

Final Project: Web Scrapping Java

Project Description: This program objective is to extract data from online websites using URL and manipulate the data as we want such as sorting and searching data that we retrieved. In this program, it now works on 2 webpages:

CCBC General Courses URL:

http://catalog.ccbcmd.edu/preview_program.php?catoid=28&poid=13923

IMDB Top 250 movies URL:

<http://www.imdb.com/chart/top>

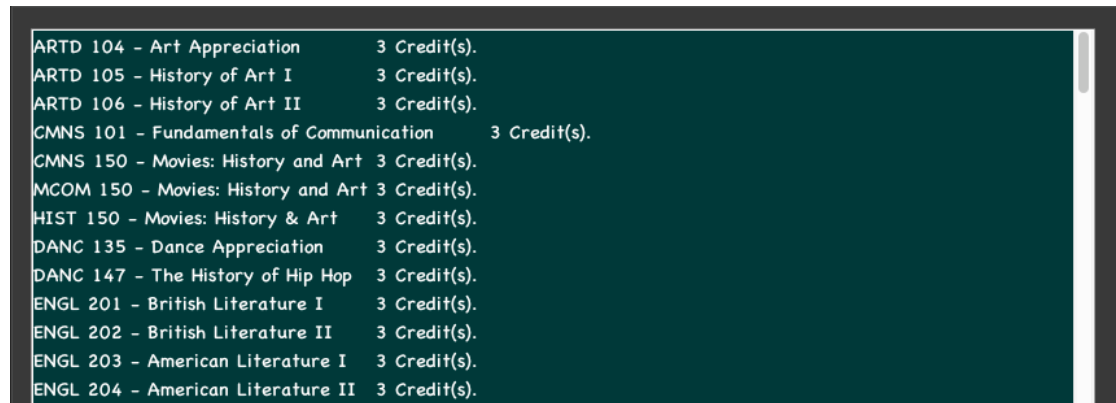
How each component works: Here is the look of the Web Scrapping Program.



1. Paste the URL on the URL's text field on the top left of the program, then click submit.
2. The program will retrieve and print all the target data that we want to have on the big green empty board. For example, all CCBC general courses will show

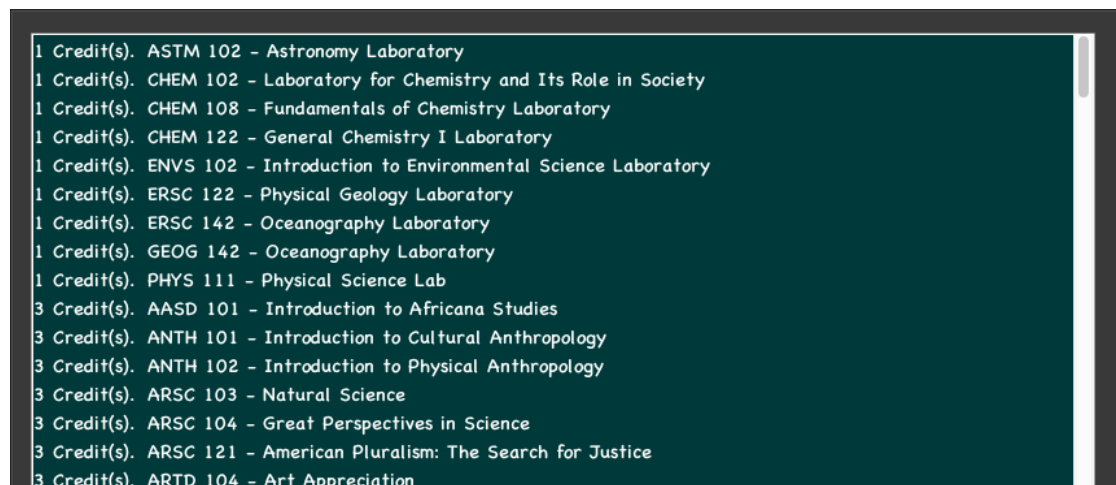
Final Project: Web Scrapping Java

up including short course titles (ENGL 101), full course names, and their credits when you submit CCBC's URL, or if you submit IMDB's URL you will get all the data from top 250 movies including ranks, movie names, and rating.



ARTD 104 - Art Appreciation	3 Credit(s).
ARTD 105 - History of Art I	3 Credit(s).
ARTD 106 - History of Art II	3 Credit(s).
CMNS 101 - Fundamentals of Communication	3 Credit(s).
CMNS 150 - Movies: History and Art	3 Credit(s).
MCOM 150 - Movies: History and Art	3 Credit(s).
HIST 150 - Movies: History & Art	3 Credit(s).
DANC 135 - Dance Appreciation	3 Credit(s).
DANC 147 - The History of Hip Hop	3 Credit(s).
ENGL 201 - British Literature I	3 Credit(s).
ENGL 202 - British Literature II	3 Credit(s).
ENGL 203 - American Literature I	3 Credit(s).
ENGL 204 - American Literature II	3 Credit(s).

3. Sorting – You can sort by click on the sorting buttons either from A-Z or Z-A, or sort Credits from Low-High or High-Low once you receive all the data from CCBC's website. You can also sort Rank from 1-250 or 250-1 or sort Rating from Low-High or High-Low from IMDB's URL as well.



1 Credit(s).	ASTM 102 - Astronomy Laboratory
1 Credit(s).	CHEM 102 - Laboratory for Chemistry and Its Role in Society
1 Credit(s).	CHEM 108 - Fundamentals of Chemistry Laboratory
1 Credit(s).	CHEM 122 - General Chemistry I Laboratory
1 Credit(s).	ENVS 102 - Introduction to Environmental Science Laboratory
1 Credit(s).	ERSC 122 - Physical Geology Laboratory
1 Credit(s).	ERSC 142 - Oceanography Laboratory
1 Credit(s).	GEOG 142 - Oceanography Laboratory
1 Credit(s).	PHYS 111 - Physical Science Lab
3 Credit(s).	AASD 101 - Introduction to Africana Studies
3 Credit(s).	ANTH 101 - Introduction to Cultural Anthropology
3 Credit(s).	ANTH 102 - Introduction to Physical Anthropology
3 Credit(s).	ARSC 103 - Natural Science
3 Credit(s).	ARSC 104 - Great Perspectives in Science
3 Credit(s).	ARSC 121 - American Pluralism: The Search for Justice
3 Credit(s).	ARTD 104 - Art Appreciation

4. Searching – A lot of data that are retrieved at once can be a bit overwhelming. In this program, you can search for just the information you want to see by typing the words or partial words in the searching area on the top right of the program, then click on search button, then the program will show just the

Final Project: Web Scrapping Java

results that include the words you search for.



5. Export – This part we still can't make it work at this moment.

Testing Plan: We tested every time new code has been added to make sure that the new code didn't ruin the original code. Most of the time, we tested everything, not just to see if the data still can be retrieved, but if the sorting and searching methods still work as well. Therefore, when we found bugs, we knew where to fix it.

Difficulties: As we were developing the project, we encountered a few difficulties.

1. When we started this project, all of us had only little idea about what web scrapping was. So, we took sometimes to learn and understand the idea and its purpose of making this project.
2. Rachada found some good Youtube videos that taught us how to do web scrapping using JSOUP in Java, she sent the links to the group, but Toby had

Final Project: Web Scrapping Java

no background in HTML, so it was very difficult for him to understand and manipulate it for our program. It took very long time for him to be able to read HTML tags from the websites.

3. As the program worked very well and we still had a few days left before the final day, we wanted to add the export function into our program to make it more useful for users, so we tried multiple times to make the export function works, but we still can't figure it out.

Big-O Notation:

Scrapping Data $\rightarrow O(n)$

Sorting $\rightarrow O(n^2)$ (Bubble sort with two nested loops)

Searching $\rightarrow O(n)$

Lessons we used in this project:

1. Abstract
2. Inheritance
3. Stack
4. Queue
5. Sorting
6. Searching
7. GUI

New things we learned from this project:

JSOUP, HTML, GUI (Eclipse's WindowBuilder)

Final Project: Web Scrapping Java

Citations:

JSOUP

- <https://jsoup.org>
- <https://www.youtube.com/watch?v=vqGPSGNe-dQ>

Sorting

- <https://mathbits.com/MathBits/Java/arrays/ABCSort.htm>

GUI (WindowBuilder)

- <https://www.youtube.com/watch?v=r8Qiz9Bn1Ag&t=1041s>
- <https://www.youtube.com/watch?v=-GoqPrxM8TQ&t=582s>