	INTEGER	The page number of the book
Seller	TEXT	The seller of the book
Size	REAL	How many 'MB' of the book

## Database

A file named 'best\_seller\_book.sqlite' has been created by the program 'final\_proj.py', when it was opened by DB Browser, there are two tables in this file: 'Best\_seller' and 'Apple\_book':

1. 'Best\_seller' has 9 columns and 140 rows. Each row represents the information of a book, the information of the columns is listed here:

Column name	Data type
Id	INTEGER
Title	Text
Category	TEXT
Author	TEXT
Publisher	TEXT
Rank	INTEGER
Weeks_on_the_list	INTEGER
Description	TEXT
Apple_url	TEXT

2. 'Apple\_book' has 10 columns and 140 rows. Each row represents the information of a book on "Apple Books" and refer to the same book on the table 'Best\_seller' with the same order. The information of the columns is listed here:

Column name	Data type
Id	INTEGER
Title	TEXT

Rating	REAL
Price	REAL
Genre	TEXT
Released_date	TEXT
Language	INTEGER
Length	INTEGER
Seller	TEXT
Size	REAL

- The primary key is the 'Id' on 'Best\_seller' table and the foreign key is the 'Id' on 'Apple\_book' table. The sequences of these two sets of 'Id's are the same.
- Screen shot:

Id	Title	Category	Author	Publisher	Rank	Weeks_on_the_list	Description	Apple_url
1	LITTLE FIRES EV...	combined-print...	Celeste Ng	Penguin Press	1	60	An artist upend...	https://du-gae...
2	WHERE THE CRA...	combined-print...	Della Owens	Putnam	2	82	In a quiet town ...	https://du-gae...
3	VALENTINE	combined-print...	Elizabeth Wetm...	Harper	3	0	A Texas town o...	https://du-gae...
4	TEXAS OUTLAW	combined-print...	James Patterson...	Little, Brown	4	0	A Texas Ranger	https://du-gae...
5	THE BOY FROM ...	combined-print...	Harlan Coben	Grand Central	5	3	When a girl gae...	https://du-gae...
6	AMERICAN DIRT	combined-print...	Jeanine Cummins	Flatiron	6	11	A bookseller fle...	https://du-gae...
7	THE SILENT PATI...	combined-print...	Alex Michaelides	Celadon	7	29	Theo Faber look...	https://du-gae...
8	THE GIVER OF S...	combined-print...	Jojo Moyes	Pamela Dorman...	8	21	In Depression-e...	https://du-gae...
9	IN FIVE YEARS	combined-print...	Rebecca Serie	Atria	9	4	A Manhattan la...	https://du-gae...
10	THE DUTCH HO...	combined-print...	Ann Patchett	Harper	10	26	A sibling relatio...	https://du-gae...
11	THE GLASS HOTEL	combined-print...	Emily St. John M...	Knopf	11	2	Years after an in...	https://du-gae...
12	NORMAL PEOPLE	combined-print...	Sally Rooney	Hogarth	12	3	The connection ...	https://du-gae...
13	THE MIRROR & ...	combined-print...	Hilary Mantel	Holt	13	4	The third book i...	https://du-gae...
14	FATE	combined-print...	Helen Hardt	Waterhouse	14	0	The 13th book i...	https://du-gae...
15	THE NIGHT WAT...	combined-print...	Louise Erdrich	Harper	15	0	As a bill that ma...	https://du-gae...
16	THE SPLENDID A...	combined-print...	Erik Larson	Crown	1	6	An examination ...	https://du-gae...
17	UNTAMED	combined-print...	Glenon Doyle	The Dial Press	2	4	The activist and ...	https://du-gae...
18	FRONT ROW AT ...	combined-print...	Jonathan Karl	Dutton	3	0	The ABC News c...	https://du-gae...
19	MORE MYSELF	combined-print...	Alicia Keys with ...	Flatiron	4	0	The Grammy A...	https://du-gae...
20	THE FIRST TIME	combined-print...	Colton Underwo...	Gallery	5	0	A memoir by a f...	https://du-gae...
21	EDUCATED	combined-print...	Tara Westover	Random House	6	111	The daughter of...	https://du-gae...
22	WOW, NO THANK ...	combined-print...	Samantha Irby	Vintage	7	0	Comedic essays ...	https://du-gae...
23	OPEN BOOK	combined-print...	Jessica Simpson	Day St.	8	9	The singer, actr...	https://du-gae...
24	THE GREAT INFLU...	combined-print...	John M. Barry	Penguin	9	4	An overview of t...	https://du-gae...
25	BECOMING	combined-print...	Michelle Obama	Crown	10	70	The former first ...	https://du-gae...
26	THE MAMBA ME...	combined-print...	Kobe Bryant	MeLcher Media/...	11	13	Various skills an...	https://du-gae...
27	UNORTHODOX	combined-print...	Deborah Feldman	Simon & Schuster	12	5	A woman break...	https://du-gae...

Id	Title	Rating	Price	Genre	Released_date	Language	Length	Seller	Size
1	Little Fires Every...	4.4	9.99	Fiction & Literat...	2017_09_12	English	352	PENGUIN GROU...	1.7
2	Where the Craw...	4.6	14.99	Fiction & Literat...	2018_08_14	English	384	PENGUIN GROU...	4.5
3	Valentine	3.9	13.99	Fiction & Literat...	2020_03_31	English	320	HARPERCOLLINS...	1.6
4	Texas Outlaw	4.5	14.99	Mysteries & Thr...	2020_03_30	English	448	Hachette Digital...	1.1
5	The Boy from th...	4.1	14.99	Mysteries & Thr...	2020_03_17	English	384	Hachette Digital...	1.0
6	American Dirt (...)	3.8	14.99	Fiction & Literat...	2020_01_21	English	400	Macmillan	5.5
7	The Silent Patient	4.3	13.99	Mysteries & Thr...	2019_02_05	English	304	Macmillan	5.5
8	The Giver of Stars	4.5	13.99	Fiction & Literat...	2019_10_08	English	400	PENGUIN GROU...	3.0
9	In Five Years	4.2	12.99	Fiction & Literat...	2020_03_10	English	272	SIMON AND SC...	3.8
10	The Dutch House	4.3	14.99	Fiction & Literat...	2019_09_24	English	352	HARPERCOLLINS...	2.0
11	The Glass Hotel	4.0	13.99	Fiction & Literat...	2020_03_24	English	320	Penguin Rando...	1.9
12	Normal People	3.9	11.99	Fiction & Literat...	2019_04_16	English	288	Penguin Rando...	2.5
13	The Mirror & th...	4.3	14.99	Fiction & Literat...	2020_03_31	English	480	Macmillan	7.2
14	Fate	4.3	6.99	Romance	2020_03_31	English	330	Meredith Wild LLC	1.8
15	The Night Watc...	4.3	14.99	Fiction & Literat...	2020_01_03	English	464	HARPERCOLLINS...	3.9
16	The Splendid an...	4.5	14.99	History	2020_02_25	English	608	Penguin Rando...	5.3
17	Untamed	4.5	14.99	Biographies & M...	2020_01_10	English	352	Penguin Rando...	4.7
18	Front Row at th...	4.4	3.99	Reference	2020_10_12	English	352	Macmillan	1.1
19	More Myself	4.4	14.99	Biographies & M...	2020_03_31	English	320	Macmillan	2.5
20	The First Time	4.6	12.99	Biographies & M...	2020_03_31	English	288	SIMON AND SC...	21.5
21	Educated	4.6	14.99	Biographies & M...	2018_02_20	English	352	Penguin Rando...	2.7
22	Wow, No Thank ...	4.3	9.99	Humor	2020_03_31	English	336	Penguin Rando...	2.1
23	Open Book	4.5	14.99	Biographies & M...	2020_02_04	English	304	HARPERCOLLINS...	2.4
24	The Great Influe...	4.2	13.99	History	2020_02_09	English	560	PENGUIN GROU...	3.8
25	Becoming	4.6	14.99	Biographies & M...	2018_11_13	English	448	Penguin Rando...	114.4
26	The Mamba Me...	4.7	16.99	Sports & Outdoors	2018_10_23	English	208	Macmillan	219.2
27	Unorthodox	4.3	12.99	Biographies & M...	2012_02_14	English	272	SIMON AND SC...	3.5
28	Sapiens	4.5	4.99	Science & Nature	2015_02_10	English	464	HARPERCOLLINS...	29.9
29	Maybe You Sho...	4.6	12.99	Biographies & M...	2019_04_02	English	464	Houghton Miffl...	6.6
30	Summary of 'Tal...	3.8	11.99	Reference	2019_12_14	English	271	Drach77/Thrall...	100.6

## Interaction and Presentation

Interactive and presentation technologies used: Flask and Plotly

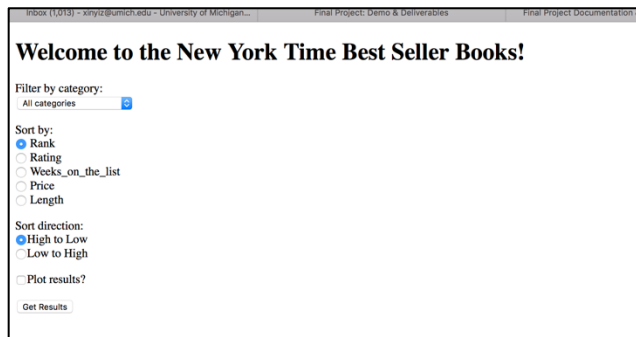
User-facing capabilities:

1. Users can select what kind of books they are interested in (there are 11 different categories can be selected).
2. Users can choose how to sort the book list by five kind of ways (by Rank, Rating, Weeks\_on\_the\_list, Price or Length).
3. Users can also select how to list the books by the values they selects (from High to Low or from Low to High).
4. Users can also choose what kind of results they want to see (a table with 10 books with detailed information or a bar chart of the 10 books with the sorted values that were selected)

Brief instructions for how a user would interact with my program:

1. Run the 'app.py' program and open the link '<http://127.0.0.1:5000/>' in the browser.
2. Click the box below the "Filter by category", user can select what kind of books he is interested in (there are 11 different categories).
3. Select how to sort the book list by five kind of ways (by Rank, Rating, Weeks\_on\_the\_list, Price or Length).
4. Select how to list the books (from High to Low or from Low to High).
5. Select how to present the results (a table with 10 books with detailed information or a bar chart of the 10 books with the value of the standard you selected). If the user wants to see the bar chart, select 'Plot results?'
6. Then click 'Get Results', then the results will be shown in a new website in the browser.

Here is what the submission form looks like:



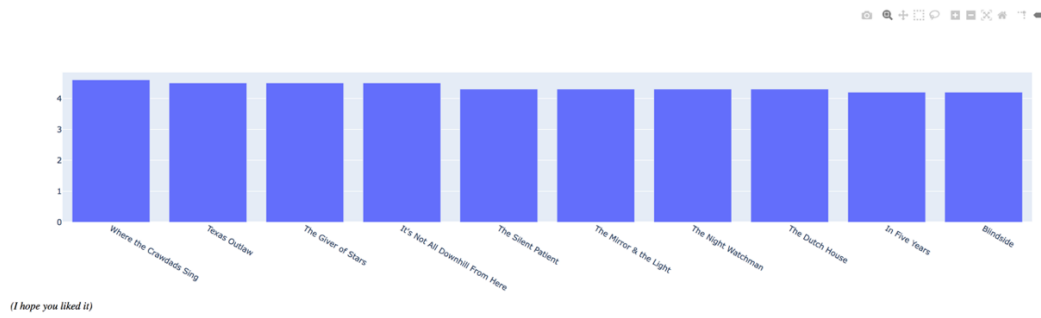
Here is what the result looks like if the user doesn't select "Plot results":

### Here are your results!

Title	Genre	Author	Rating	Released_date	Seller
Where the Crawdads Sing	Fiction & Literature	Delia Owens	4.6	2018_08_14	PENGUIN GROUP USA, INC.
Texas Outlaw	Mysteries & Thrillers	James Patterson and Andrew Bourelle	4.5	2020_03_30	Hachette Digital, Inc.
The Giver of Stars	Fiction & Literature	Jojo Moyes	4.5	2019_10_08	PENGUIN GROUP USA, INC.
It's Not All Downhill From Here	Fiction & Literature	Terry McMillan	4.5	2020_03_31	Penguin Random House LLC
The Silent Patient	Mysteries & Thrillers	Alex Michaelides	4.3	2019_02_05	Macmillan
The Mirror & the Light	Fiction & Literature	Hilary Mantel	4.3	2020_03_10	Macmillan
The Night Watchman	Fiction & Literature	Louise Erdrich	4.3	2020_03_03	HARPERCOLLINS PUBLISHERS
The Dutch House	Fiction & Literature	Ann Patchett	4.3	2019_09_24	HARPERCOLLINS PUBLISHERS
In Five Years	Fiction & Literature	Rebecca Serle	4.2	2020_03_10	SIMON AND SCHUSTER DIGITAL SALES INC
Blindside	Mysteries & Thrillers	James Patterson and James O. Born	4.2	2020_02_24	Hachette Digital, Inc.

Here is what the result looks like if the user selects “Plot results”:

**Here is your graph!**



## Demo Link

<https://www.loom.com/share/b298e95e034e4eb49be94e18f2a7327a>