CoC News Article Downloader and Viewer

Miasia Jones

CS 6675-Homework 2-Option 1.2: Write a Web Crawler of your own

Introduction

For this assignment I decided to crawl the news articles on the College of Computing (CoC) website. The seed URL is the main page of the CoC news page. Each page list news articles for a user to choose from and each article listed is linked to a URL to read the full article. I chose this seed URL because I knew I could easily access at least 1,000 URLs from it, being that there are currently 64 pages listing about 20 articles each. The publishing year range that these articles span is 2011-2021.

Crawler Design

The web crawler is designed to extract the title, date, and URL of every article listed on the CoC news pages. Each article URL is visited, and its HTML content is extracted. All this information is then saved into a .csv file. I created an app using PyQt5 that can crawl the CoC News website and manage the downloaded articles. You do not have to use the app to run the crawler. See README.md for more information. Below is the Pros and Cons for the design of my web crawler alone:

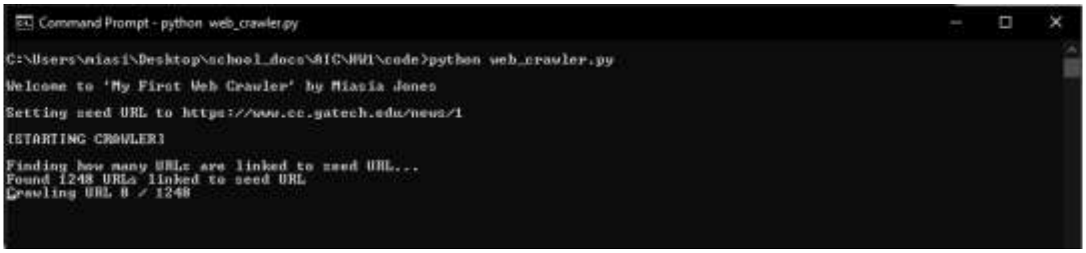
Pros:

- Crawler retrieves all articles from CoC News website, even the undated ones.
- Pages that are visited are not visited again.
- The console output continuously updates the user of its progress.

Cons:

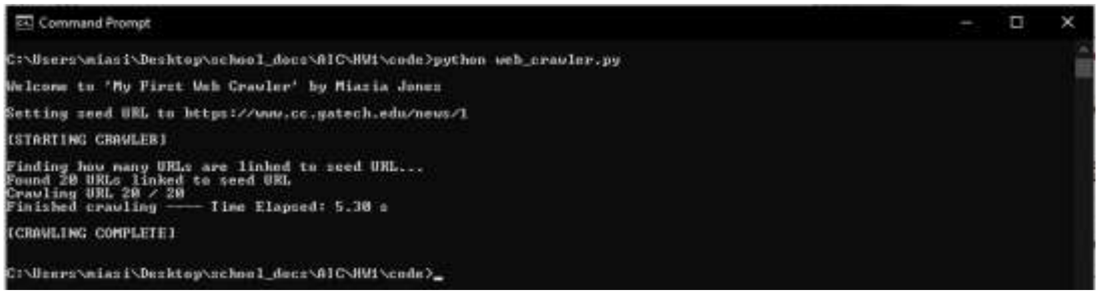
- The crawler calls specific class names in the HTML script of CoC News website to extract certain information, so if these class names were changed, then my crawler may not work.
- The crawler stores the HTML content for each article retrieved in a .csv file, which makes the file hard to read. Saving the HTML content does help within the app I created though.

Console Screenshots



```
C:\Users\miasi\Desktop\school_docs\AIC\HW1\code>python web_crawler.py
Welcome to 'My First Web Crawler' by Miasia Jones
Setting seed URL to https://www.cc.gatech.edu/news/1
[STARTING CRAWLER]
Finding how many URLs are linked to seed URL...
Found 1248 URLs linked to seed URL
Crawling URL 8 / 1248
```

Crawling all pages on CoC News page

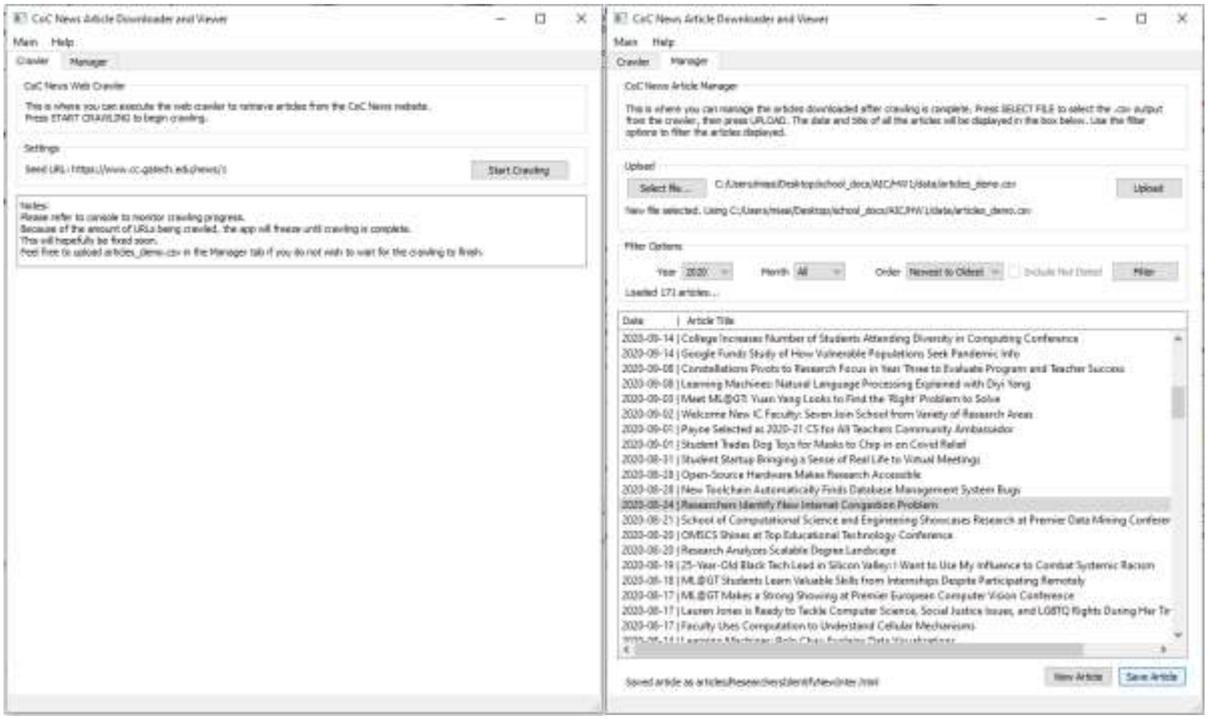


```
C:\Users\miasi\Desktop\school_docs\AIC\HW1\code>python web_crawler.py
Welcome to 'My First Web Crawler' by Miasia Jones
Setting seed URL to https://www.cc.gatech.edu/news/1
[STARTING CRAWLER]
Finding how many URLs are linked to seed URL...
Found 20 URLs linked to seed URL
Crawling URL 20 / 20
Finished crawling ---- Time Elapsed: 5.30 s
[CRAWLING COMPLETE]

C:\Users\miasi\Desktop\school_docs\AIC\HW1\code>_
```

Crawling only one page on CoC News page

App Screenshots



CoC News Web Crawler

This is where you can execute the web crawler to retrieve articles from the CoC News website.
Press START CRAWLING to begin crawling.

Settings

Seed URL: https://www.cc.gatech.edu/news/1

Notes:
Please refer to console to monitor crawling progress.
Because of the amount of URLs being crawled, the app will freeze until crawling is complete.
This will hopefully be fixed soon.
Feel free to upload articles_demo.csv in the Manager tab if you do not wish to wait for the crawling to finish.

CoC News Article Manager

This is where you can manage the articles downloaded after crawling is complete. Press SELECT FILE to select the .csv output from the crawler, then press UPLOAD. The date and title of all the articles will be displayed in the box below. Use the filter options to filter the articles displayed.

Upload

C:/Users/miasi/Desktop/school_docs/AIC/HW1/data/articles_demo.csv

New file selected. Using C:/Users/miasi/Desktop/school_docs/AIC/HW1/data/articles_demo.csv

Filter Options

Year 2020 Month All Order Newest to Oldest ☐ Include Not Dated

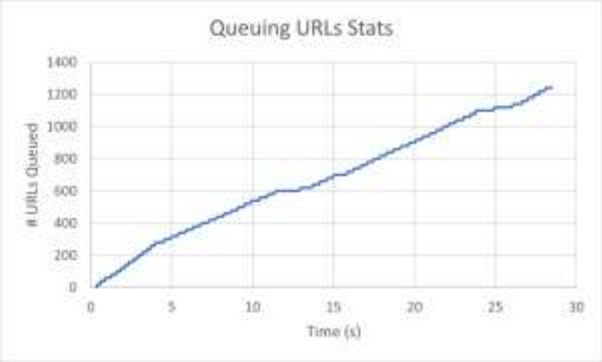
Loaded 171 articles...

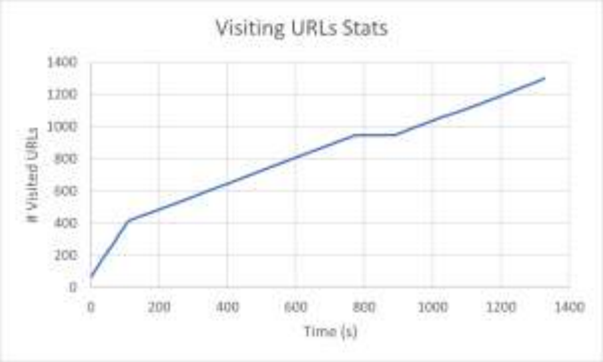
```
Date       | Article Title
2020-09-14 | College Increases Number of Students Attending Diversity in Computing Conference
2020-09-14 | Google Funds Study of How Vulnerable Populations Seek Pandemic Info
2020-09-08 | Constellations Pivots to Research Focus in Year Three to Evaluate Program and Teacher Success
2020-09-08 | Learning Machines: Natural Language Processing Explained with Diyi Yang
2020-09-03 | Meet ML@GT: Yuan Yang Looks to Find the 'Right' Problem to Solve
2020-09-02 | Welcome New IC Faculty: Seven Join School from Variety of Research Areas
2020-09-01 | Payne Selected as 2020-21 CS for All Teachers Community Ambassador
2020-09-01 | Student Trades Dog Toys for Masks to Chip in on Covid Relief
2020-08-31 | Student Startup Bringing a Sense of Real Life to Virtual Meetings
2020-08-28 | Open-Source Hardware Makes Research Accessible
2020-08-28 | New Toolchain Automatically Finds Database Management System Bugs
2020-08-24 | Researchers Identify New Internet Congestion Problem
2020-08-21 | School of Computational Science and Engineering Showcases Research at Premier Data Mining Confere
2020-08-20 | OMSCS Shines at Top Educational Technology Conference
2020-08-20 | Research Analyzes Scalable Degree Landscape
2020-08-19 | 25-Year-Old Black Tech Lead in Silicon Valley: I Want to Use My Influence to Combat Systemic Racism
2020-08-18 | ML@GT Students Learn Valuable Skills from Internships Despite Participating Remotely
2020-08-17 | ML@GT Makes a Strong Showing at Premier European Computer Vision Conference
2020-08-17 | Lauren Jones is Ready to Tackle Computer Science, Social Justice Issues, and LGBTQ Rights During Her Ti
2020-08-17 | Faculty Uses Computation to Understand Cellular Mechanisms
2020-08-13 | Learning Machines: Polo Chau Explains Data Visualizations
```

Saved article as articles/ResearchersIdentifyNewInter.html

Using the CoC News Article Downloader and Viewer

Crawl Statistics





These are statistics when all the URLs linked to CoC News website is crawled. There were 1,244 URLs found and visited. It took ≈28.47 seconds to find and queue all the URLs linked to the seed URL, in other words, ≈43.74 pages/second. It took ≈1326.76 seconds (≈22.11 minutes) to visit and scrape each URL in the queue, in other words, ≈0.94 pages/second.

Lessons Learned

- Using the CoC News website was not my first choice, but the structure of the website worked well for what I wanted to do with my crawler. This was my first time implementing a web crawler. My design was simple, but I like the outcome. I was able to brush up on my HTML skill and worked with Python packages I never used before. I was very invested in the assignment and didn’t mind the challenge. I’m not sure who would make use of the information I retrieved and I app I created, but I’m very happy with my work. The only thing new that I added that wasn’t present on the CoC News page is a news article filtering feature.

- To crawl 10 million pages, I predict that my crawler would take about 4 months. To crawl 1 billion pages, I predict that my crawler would take about 30 years.