CSGY-9053: Introduction to JAVA
Final Project Report
Multi-Spider Web Crawler

Team Members:
- **Mahati Madhira VSK (mm12032)**
- **Sayali Jathar (snj4459)**

Github Repository: *https://github.com/MahatiMadhira/MultiThreadedWebCrawler*

Abstract:
A Web Crawler is a system for downloading, storing and analyzing web pages. They are used for various purposes like compiling web pages, indexing them and finding web pages that match their queries.
A Web Crawler bot's job is to go through all the web pages that it comes across starting from a seed URL and organize them in a systematic way to make it easier for anyone visiting these webpages to find content and related information easily.

Use Cases:
- Search Engine Indexing: Web Crawler generates a local index for search engines. Example: GoogleBot which is the web crawler for Google's search engine.
- Web Monitoring: Crawlers help monitor copyright and trademark violations on the Internet. Example: DigiMarc is a copyright institution that uses crawlers to uncover pirated resources.

Features of JAVA used:
- Jsoup JAVA library for parsing the HTML files and URLs.
- SQLite to store the scraped URLs in a database.
- Swing framework to create a GUI to display the results of scraping.
- Multi-Threading to implement multiple bots concurrently.
- Depth First Search method of traversing links.

How to Execute the code:
1. Download the code as a .zip or .tar file and unzip in the IDE workspace (I used Eclipse).
2. Add the Jsoup and SQLite .jar files to your project classpath so you can import the corresponding libraries.
3. Navigate to webCrawler/src/webCrawler/MainTest.java file and run the file to start the web crawling process.
4. If you don't have SQLite installed, please follow the installation guide based on your OS and click on the crawler.db file to access the database containing the scraped URLs in the crawler table which contains 3 columns: id, URL, PURL which describe the kind of data being scraped.

Issues Faced:

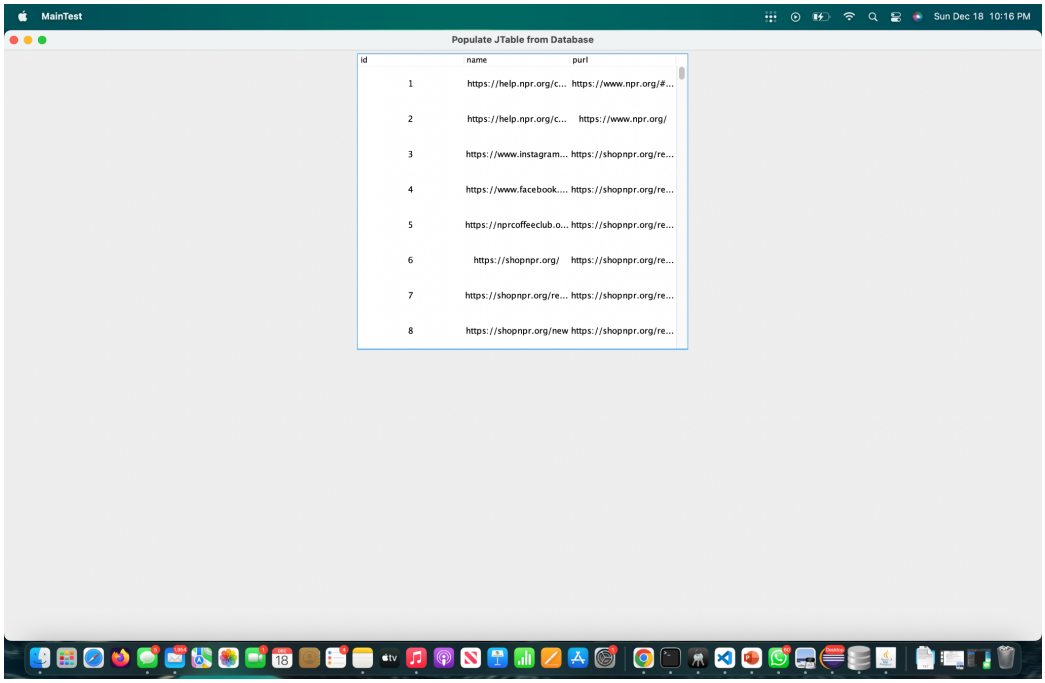
- We were unable to implement MD5 hashing checks to prevent looping in our web crawlers.

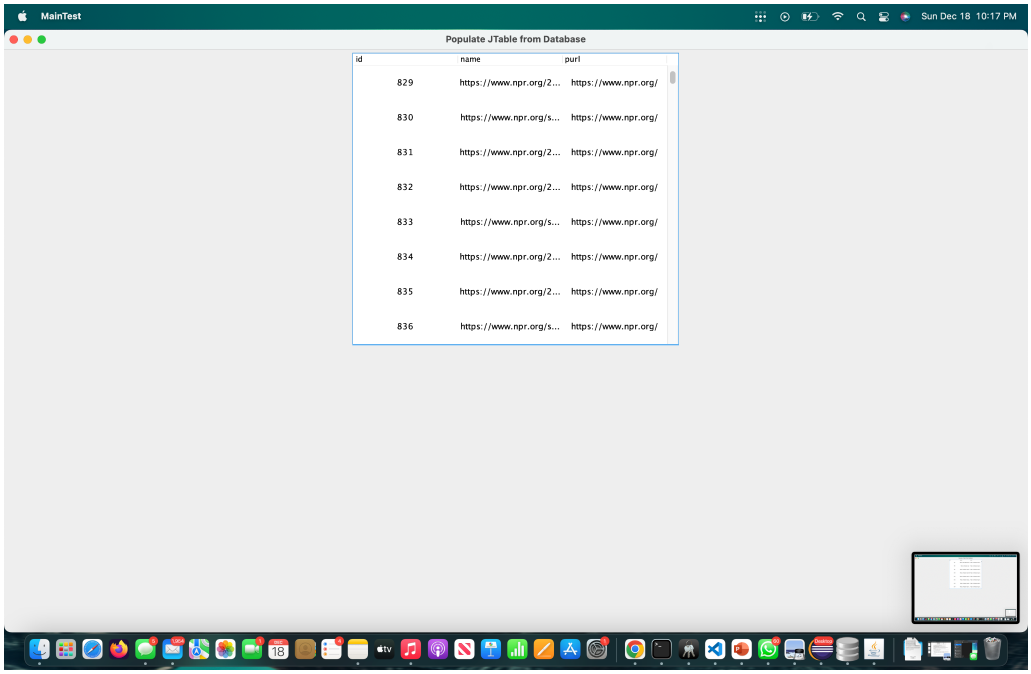
Results:

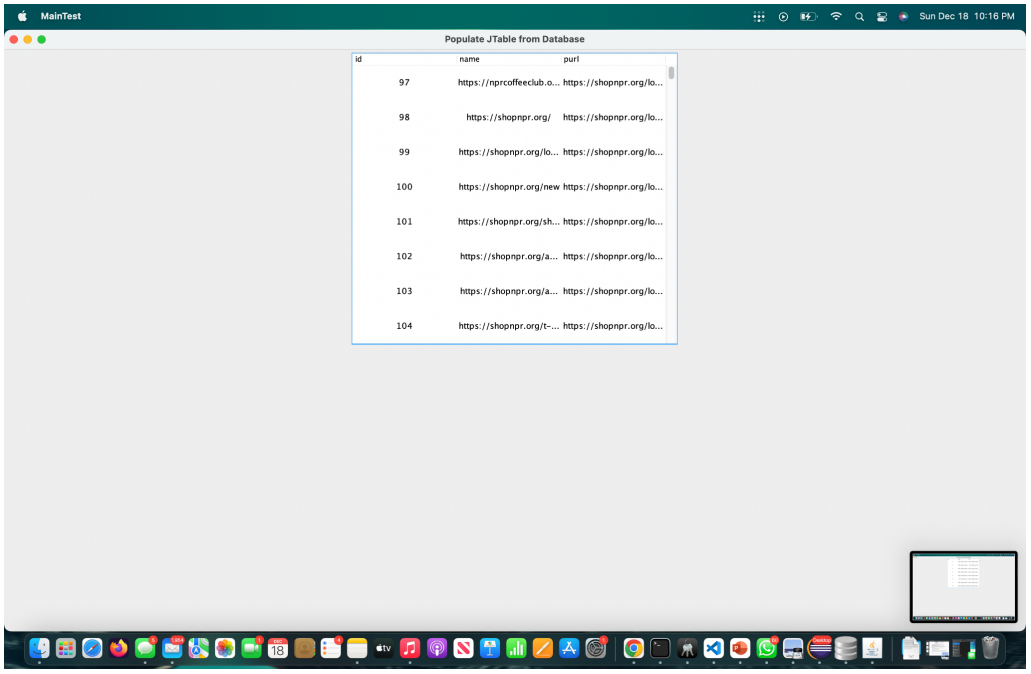
These results are for 4 bots with seed URLs:

1. https://abcnews.go.com
2. https://npr.org
3. https://nytimes.com
4. https://foxnews.com

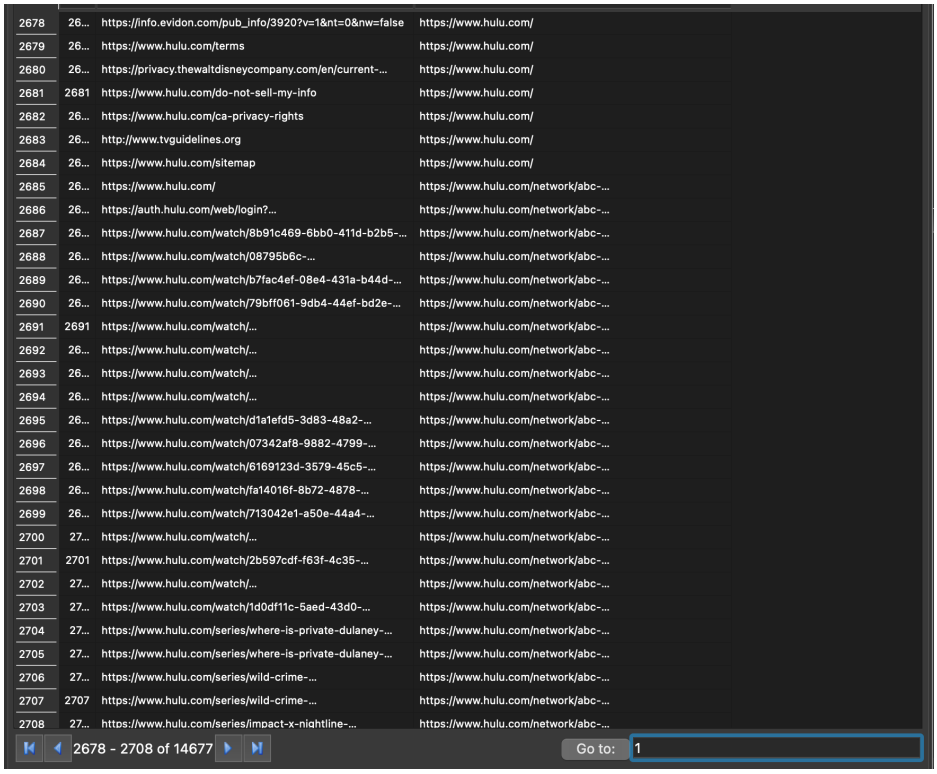
GUI Output:




MainTest
Sun Dec 18 10:17 PM
Populate JTable from Database
id name purl
829 https://www.npr.org/2... https://www.npr.org/
830 https://www.npr.org/s... https://www.npr.org/
831 https://www.npr.org/2... https://www.npr.org/
832 https://www.npr.org/2... https://www.npr.org/
833 https://www.npr.org/s... https://www.npr.org/
834 https://www.npr.org/2... https://www.npr.org/
835 https://www.npr.org/2... https://www.npr.org/
836 https://www.npr.org/s... https://www.npr.org/

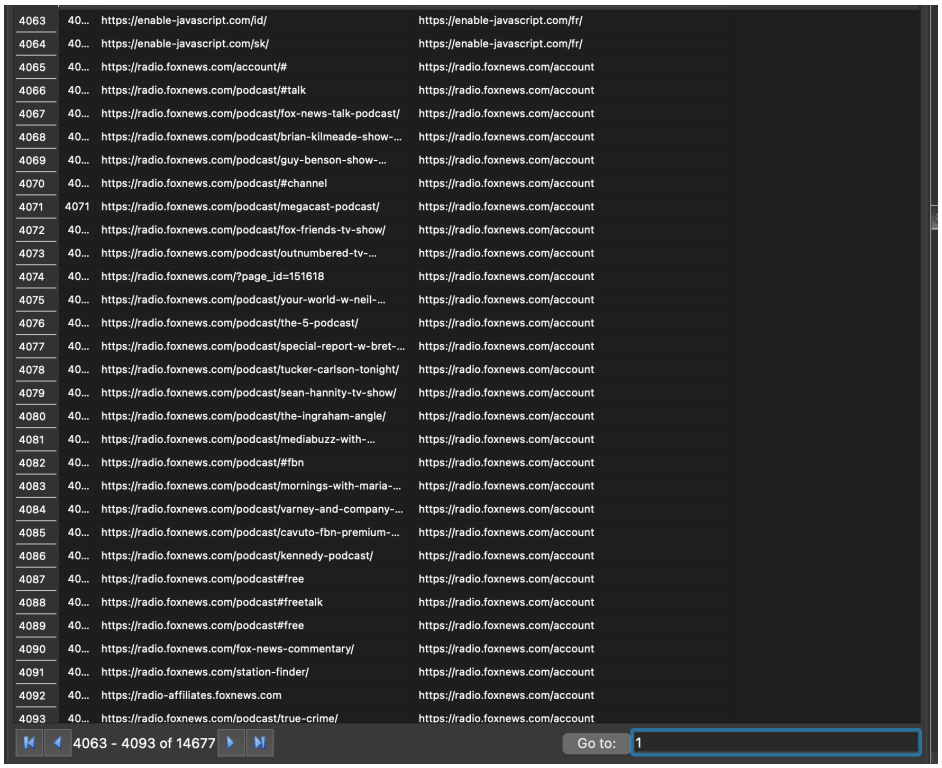

MainTest
Sun Dec 18 10:16 PM
Populate JTable from Database
id name purl
97 https://nprcoffeeclub.o... https://shopnpr.org/lo...
98 https://shopnpr.org/ https://shopnpr.org/lo...
99 https://shopnpr.org/lo... https://shopnpr.org/lo...
100 https://shopnpr.org/new https://shopnpr.org/lo...
101 https://shopnpr.org/sh... https://shopnpr.org/lo...
102 https://shopnpr.org/a... https://shopnpr.org/lo...
103 https://shopnpr.org/a... https://shopnpr.org/lo...
104 https://shopnpr.org/t-... https://shopnpr.org/lo...

SQLite database output:



2678	26...	https://info.evidon.com/pub_info/3920?v=1&nt=0&nw=false	https://www.hulu.com/
2679	26...	https://www.hulu.com/terms	https://www.hulu.com/
2680	26...	https://privacy.thewaltdisneycompany.com/en/current-...	https://www.hulu.com/
2681	2681	https://www.hulu.com/do-not-sell-my-info	https://www.hulu.com/
2682	26...	https://www.hulu.com/ca-privacy-rights	https://www.hulu.com/
2683	26...	http://www.tvguidelines.org	https://www.hulu.com/
2684	26...	https://www.hulu.com/sitemap	https://www.hulu.com/
2685	26...	https://www.hulu.com/	https://www.hulu.com/network/abc-...
2686	26...	https://auth.hulu.com/web/login?...	https://www.hulu.com/network/abc-...
2687	26...	https://www.hulu.com/watch/8b91c469-6bb0-411d-b2b5-...	https://www.hulu.com/network/abc-...
2688	26...	https://www.hulu.com/watch/08795b6c-...	https://www.hulu.com/network/abc-...
2689	26...	https://www.hulu.com/watch/b7fac4ef-08e4-431a-b44d-...	https://www.hulu.com/network/abc-...
2690	26...	https://www.hulu.com/watch/79bff061-9db4-44ef-bd2e-...	https://www.hulu.com/network/abc-...
2691	2691	https://www.hulu.com/watch/...	https://www.hulu.com/network/abc-...
2692	26...	https://www.hulu.com/watch/...	https://www.hulu.com/network/abc-...
2693	26...	https://www.hulu.com/watch/...	https://www.hulu.com/network/abc-...
2694	26...	https://www.hulu.com/watch/...	https://www.hulu.com/network/abc-...
2695	26...	https://www.hulu.com/watch/d1a1efd5-3d83-48a2-...	https://www.hulu.com/network/abc-...
2696	26...	https://www.hulu.com/watch/07342af8-9882-4799-...	https://www.hulu.com/network/abc-...
2697	26...	https://www.hulu.com/watch/6169123d-3579-45c5-...	https://www.hulu.com/network/abc-...
2698	26...	https://www.hulu.com/watch/fa14016f-8b72-4878-...	https://www.hulu.com/network/abc-...
2699	26...	https://www.hulu.com/watch/713042e1-a50e-44a4-...	https://www.hulu.com/network/abc-...
2700	27...	https://www.hulu.com/watch/...	https://www.hulu.com/network/abc-...
2701	2701	https://www.hulu.com/watch/2b597cdf-f63f-4c35-...	https://www.hulu.com/network/abc-...
2702	27...	https://www.hulu.com/watch/...	https://www.hulu.com/network/abc-...
2703	27...	https://www.hulu.com/watch/1d0df11c-5aed-43d0-...	https://www.hulu.com/network/abc-...
2704	27...	https://www.hulu.com/series/where-is-private-dulaney-...	https://www.hulu.com/network/abc-...
2705	27...	https://www.hulu.com/series/where-is-private-dulaney-...	https://www.hulu.com/network/abc-...
2706	27...	https://www.hulu.com/series/wild-crime-...	https://www.hulu.com/network/abc-...
2707	2707	https://www.hulu.com/series/wild-crime-...	https://www.hulu.com/network/abc-...
2708	27...	https://www.hulu.com/series/impact-x-nightline-...	https://www.hulu.com/network/abc-...

2678 - 2708 of 14677 Go to: 1



4063	40...	https://enable-javascript.com/id/	https://enable-javascript.com/fr/
4064	40...	https://enable-javascript.com/sk/	https://enable-javascript.com/fr/
4065	40...	https://radio.foxnews.com/account/#	https://radio.foxnews.com/account
4066	40...	https://radio.foxnews.com/podcast/#talk	https://radio.foxnews.com/account
4067	40...	https://radio.foxnews.com/podcast/fox-news-talk-podcast/	https://radio.foxnews.com/account
4068	40...	https://radio.foxnews.com/podcast/brian-kilmeade-show-...	https://radio.foxnews.com/account
4069	40...	https://radio.foxnews.com/podcast/guy-benson-show-...	https://radio.foxnews.com/account
4070	40...	https://radio.foxnews.com/podcast/#channel	https://radio.foxnews.com/account
4071	4071	https://radio.foxnews.com/podcast/megacast-podcast/	https://radio.foxnews.com/account
4072	40...	https://radio.foxnews.com/podcast/fox-friends-tv-show/	https://radio.foxnews.com/account
4073	40...	https://radio.foxnews.com/podcast/outnumbered-tv-...	https://radio.foxnews.com/account
4074	40...	https://radio.foxnews.com/?page_id=151618	https://radio.foxnews.com/account
4075	40...	https://radio.foxnews.com/podcast/your-world-w-neil-...	https://radio.foxnews.com/account
4076	40...	https://radio.foxnews.com/podcast/the-5-podcast/	https://radio.foxnews.com/account
4077	40...	https://radio.foxnews.com/podcast/special-report-w-bret-...	https://radio.foxnews.com/account
4078	40...	https://radio.foxnews.com/podcast/tucker-carlson-tonight/	https://radio.foxnews.com/account
4079	40...	https://radio.foxnews.com/podcast/sean-hannity-tv-show/	https://radio.foxnews.com/account
4080	40...	https://radio.foxnews.com/podcast/the-ingraham-angle/	https://radio.foxnews.com/account
4081	40...	https://radio.foxnews.com/podcast/mediabuzz-with-...	https://radio.foxnews.com/account
4082	40...	https://radio.foxnews.com/podcast/#fbn	https://radio.foxnews.com/account
4083	40...	https://radio.foxnews.com/podcast/mornings-with-maria-...	https://radio.foxnews.com/account
4084	40...	https://radio.foxnews.com/podcast/varney-and-company-...	https://radio.foxnews.com/account
4085	40...	https://radio.foxnews.com/podcast/cavuto-fbn-premium-...	https://radio.foxnews.com/account
4086	40...	https://radio.foxnews.com/podcast/kennedy-podcast/	https://radio.foxnews.com/account
4087	40...	https://radio.foxnews.com/podcast#free	https://radio.foxnews.com/account
4088	40...	https://radio.foxnews.com/podcast#freetalk	https://radio.foxnews.com/account
4089	40...	https://radio.foxnews.com/podcast#free	https://radio.foxnews.com/account
4090	40...	https://radio.foxnews.com/fox-news-commentary/	https://radio.foxnews.com/account
4091	40...	https://radio.foxnews.com/station-finder/	https://radio.foxnews.com/account
4092	40...	https://radio-affiliates.foxnews.com	https://radio.foxnews.com/account
4093	40...	https://radio.foxnews.com/podcast/true-crime/	https://radio.foxnews.com/account

4063 - 4093 of 14677 Go to: 1

Terminal Output: