Team Name: Eric
Team Member: Eric Truong
Project Title: Book Bastion

Report on Project 3

Github URL: https://github.com/EricUF/Project3
Video URL: https://www.youtube.com/watch?v=0Aa8nFEpmh0&ab_channel=EricTruong

Proposal

Problem

The problem my project was seeking to solve is which data structure would be the most effective as storing and then searching through for storing data about books. The books I used came from the dataset, https://www.kaggle.com/code/mohamedbakhet/eda-for-amazon-books-reviews/data , which contains over 200,000 books which came from the Amazon book store with data gathered from May 1996 to July 204.

Motivation

Searching through rows of shelves can be a daunting task when you're searching for the book that you're looking forward to reading, especially with the amount of books available today. I enjoy reading books and oftentimes I try to look around for books with only a few characteristics in mind. If one were to search from the list of all these books it can be difficult to find exactly what they are looking for. It can be an even harder problem when the user doesn't know exactly what they are looking for, but might have a general idea of what they are trying to find. With a good data structure the user could search for a specific title or author or even search for a list of books based on the parameters they wish to use.

Features

For my project the futures implemented are the ability to search through two different data structures using either the titles, author names, or with the parameters of the genre, the range of year published, and the amount of ratings. The user also has the option to use two different data structures, an ordered_map and a unordered_map. The results will print out a list of options that fit the search options of the user and from there the user can select the book they wish to learn more information on. The program will also display the time spent in a function using the chrono library.

Data

The dataset I chose came from Kaggle which has many publicly available datasets. The one I chose was about the information about books from the Amazon book store. The data used a csv to hold the information which contained information on the titles, descriptions, authors, image link, preview link, publisher name, publish data, information link, the genre, and the amount of ratings. The original file contained over 200,000 unique books, however for the csv file many of the data points in some of the rows were blank so I decided to not use any books with a complete row of data which then brought the dataset to contain only about 118,000 unique books to analyze. For my project I am analyzing two different data structures, an

ordered_map and an unordered_map. For both map structures I made three of each type. For both structures, I made a map with the key being the order the books came in the dataset, which has no particular order. Then I made a map with the key being the title of the books and then also a map with the key being the author names.

Tools

For my programming language I used C++ as it is the language I am most familiar with and have used very often especially for my classes. I used Visual Studios 2022 Community Edition as it is an IDE I am the most familiar with for C++. I used only the libraries included in the C++ Standard Library. For part of my project that I was unable to fully complete I used CLR Net Framework to build a GUI for my program in which the GUI was fully completed however I was unable to connect the code I had made to the GUI I had created. I also used Excel to look at and analyze the data I was using.

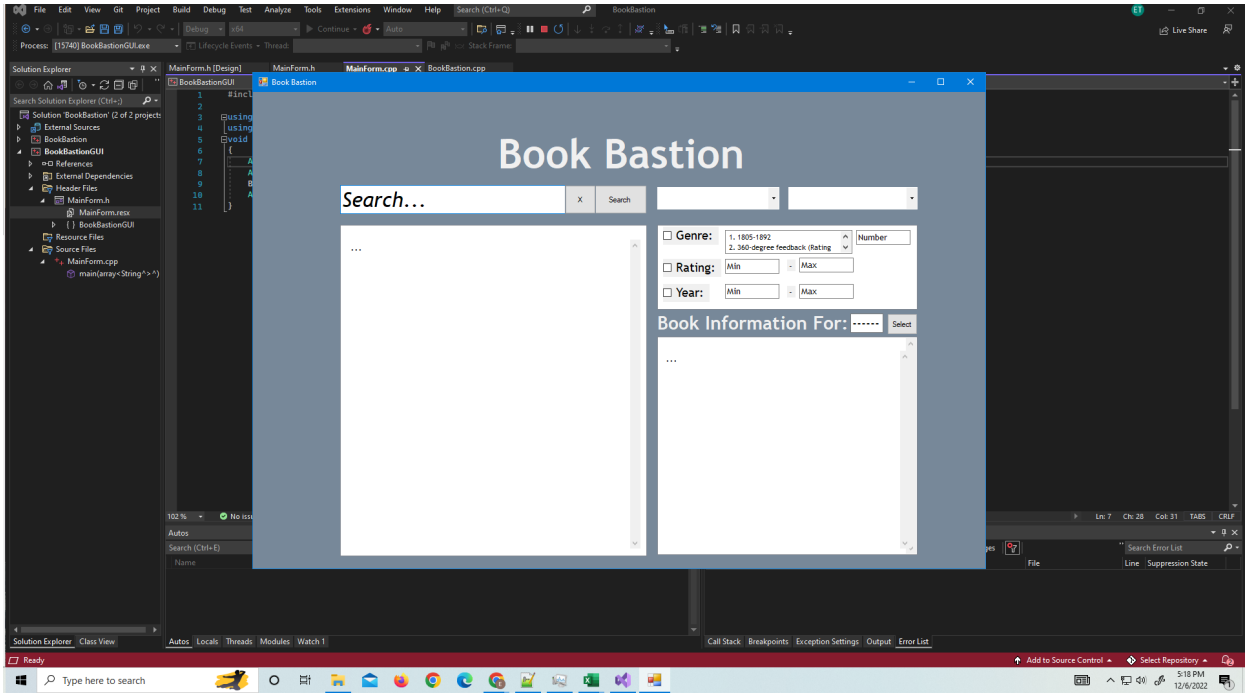
Responsibilities

This section will be brief as this was a solo project in which I completed by myself thus all of the responsibilities and roles fall on to me.

Analysis

Some changes I made to the proposal during my work was that I had to not include the rating csv file that was also available to me. The rating csv file which includes the actual rating of the books, the prices of the books, and the comments or reviews left behind users was a csv file with over 3,000,000 lines of data. I was unable to even open the file in Excel. I used multiple programs to split or splice the data into smaller sheets however the data would only contain a small portion of the rating csv file, leaving out the rest of the data. After a few days of no success with this I chose to not include it as I still had over 100,000 points of data in the book data csv file. Another change is the user interface. I coded the program entirely on Visual Studios using the command line interface just to get an idea of how it would work. From there I attempted to learn how to use the CLR Net Framework to build the GUI. Creating the visual interface itself was easier than getting my code to work with the GUI. The search bar, drop down options, and input options all work however I was unable to figure out how to import the code I had already created before into the code for the GUI.

An image of the GUI made.



For my project I had a total of 9 functions including the main function. I will go over each function in ascending order starting from the main function.

For the main function the code here initializes the data structures necessary for the program and then runs a while loop that will allow the user to use the actual program. For some of the menu options in the main function a list of the genres available are also printed out using a for loop. Thus the worst case time complexity of the main function is O(n^2). The first n comes from the while loop which is determined by how many times the user chooses to keep using the program and can be as short as 1 run. The second n comes the list of genres that is printed out from the dataset with n being the amount of genres. For this dataset there are 3,069 genres but with a different dataset it can vary.

For the next function which is the search function for the unordered map of author names the function uses a while loop to iterate through the map and then a for loop outside of the while loop for the list of books that were a match with the parameter. Thus the complexity time is O(n) with the n being the amount of books in the map and also n is the amount of books that were a match for the for loop. The n for the while loop however will always be bigger as it is all the points in the dataset.

After this I have 3 more search functions each built specifically for each method but follow a similar structure to the function above, the unordered map of author names. Because of the similar structure with the while loop and the for loop each of these time complexities are also O(n) with n being the same variable.

Then I have two search functions that use the genre, year, and ratings as its parameters. Here it also uses the while loop and then the for loop, but for these functions the while loop also contains a for loop in order to search for the year data point. The nested for loop goes through the length of the publish date as some dates included a month and day and some only included

the year and this for loop helped break down the string into just the year. This leads to this function having a worst case time complexity of O(n^2) with n being the amount of points in the data set and the second n being the length of the date string which in some cases were 12 characters long but could also be as short as 4 characters.

The next function is a function I used from one of my references to detect if a string is only digits and does not contain any characters. This function has an O(1) time as it only contains one line that returns a true or false if a string is just digits.

The last function is called bookDataImport which does as the names suggest, imports the dataset into each data structure I have. This function contains only one loop which is a while loop that will iterate through every line of data in the dataset. This function has a worst case time complexity of O(n) with n being the lines of data in the dataset.

Reflection

As a student working solo it was very time consuming to work on the project. I was originally in a group but I had left due to some personal issues causing my schedule to be very disordered. Overall, the experience of building this code was fun and frustrating. Getting the dataset and to import into the data structures was extremely difficult due to how the datasets were. There were many blocks I faced while making my code. The csv file that I used contained no distinct characters to use as a delimiter when separating a line of data because some of the texts contained commas at the start or end of the string which made it impossible to separate a line using any character. From there I had to edit the dataset to include recognizable characters in order to separate them properly.

A large issue I faced as noted earlier was the inability to access the second csv file that contained the data I initially wanted to use as search parameters as well as information that would have been extremely beneficial to see when looking at a book. I was unable to overcome this obstacle. The last and most difficult obstacle that I also did not overcome was that I could not connect my code with the GUI I had created. It was my first time ever really creating a GUI for C++ and I did not know where to start. After doing some research I opted to try using Visual C++ with a CLR Net Framework and from there I was able to create the visual aspect of the GUI along with the buttons, drop down menus, and scrollable list.

If I were to start this project again I would instead focus on building the GUI and the code for it first and from there I would actually build the code in accordance to that. Seeing the code now for the GUI file I feel as if I could code my project into but due to time constraints I did not have time to recode my whole entire project. Another idea is that perhaps I would also try another language that may be easier for me to use with an interface that is user friendly. As for workflow I had felt like I was working through the steps on time however I had gotten caught up on a few obstacles that slowed me down. As I said before I create and code from the CRL Net Framework first but as for the other csv file I still am unsure as to how I could access its data.

From this project I learned how to analyze large datasets and properly import them into data structures and I also learned how to search through these programs and display the correct results.

References

https://www.geeksforgeeks.org/unordered_map-insert-in-c-stl/
https://www.geeksforgeeks.org/map-insert-in-c-stl/
https://www.geeksforgeeks.org/map-vs-unordered_map-c/
https://www.geeksforgeeks.org/unordered_map-in-cpp-stl/
https://absentdata.com/delete-blank-rows/
https://learn.microsoft.com/en-us/office/vba/library-reference/concepts/getting-started-with-vba-in-office
https://www.tutorialkart.com/cpp/cpp-string-find/#:~:text=string%3A%3Afind()%20function,%2C%20find()%20returns%20%2D1.
https://www.geeksforgeeks.org/search-by-value-in-a-map-in-c/
https://stackoverflow.com/questions/2340281/check-if-a-string-contains-a-string-in-c
https://www.youtube.com/watch?v=LF1cI7zeFm4&ab_channel=Simplilearn
https://www.kaggle.com/code/mohamedbakhet/eda-for-amazon-books-reviews/data?select=books_data.csv
https://stackoverflow.com/questions/8888748/how-to-check-if-given-c-string-or-char-contains-only-digits
https://www.codeproject.com/Questions/141323/How-to-check-if-a-button-is-being-pressed-in-C
https://social.msdn.microsoft.com/Forums/en-US/e9ada815-cd3a-4280-812c-bbe913f3750e/populate-listbox-from-db-using-only-value-field?forum=aspgettingstarted
https://www.geeksforgeeks.org/measure-execution-time-function-cpp/
https://learn.microsoft.com/en-us/cpp/dotnet/how-to-create-clr-empty-projects?view=msvc-170
https://stackoverflow.com/questions/43629363/how-to-check-if-a-string-contains-a-char
https://support.microsoft.com/en-us/office/filter-for-unique-values-or-remove-duplicate-values-ccf664b0-81d6-449b-bbe1-8daaec1e83c2#:~:text=To%20filter%20for%20unique%20values%2C%20click%20Data%20%3E%20Sort%20%26%20Filter,group%20on%20the%20Home%20tab.
https://learn.microsoft.com/en-us/windows/apps/design/controls/combo-box
https://www.codeproject.com/Questions/554554/Dropplusdownplusmenuplusinplusvisualplusc-2b-2b
https://learn.microsoft.com/en-us/windows/apps/design/controls/item-containers-templates
https://www.functionx.com/vccli/controls/textbox.htm
https://social.msdn.microsoft.com/Forums/en-US/fcf3eed7-e2a8-49d5-a098-21cbda47e12c/how-do-i-create-a-clear-button-loop-to-clear-all-textboxes?forum=vbgeneral