Alexander De Laurentiis
40050385

Report

Crawler: -> selfCrawler.py

So for the crawler I decided to only use the requests and urllib.parse libraries. The primary reasons being that I wanted to look into more how a crawler actually functions than use a crawler library, and because of my bad experiences of using libraries and being unable to debug… aka BeautifulSoup on project 1 would only give me bugs. I have an embedded function in the main crawling function to scrape the page that will check first if it is a file, if it is not a file type that I have blacklisted, and the url is of the correct “https://” format it will check if it had crawled it. If it hasn’t crawled it it will then requests.get() it, parse out any urls by retrieving all the href ID values, then yield the object holding its ID, URL, and Raw text to be preprocessed and passed into SPIMI. This way since it will always be yielding the data right into the stream it will never all be in memory, I had tested this by crawling 10,000 pages from concordia with a politeness interval(So it took a day…) and it successfully and at a very constant rate downloaded the desired pages discarding any faulty links.

Extractor: - extractor.py

I added a few methods to my file for page extraction/parsing to assist with this project. The most complicated addition was one to extract text based on the value of an ID,this way I could decide later what the most important class values or ID values were to extract from the concordia page everything between the tags with those IDs, such as class=”text-main”. I also added a function to remove any tag values from a body of text and combine all the center info, this way I could retrieve all the main body text then remove all the inner tags and keep their values. Then I made a method that would extract the tag values import just to the concordia raw page data, as well as the keywords in the meta tags at the top of each page for searching.

Indexer: indexing.py

I changed up the tokenStream method to take Page objects and correctly discern the needed page info, parse the Raw text, save some info for BM25, and yield tokens to SPIMI. SPIMI remained unchanged from last project.

Page Objects: -> fileData.py

Here I created a class of Page, which when passed a URL, ID, and Raw text would create a page object and save it all, as well as parse all the URLs from the text. This way I could neatly packaged yield the data between the crawler and tokenizer.

Front/BackEnd: -> server.py, base.html

So I decided to create a very simple flask webpage for a gui to take input, then using AJAX send it to the server, call the correct query processing method, and then display the links on the web page. I decided this would make it much simpler and be more practical in the long run, and allowed ease of querying. This can be run by just python server.py in the command line.

Query Processer: queryProcessor.py

For the queries I first collected the OR query of the documents before ranking each document in the query. Once ranked with BM25 properly this time I would order the list based on ranking of documents in descending order then remove the ranking, finally returning the ordered list to have the first 15 displayed.

Problems that Could be improved further

One problem faced in the end that if I had time to solve would be the lack of spaces between certain words when preprocessing which caused on occasion large conglomerates of 50+letters in the index, I have fixed it in the code I believe now, but in order to be sure I would need to reparse the index, which while following politeness principles and not multithreading would take a while. There could also be a time.sleep() function added into the crawler for when it is crawling a singular domain, as right now it will continue to cycle through possible URLS until the politeness interval is up, this would save processing power and may still be implemented.

Some of the Sorting algorithms could be optimized for speed for finding postings lists, and if allowed to use other data formats where the index may be encrypted would be much faster than json or parsing a txt file. I also did not pass the df through with the term, as that would create an impossible if not troublesome situation to search the exact pair, especially with the dict data structure. Also the df can be determined by the length of the posting list so it is already present data. If using a sql type structure this would be easy, and scalable to likely the whole site of concordia since search would be on disk rather than in memory. Also write time and load time would likely be more efficient than using json/txt files.

The indexing process can be called by calling the indexer.py file with any of my designated cmd parameters, there are errors where the values are saved and not passed through to the actual file though for no seen reason so I have refrained from fully implementing them. Ex: passing it through with "python indexer.py -M 10" the -M indicating max download 10 pages, the value would be successfully passed and able to be print yet the crawler would continue much much farther seemingly infinitely, yet if I hardcoded passing 10 to the crawler as a parameter it would stop at 10 pages. For this reason the arg.parsers are in the file yet not fully implemented once they may reliably work. No error was given too so there is nothing visible to debug...

For ranking schemes I find BM25 vs just a term frequency one is much better. Especially if allowing an OR query and ordering by how many terms per doc, that can lead to inaccurate results when a term isn't in the document and the query has a few common terms. While using the BM25 documents with the more common terms but more of the terms would be ranked less than the docs with the rarer terms.

QUERIES

So in the end my index did something funny yet especially inconvenient, it should be fixed but I didn't have the time to reconstruct my index without removing the politeness timer. Everytime the word "faculty" it got smushed together with a bunch of useless letters in the parsing……… So the challenge queries may have a skewed accuracy since they didn't include that word. Since the initial approach I had was collapsing hyphens as well no document with "sars-cov" appeared, there are multiple documents indexed in the 10,000 though with "sars-cov-2" and a singular document with "sars", so if I rather replaced the hyphen with a space rather than collapse it it would have had more matches.

1. covid-19 research academics
   https://www.concordia.ca/cunews/offices/advancement/2020/05/11/meet-angelo-filosa-a-grad-whose-company-provides-covid-19-test-kits-and-research-solutions.html?c=/artsci/chemistry/news
   https://www.concordia.ca/academics.html
   https://www.concordia.ca/coronavirus/students.htm
   https://www.concordia.ca/coronavirus/students.html#tuition
   https://www.concordia.ca/coronavirus/students.html#in-person-activities
   https://www.concordia.ca/news/stories/2020/10/27/concordia-launches-its-sustainability-action-plan.html?c=/jmsb/news/archive
   https://www.concordia.ca/news/stories/2020/10/27/concordia-launches-its-sustainability-action-plan.html
   https://www.concordia.ca/news/topic.html?topic=topics:academic_disciplines/environment
   https://www.concordia.ca/fr/coronavirus/annonces.html
   https://www.concordia.ca/cunews/encs/2020/11/20/Funding-for-ultrasound-technology-research.html?c=/research/perform/news-events
   https://www.concordia.ca/coronavirus.html
   https://www.concordia.ca/cunews/main/stories/2020/04/23/concordias-institute-for-urban-futures-moves-its-sex-workers-rights-conference-online.html?c=/news/authors/colin-throness
   https://www.concordia.ca/research/coronavirus-faq.html
   https://www.concordia.ca/students/registration.html
   https://www.concordia.ca/content/concordia/en/students/registration

   1. researchers covid-19 covid staff

   https://www.concordia.ca/fr/coronavirus/annonces.html

https://www.concordia.ca/cunews/offices/advancement/2020/11/04/meet-two-concordia-grads-on-the-front-lines-of-covid-19-care.html?c=/alumni-friends/news/community-vs-covid-19
https://www.concordia.ca/alumni-friends/news/community-vs-covid-19.html
https://www.concordia.ca/coronavirus.html
https://www.concordia.ca/coronavirus.html#dates
https://www.concordia.ca/news/stories/2020/12/03/concordias-reuse-centre-wants-to-deliver-some-free-inspiration.html?c=/alumni-friends/news/community-vs-covid-19
https://www.concordia.ca/news/stories/2020/12/03/concordias-reuse-centre-wants-to-deliver-some-free-inspiration.html?c=/news/archive
https://www.concordia.ca/research/coronavirus-faq.html
https://www.concordia.ca/finearts/facilities/academic-research-support.html
https://www.concordia.ca/academics/experiential-learning.html
https://www.concordia.ca/news/stories/2020/11/02/concordia-researchers-build-web-based-tool-to-reduce-risk-of-indoor-sars-cov2-transmission.html?c=/coronavirus
https://www.concordia.ca/news/stories/2020/04/29/a-massive-move-to-online-instruction-provides-a-challenge-and-an-opportunity-for-concordias-educational-technology-students.html?c=/news/topic
https://www.concordia.ca/offices/ci.html
https://www.concordia.ca/it/support/learn-teach-work-from-home.html
https://www.concordia.ca/finearts/facilities/studio-support/core-technical-centres.html

2. departments research environmental issues

https://www.concordia.ca/campus-life/security/emergency/cert/erp.html
https://www.concordia.ca/artsci/loyola-college-diversity-sustainability/Sustainability-and-the-climate-crisis/lca.html
https://www.concordia.ca/campus-life/safety/coronavirus.html
http://www.concordia.ca/campus-life/safety/coronavirus.html
https://www.concordia.ca/cunews/main/stories/2019/06/17/the-canadian-society-for-civil-engineering-honours-osama-moselhis-40-years-of-contributions.html?c=/news/topic
https://www.concordia.ca/artsci/loyola-college-diversity-sustainability/about.html
https://www.concordia.ca/about/administration-governance/office-vp-services/vp-services-annual-report/our-leadership-team/health-safety.html
https://www.concordia.ca/artsci/economics/econ-policy.html
https://www.concordia.ca/artsci/geography-planning-environment/programs/graduate/geography-urban-environmental-phd.html
https://www.concordia.ca/academics/experiential-learning/opportunities/work/internships.html
https://www.concordia.ca/hr/training-development/workshops-courses.html

https://www.concordia.ca/academics/undergraduate/calendar/current/sec71/71-50.html#b71.50.2
https://www.concordia.ca/academics/undergraduate/calendar/current/sec71/71-50.html
https://www.concordia.ca/academics/undergraduate/calendar/current/sec71/71-50.html#options
https://www.concordia.ca/academics/undergraduate/calendar/current/sec71/71-50.html#b71.50.1

2. departments sustainability energy water conservation

https://www.concordia.ca/about/sustainability/contacts.html
https://www.concordia.ca/campus-life/food-services/eating-responsibly/sustainable-eating.html
https://www.concordia.ca/academics/undergraduate/calendar/current/sec71/71-60.html#iadi
https://www.concordia.ca/artsci/loyola-college-diversity-sustainability/Sustainability-and-the-climate-crisis/lca.html
https://www.concordia.ca/academics/graduate/calendar/current/encs/engineering-courses.html#INDU-MECH
https://www.concordia.ca/about/administration-governance/office-vp-services/vp-services-annual-report/our-leadership-team/health-safety.html
https://www.concordia.ca/academics/undergraduate/calendar/current/sec31/31-130.html#Geography
https://www.concordia.ca/academics/undergraduate/calendar/current/sec31/31-130.html#Geology
https://www.concordia.ca/academics/undergraduate/calendar/current/sec31/31-130.html
https://www.concordia.ca/research/water-energy/expertise.html
https://www.concordia.ca/artsci/biology/students/advising-support.html
https://www.concordia.ca/artsci/loyola-college-diversity-sustainability/Sustainability-in-the-city-and-beyond.html
https://www.concordia.ca/news/stories/2020/07/29/urban-water-consumption-will-increase-due-to-climate-change-concordia-research-shows.html?c=/news/topic
https://www.concordia.ca/research/water-energy.html
https://www.concordia.ca/about/sustainability/study-teach/sustainable-courses.html

Challenge Queries

1. Q: water management sustainability concordia

-https://www.concordia.ca/artsci/research/loyola-sustainability/people.html#students
-https://www.concordia.ca/about/sustainability/contacts.html
-https://www.concordia.ca/research/water-energy/expertise.html
-https://www.concordia.ca/faculty/chunjiang-an.html
-https://www.concordia.ca/ginacody/building-civil-environmental-eng/faculty.html?fpid=chunjiang-an
-https://www.concordia.ca/cunews/main/stories/2020/05/14/concordia-invests-in-a-zero-waste-culture-change-campaign.html
-https://www.concordia.ca/cunews/main/stories/2020/05/14/concordia-invests-in-a-zero-waste-culture-change-campaign.html?rootnav=news/stories
-https://www.concordia.ca/artsci/research/loyola-sustainability/people.html
-https://www.concordia.ca/about/sustainability/study-teach/sustainable-courses.html
-https://www.concordia.ca/offices/facilities.html
-https://www.concordia.ca/about/sustainability/action-plan/2020-2025/research.html
-https://www.concordia.ca/academics/undergraduate/calendar/current/sec31/31-525.html
-https://www.concordia.ca/ginacody/building-civil-environmental-eng/programs/environmental-eng/meng.html
-https://www.concordia.ca/campus-life/security/emergency/cert/erp.html
-https://www.concordia.ca/jmsb/programs/undergraduate/sustainable-investing-practicum/program-overview.html

2. Concordia covid-19 faculty

-https://www.concordia.ca/students/international/coronavirus.html
-https://www.concordia.ca/fr/coronavirus/annonces.html
-https://www.concordia.ca/coronavirus.html
-https://www.concordia.ca/coronavirus.html#dates
-https://www.concordia.ca/cunews/encs/2020/11/20/Funding-for-ultrasound-technology-research.html?c=/research/perform/news-events
-https://www.concordia.ca/offices/archives/covid-19-web-collection.html
-https://www.concordia.ca/coronavirus/faculty-staff.html
-https://www.concordia.ca/fr/coronavirus.html
-https://www.concordia.ca/fr/coronavirus.html#content-main_grid_container_170497148
-https://www.concordia.ca/cunews/main/stories/2020/07/06/concordia-supports-a-stranded-international-student-through-the-covid-19-pandemic.html?c=/artsci/economics/news
-https://www.concordia.ca/cunews/offices/advancement/2020/05/11/meet-angelo-filosa-a-grad-whose-company-provides-covid-19-test-kits-and-research-solutions.html?c=/artsci/chemistry/news
-https://www.concordia.ca/cunews/artsci/2020/04/22/how-the-covid-19-crisis-transformed-a-concordia-health-course.html?c=artsci/applied-human-sciences/news

-https://www.concordia.ca/alumni-friends/news/community-vs-covid-19.html
-https://www.concordia.ca/cunews/main/stories/2020/12/07/mitacs-and-concordia-partner-to-offer-the-first-ever-business-strategy-internship.html?c=/jmsb/about
-https://www.concordia.ca/coronavirus/wellbeing/symptoms.html

3. SARS-CoV Concordia faculty
-https://www.concordia.ca/content/concordia/en/campus-life/recreation
-https://www.concordia.ca/news/stories/2019/10/24/2-concordia-students-land-at-the-2019-paris-air-show.html?c=/news/topic
-https://www.concordia.ca/fr/actualites/nouvelles/2019/05/15/des-jeunes-de-dix-communautes-autochtones-se-reunissent-a-concordia-dans-le-cadre-de-startup-nations.html?c=/a-propos/engagement-communautaire/nouvelles
-https://www.concordia.ca/research/perform/about/jobs.html
-https://www.concordia.ca/hr/benefits.html
-https://www.concordia.ca/artsci/coms/research/phd-bios.html
-https://www.concordia.ca/ctl/about.html
-https://www.concordia.ca/cunews/jmsb/executive-centre/blog/2020/09/16/a-new-concordia-project-is-aggregating-environmental-knowledge.html
-https://www.concordia.ca/academics/undergraduate/calendar/current/sec16/16.html#b16.1.9
-https://www.concordia.ca/finearts/art-education/research/faculty-projects.html
-https://www.concordia.ca/fr/actualites/nouvelles/2019/06/11/les-parents-doues-pour-la-langue-complimentent-davantage-leurs-enfants-et-contribuent-au-developpement-de-leurs-competences-en-lecture-montre-une-etude-de-concordia.html?c=nouvelles/medias/communiques-de-presse
-https://www.concordia.ca/research/lifestyle-addiction/research/publications.html
-https://www.concordia.ca/hr/contact.html#payroll
-https://www.concordia.ca/offices/archives/honorary-degree-recipients/2006/06/michele-thibodeau-deguire.html
-https://www.concordia.ca/alumni-friends/news/podcasts.html