

# CS 4675 Project 1


## Executable and code bade

Executable: located in ./postscape/output/scrapy/scrapy.exe

Run from ./postscape scrapy crawl extractor

Python code for scraper in postscrape\postscrape\spiders\post\_spyder.py

Keyword list is in postscrape\current\_stats\_keyword\_current\_output as a csv mapping keyword to list of keywords



```
File Edit Format View Help
computing,"['https://www.cc.gatech.edu', 'https://www.cc.gatech.edu/index.php/division-computing-instruction',
'https://www.cc.gatech.edu/index.php/school-computer-science', 'https://www.cc.gatech.edu/index.php/calendar',
'https://www.cc.gatech.edu/index.php/school-computational-science-and-engineering',
'https://www.cc.gatech.edu/index.php/events', 'https://www.cc.gatech.edu/index.php/school-interactive-computing',
'https://www.cc.gatech.edu/index.php/computing-career-services', 'https://www.cc.gatech.edu/index.php/mentoring-program',
'https://www.cc.gatech.edu/index.php/office-outreach-enrollment-and-community',
'https://www.cc.gatech.edu/index.php/international-study', 'https://www.cc.gatech.edu/index.php/tutoring-assistance',
'https://www.cc.gatech.edu/index.php/entrepreneurship-gt-computing', 'https://www.cc.gatech.edu/index.php/content/career-
fairs-0', 'https://www.cc.gatech.edu/news/nsf-grant-could-lead-better-computational-tools-human-fact-checkers',
'https://www.cc.gatech.edu/news/new-women-technology-scholarship-recipient-had-early-interest-ai',
'https://www.cc.gatech.edu/news/newly-named-founding-chair-recognizes-cybersecurity-revolves-around-problems-not-academic',
'https://www.cc.gatech.edu/index.php/school-cybersecurity-and-privacy', 'https://www.cc.gatech.edu/index.php/news',
'https://www.cc.gatech.edu/computing-career-services', 'https://www.cc.gatech.edu/tutoring-assistance',
'https://www.cc.gatech.edu/index.php/support-college', 'https://www.cc.gatech.edu/news/648961/new-global-top-5-ranking-
based-faculty-research-success', 'https://www.cc.gatech.edu/international-study', 'https://www.cc.gatech.edu/author/ann-
claycombe', 'https://www.cc.gatech.edu/index.php/events/2022/02/10/school-cybersecurity-privacy-student-town-hall',
'https://www.cc.gatech.edu/unit/computational-science-engineering',
'https://www.cc.gatech.edu/index.php/events/2022/02/22/power-two',
'https://www.cc.gatech.edu/index.php/calendar/day/20220222', 'https://www.cc.gatech.edu/event/category/training-workshop',
'https://www.cc.gatech.edu/event/category/conference-symposium', 'https://www.cc.gatech.edu/event/category/seminar-
lecture-colloquium', 'https://www.cc.gatech.edu/unit/school-cybersecurity-and-privacy',
'https://www.cc.gatech.edu/event/category/career-professional-development', 'https://www.cc.gatech.edu/author/david-
mitchell', 'https://www.cc.gatech.edu/unit/college-computing', 'https://www.cc.gatech.edu/index.php/news?page=56',
'https://www.cc.gatech.edu/author/ben-snedeker', 'https://www.cc.gatech.edu/unit/omscs',
'https://www.cc.gatech.edu/index.php/calendar/day/20220223', 'https://www.cc.gatech.edu/index.php/calendar/day/20220221',
'https://www.cc.gatech.edu/unit/computational-science-engineering?page=1',
'https://www.cc.gatech.edu/taxonomy/term/301/feed', 'https://www.cc.gatech.edu/author/ann-claycombe?page=1',
'https://www.cc.gatech.edu/taxonomy/term/293/feed', 'https://www.cc.gatech.edu/taxonomy/term/63/feed',
'https://www.cc.gatech.edu/taxonomy/term/300/feed', 'https://www.cc.gatech.edu/author/david-mitchell?page=1',
'https://www.cc.gatech.edu/index.php/news?page=53', 'https://www.cc.gatech.edu/event/category/career-professional-
development?page=1', 'https://www.cc.gatech.edu/taxonomy/term/308/feed',
'https://www.cc.gatech.edu/taxonomy/term/298/feed', 'https://www.cc.gatech.edu/taxonomy/term/307/feed',
'https://www.cc.gatech.edu/index.php/news?page=54', 'https://www.cc.gatech.edu/index.php/calendar/day/20220220'.
```

## Design

The algorithm runs BFS by enqueueing new links extracted from web pages. I started with a seed URL of <https://www.cc.gatech.edu>. Initially, I limited the scraper to only enqueue pages to visit in the domain of 'cc.gatech.edu'; however, after long enough, we were scraping archived info from early 2000s class websites- I discovered a whole subset of sites with a prefix (ABC.cc.gatech.edu, XYZ.cc.gatech.edu). After ~40,000 results (in postscrape/ oldstats), I realized very few pages were being scraped with pertinent info, and shifted my strategy.

I then forced the domain of the website to be [www.cc.gatech.edu](http://www.cc.gatech.edu), which helped somewhat in limiting my results.

Checks for keywords from dictionary, if found keywords are added to a CSV updated every 100 pages crawled containing mappings of keyword to list of website containing said words

The pros of running BFS is that we will reach a wide subset of pages, and explore many avenues we would not have reached initially. "Important pages" were reached quickly as we explored the links on our first page first, then on each subsequent page without running down a rabbit hole on the first link we found.

For cons, since I did not have a max recursion like I would in DFS, all links from each page were added to the queue. This ended up being time consuming and resulted in a lot of pages being visited. As there was also no filtering past the '[www.cc.gatech.edu](https://www.cc.gatech.edu)', many of these pages were of little interest- for example, `index.php/calender` links did not need to be scraped, but were.

## Screenshots of crawler

```
https://www.cc.gatech.edu/calendar/week/203230)
2022-02-03 11:34:00 [extractor] DEBUG: crawling
2022-02-03 11:34:00 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/calendar/203230> (referer: https://www.cc.gatech.edu/calendar/week/203230)
2022-02-03 11:34:01 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/index.php/calendar/day/200947> (referer: https://www.cc.gatech.edu/index.php/calendar/week/200947)
2022-02-03 11:34:01 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/calendar/day/203230> (referer: https://www.cc.gatech.edu/calendar/week/203230)
2022-02-03 11:34:01 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/index.php/calendar/year/200947> (referer: https://www.cc.gatech.edu/index.php/calendar/week/200947)
2022-02-03 11:34:01 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/calendar/year/203230> (referer: https://www.cc.gatech.edu/calendar/week/203230)
2022-02-03 11:34:01 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.cc.gatech.edu/index.php/calendar/week/200946> (referer: https://www.cc.gatech.edu/index.php/calendar/week/200947)
2022-02-03 11:34:01 [extractor] DEBUG: crawling
2022-02-03 11:34:01 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/index.php/calendar/200947> (referer: https://www.cc.gatech.edu/index.php/calendar/week/200947)
2022-02-03 11:34:03 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/calendar/year/203237> (referer: https://www.cc.gatech.edu/calendar/week/203237)
2022-02-03 11:34:03 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.cc.gatech.edu/calendar/week/203236> (referer: https://www.cc.gatech.edu/calendar/week/203237)
2022-02-03 11:34:03 [extractor] DEBUG: crawling
2022-02-03 11:34:04 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/index.php/calendar/200918> (referer: https://www.cc.gatech.edu/index.php/calendar/week/200918)
2022-02-03 11:34:04 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/calendar/day/203237> (referer: https://www.cc.gatech.edu/calendar/week/203237)
2022-02-03 11:34:04 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/calendar/203237> (referer: https://www.cc.gatech.edu/calendar/week/203237)
2022-02-03 11:34:04 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.cc.gatech.edu/index.php/calendar/year/200904> (referer: https://www.cc.gatech.edu/index.php/calendar/200904)
2022-02-03 11:34:04 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.cc.gatech.edu/index.php/calendar/200903> (referer: https://www.cc.gatech.edu/index.php/calendar/200904)
2022-02-03 11:34:04 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.cc.gatech.edu/calendar/week/203349> (referer: https://www.cc.gatech.edu/calendar/week/203350)
2022-02-03 11:34:05 [extractor] DEBUG: crawling
2022-02-03 11:34:05 [extractor] DEBUG: crawling
```

*Crawler command line running*

```
start_time,2022-02-03 06:39:31.676363
scheduler/enqueued/memory,5941
scheduler/enqueued,5941
scheduler/dequeued/memory,4809
scheduler/dequeued,4809
downloader/request_count,4860
downloader/request_method_count/GET,4860
downloader/request_bytes,1526107
robotstxt/request_count,27
downloader/response_count,4837
downloader/response_status_count/200,3335
downloader/response_bytes,371383058
```

*Intermediate CSV outputs w/ stats*

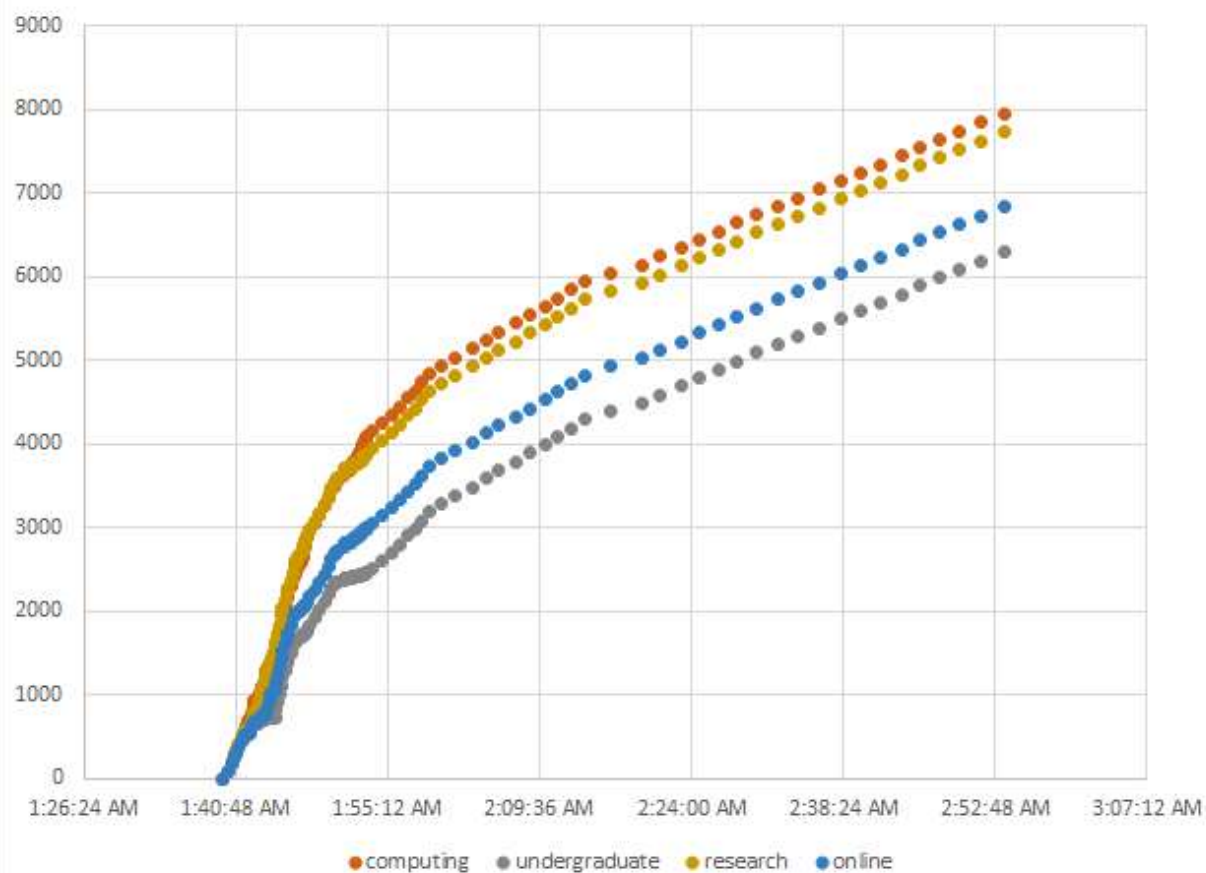
```
computing,1093
undergraduate,702
research,1046
online,746
current_time,2022-02-03 01:43:14.807638
```

*Keyword extraction stats at the same point*

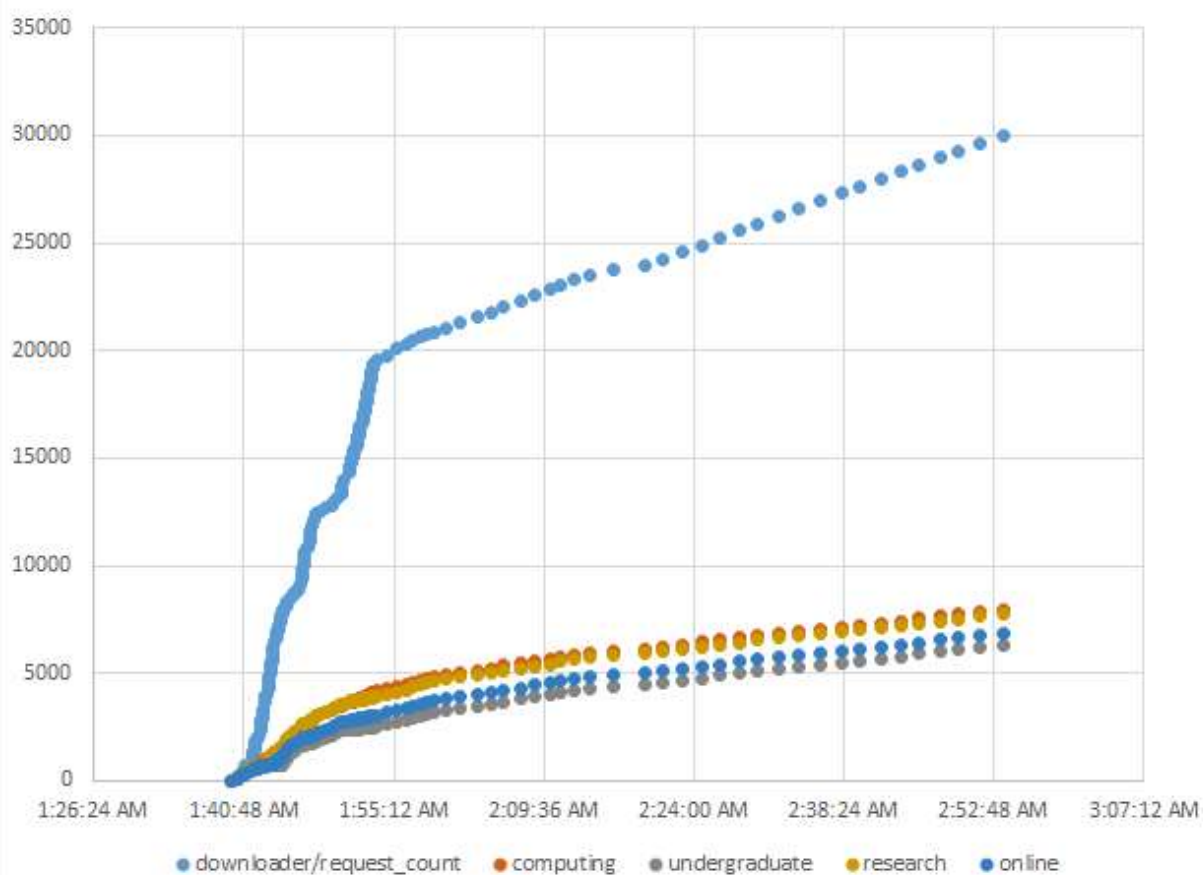
## Stats

---

Keywords Found

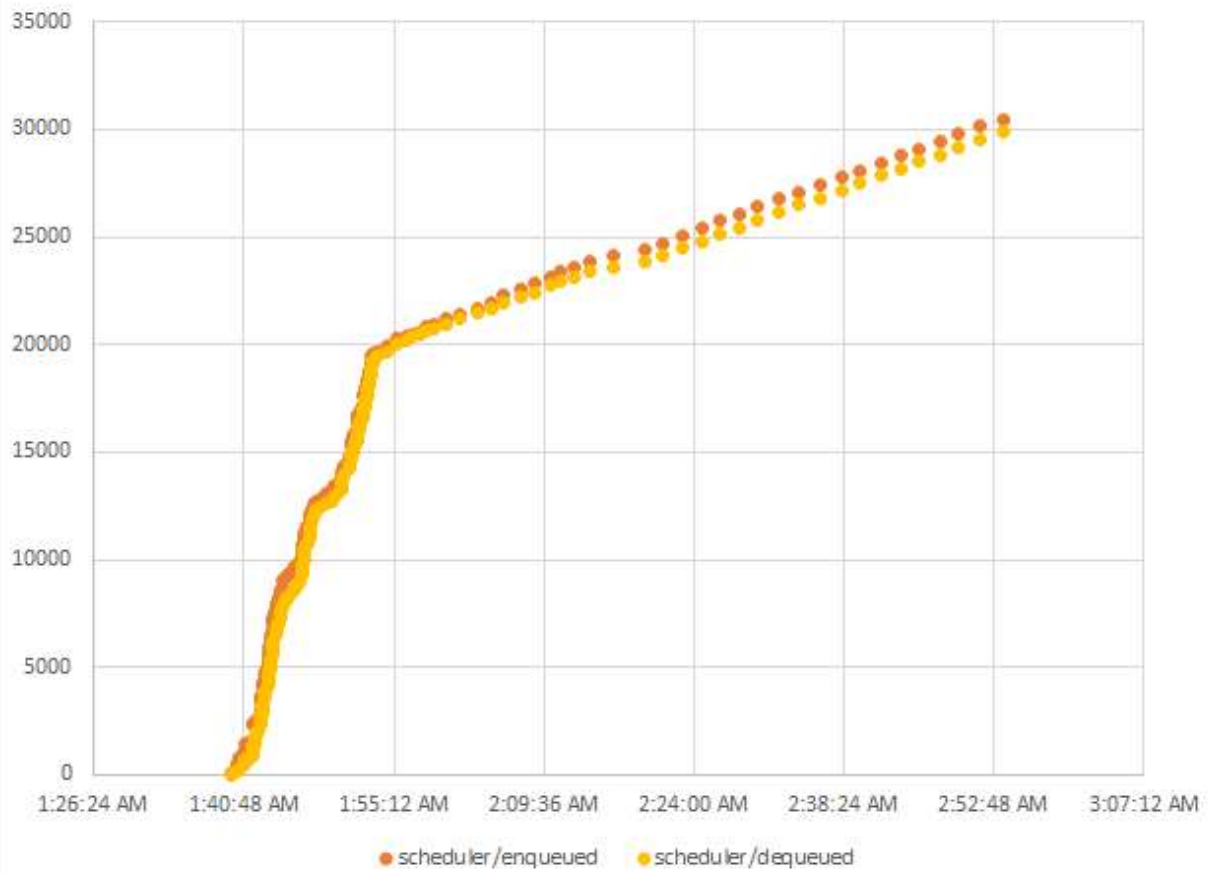


Keywords Found

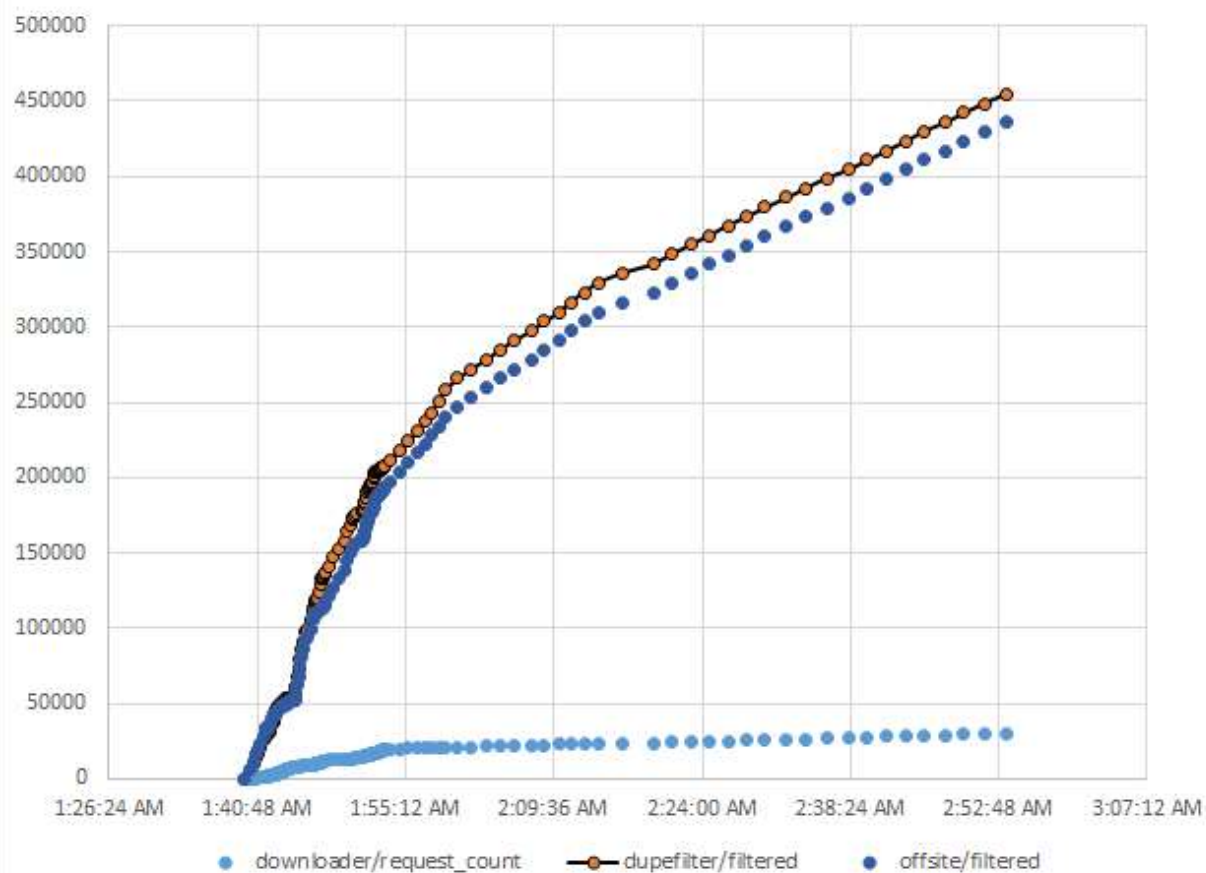




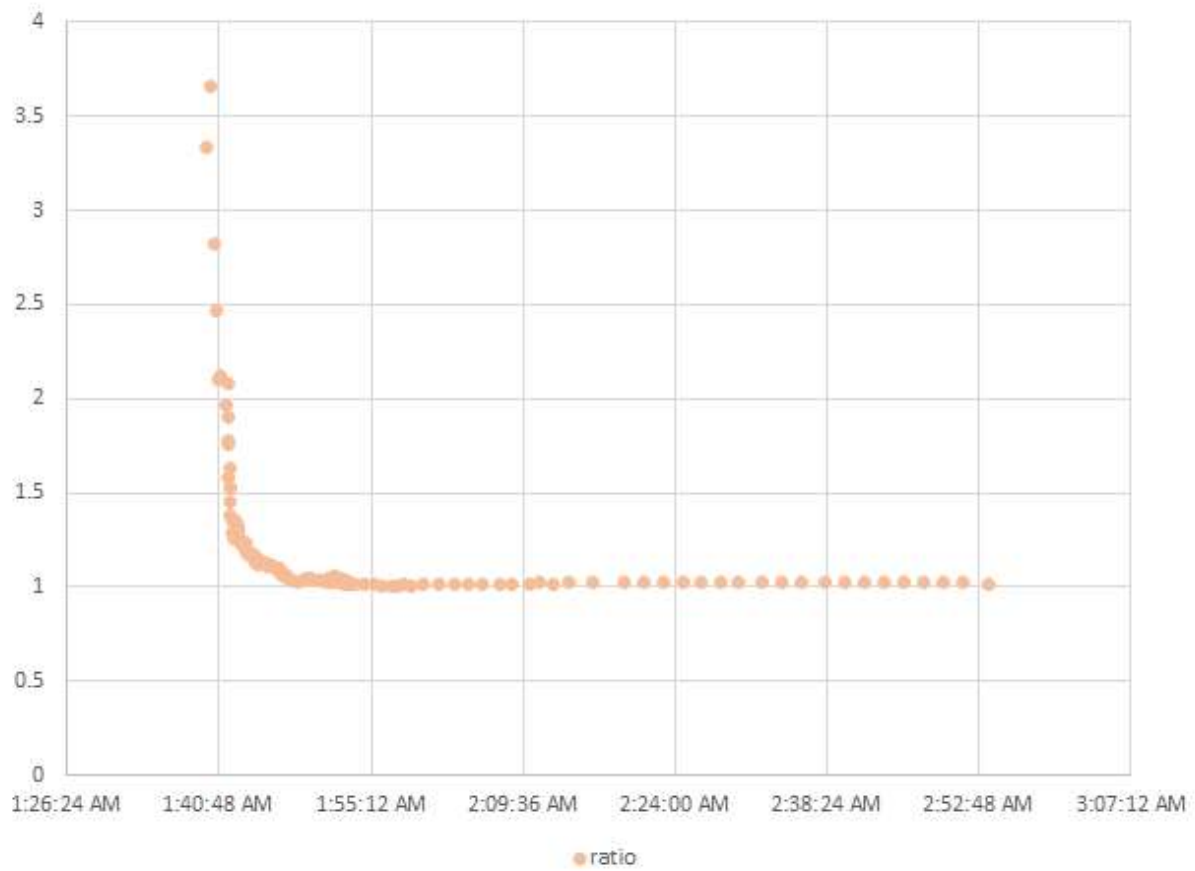
Enqueue vs. Dequeue



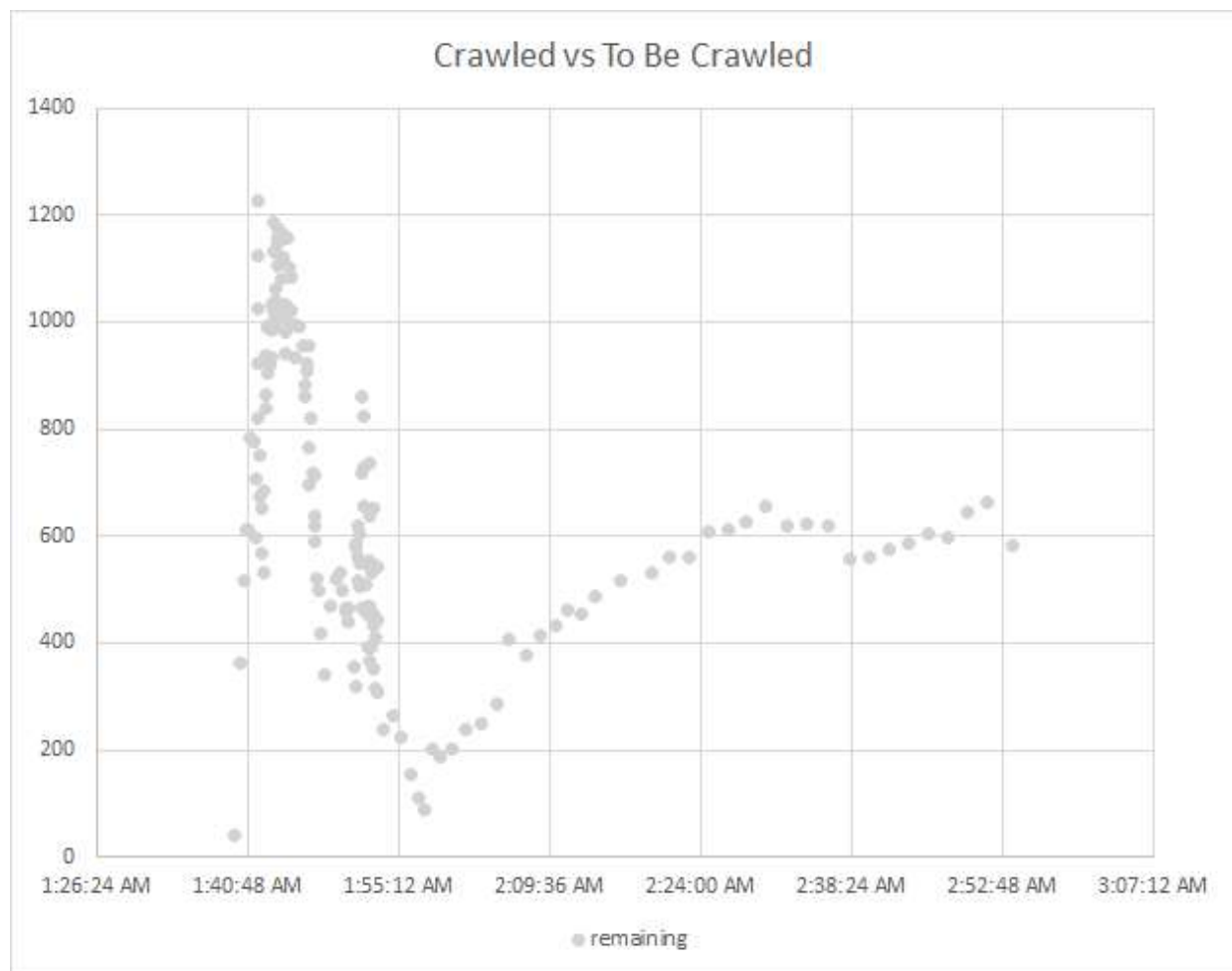
Filtered sites



Crawled vs To Be Crawled







My computer did struggle to run consistently, but I was able to monitor its status for a while. These are the key findings I discovered. As time went on, more and more sites were either already visited or were outside of the supplied domain. Computing was the most common word found, and undergrad was less common. A log pattern is seen, with initial sites containing these words very frequently and less so later on.

My crawler showed struggles with filtering through already visited links, which I think is due to running BFS and not having a max recursion property like I would on a DFS algorithm. As a result, pages linking back to popular pages were processed over and over. My ratio of enqueued vs dequeued over time gets closer and closer to 1, although this is expected as the dequeued pages builds up. More interesting is the last graph in grey, showing how many pages are remaining. It seems to steady out around 600, it did gradually go down in future data (I resumed running at a different time point, which would have broken the axis)

Overall, it would have taken a long time to complete given how quickly the filtered out websites were growing, while we were still enqueueing websites that do need to be visited at a steady rate. I would predict a log scale of time for how long this would take, with perhaps 10x as long to scrape 2x the amount of current website it has scraped. Reaching a million/ billion seems infeasible given the lack of constraint of depth of the search.

## Lessons Learned

I learned how hard scraping is, and how brute force isn't feasible. I also learned the importance of supplying good filters on sites to visit, as my first try allowed any domain preceding cc.gatech.edu to be scraped. I saw how much computing power is required to scrape, and how getting an initial set of sites isn't too hard, but scraping everything is extremely challenging due to repeated links/ external links. This project showed me just how important a smart algorithm is in scraping.

## Resources

<https://www.youtube.com/watch?v=ALizgnSFTwQ>

<https://docs.scrapy.org/en/latest/topics/spiders.html>