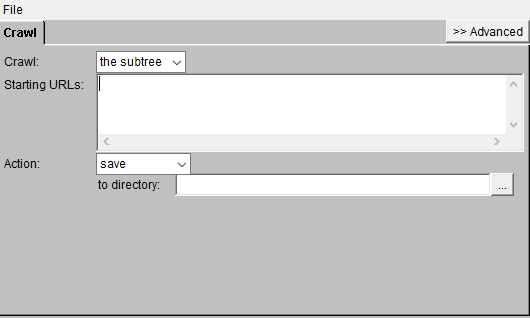
Jean-Loup Johnston

4006259

December 2nd 2019

Comp 479

Project 3 report

To begin the project, I did my web crawling using webSphinx with it’s GUI environment. 

WebSphinx takes a URL, crawls the subtree and saves all HTML files to a given directory on my computer. I saved the crawled pages in a folder called “crawlData”. WebSphinx has a built in function to obey the standard for robot exclusion. At first I couldn’t get any results when trying to crawl the Concordia page, then I tried using “http” instead of “https” in the URLand it resolved my issue.



Websphinx has no option to put an exact limit on documents being downloaded. I was also able to limit the documents for testing purposes using the depth option. This limits the downloads to only the documents that are one level below the given URL tree. During my testing, I worked with 40 documents. In my final web crawl I indexed 1106 documents.



Next, to extract the relevant text from the downloaded html files I used the Beautiful Soup python library. This library has an html parser than extracts the body and title text from the html. From the extracted text I then remove the digits, put everything to lower case, tokenize using the python nltk library, and remove the English stop words from the nltk library.

Once all my text is preprocessed, I created a dictionary and store it on the disk. While creating the dictionary I also store the document frequencies, term frequencies, and document lengths on the disk.

Testing

Information need: **Which departments have AI research?**

Query 1: “ai research department”

Top k = 10

The bm25 ranking for this query does not return the most relevant pages, the first results are an event page, a news page and lists of graduate and undergraduate programs. The first results does not give us much information, it just mentions research and departments vaguely. The second result has specific news articles about ai research but doesn’t mention which departments conduct the research. The next 2 pages just list all the programs Concordia has to offer, which does not help very much in finding the one that conducts ai research. If we extend the top k to 50, more relevant results appear further down in the list. There are a dozen pages about research in general but not ai. The result “ginacody\research\ai.html” is at position 90, which appears to be most relevant. For top k=100 most results are research news pages that don’t talk about ai.

The bm25 ranking is not much more useful for our information need with the top 10 results. The top 2 results are art & science pages, which are irrelevant. But, the 4th result does give us more information. It is the list of graduate programs in the ginacody directory. This result is closer to what we are looking for than simply a list of all the programs in the school. For the top 50 and top 100 there is a wide range of arts & sciences pages, news, pages and irrelevant research pages

In general this query was not very successful, the results were either irrelevant or gave lists of departments that were much too large. There were also many results for the arts & sciences department. This query was not very successful for our information need.

Query 2: “artificial intelligence research department”

We immediately see that the results are more relevant. More results are related to artificial intelligence, and we have specific departments that are related to artificial intelligence. The bm25 ranking seems better, it returns the pages for software engineering, computer engineering and math & stats, all departments that are related to artificial intelligence. The top result for tf-idf does not mention artificial intelligence on the page, and the top result for bm25 is one of the most relevant pages for ai at Concordia.

For the bm25 top 50, another relevant department is added (ginacody\info-systems-eng) at position 12, the rest strays from the query. For the top 100, there are just irrelevant department and research topics. For the top 50 in tf-idf, the relevant departments from b25 appear at positions 11,17,21,40. The next results are mostly other departments and research topics.

Query 3: “artificial intelligence research departments”

Making the query term “department” plural had a surprising effect on the results due to the tokens not being stemmed. In the top 10, there are less departments and more news pages for both bm25 and tf-idf. They are vague news pages that do not return much information. For both ranking methods, some departments related to artificial intelligence appear in the top 50, not much of use comes up after the to 50.

In summary, the best query for this information need was “artificial intelligence research department”. Spelling out “artificial intelligence” greatly improved the results. adding an s to “departments” was not useful, it probably got uses of the word that were too specific. By looking through the Concordia website, I found the most relevant result for the information need to be <https://www.concordia.ca/ginacody/research/ai.html>. 

This result lists the departments related to artificial intelligence.

Information need: **Which researchers are working on AI research?**

Query 1: “artificial intelligence researchers”

With this query we can see that both ranking methods are returning “/ginacody/research/ai” as the first result. Once again this seems like the most relevant results, it gives the following list of researchers working on artificial intelligence:



The other relevant result is <https://www.concordia.ca/research/news?page=4>. This page has 2 articles about artificial intelligence research with the researcher names. This result is at position 4 in bm25 ad 5 in tf-idf. The other results do not give much information related to our information need.

The word “researchers” does not seem to help in the query, the relevant pages are retrieved because of “artificial intelligence”.

Query 2: “artificial intelligence professor”

The top 2 results for bm25 are relevant for this query, then the 6th result also has some relevance by listing researchers. The top 2 results are also relevant in tf-idf but the 4th news page with relevant ai articles is not returned. This result comes in at position 12. There do not seem to be other relevant results for this information need.

Query 3: “artificial intelligence ai professor professors”

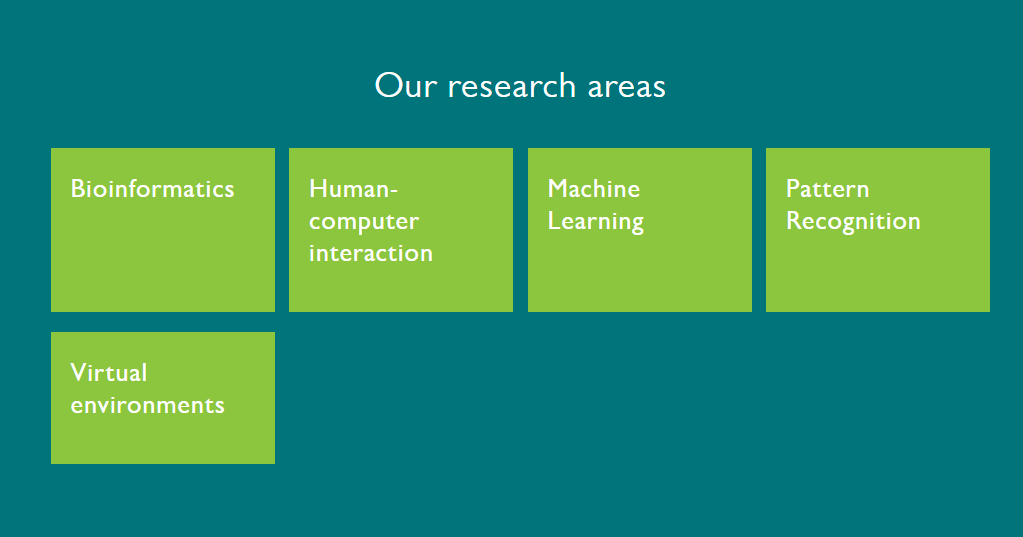
In this query I expanded the previous query by adding “ai” and “professors” to try and add more terms that could get relevant results. The results did not change very much for top k = 10, 50,100. Bm25 is still returning more relevant results than tf-idf.

In summary, there were only a few results that were relevant to this information need. Overall, the best query for this information need appears to be “artificial intelligence researchers”. The term “researchers” was not very useful, “professors” added more relevant results. For the 3rd query, adding more terms did not prove useful, possibly because there were few relevant results in the corpus.

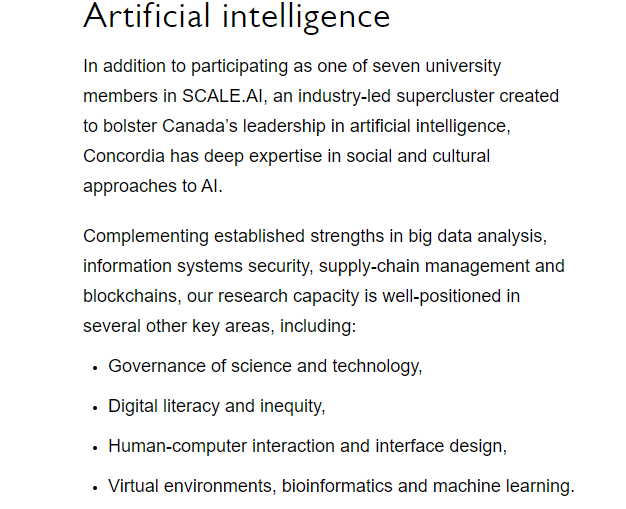
Information need: **What AI research is being conducted at Concordia?**

query 1: “artificial intelligence research areas”

Once again the most relevant page is <https://www.concordia.ca/ginacody/research/ai.html>:



This result comes as the first result for both ranking methods, and satisfies our information need. The second result for bm25 is also relevant to our information need:



The 5th and 6th results also have some relevance. They mention specific research articles and research centers that are relevant to artificial intelligence. The top 100 results bring in very similar results to other queries related to artificial intelligence. Using tf-idf the same relevant results that were at positions 1,2,5,6 are now at positions 1,2,7,8. Both ranking method bring similar relevant results in the top 10.

query 2: “ai research areas”

In my second query I used “ai” instead of artificial intelligence in hope of getting different Concordia news articles that are different than the ones retrieved in the first query. Both rankings returned the following page as their 3rd result, but no other relevant news in the top 10.



This articles about ai would not have been retrieved with “artificial intelligence”. Other than that bm25 retrieved irrelevant articles for positions 2 and 4. The top result for tf-idf was irrelevant, then it only has relevant results for 2 and 3. The research/ai page does not come up in the top 10 results for tf-idf. Bm25 returns several more news pages in the top 50, and from 50-100 it mostly returns alumni pages. The results are similar for tf-idf. Overall, this query was less informative than writing out “artificial intelligence”.

query 3: “artificial intelligence research fields”

Both bm25 and tf-idf are returning “ginacody/research/ai” , “research/cenparmi”, “ginacody/research/centers”,” research/spotlights” as relevant results in the top 10. This appears to be the most useful query for our information need. Once again, tf-idf is slightly less good for ranking because it is some completely irrelevant results. The top 50 and 100 returns some relevant news pages similarly to our first query.

AiTopics crawl

For the list of ai related terms, I used <https://relatedwords.org/relatedto/artificial%20intelligence>. From there I copied the top 500 terms and saved them in aiTerms.txt. I then crawled about 1300 pages from the aitopics root and kept the document frequencies of those 500 terms in the 1300 aitopics pages. Then in my query code, if the word being queried is in the list of 500 ai terms I use the new document frequency, otherwise I keep the Concordia document frequency. I had some issues when trying to crawl aitopics with webSphinx, at first it would crash on he first page. The I set it to crawl the web and not just pages in the aiTopics domain, and this worked. I did not realized at first that aiTopics was just a search engine that retrieved pages from other sites.

Teamwork

For this project, my teammate and I mostly worked on our own code individually. We consulted each other for small details when we weren’t sure what to do. A difference we have is that she has an index that is about twice as large as mine. When we were comparing queries she had son AND queries she had test, but I could not test them on my system. I do not have enough articles to get results for AND phrase queries, but she does. My partner also used stemming in her preprocessing and I didn’t. This lead my to have different queries for “department” and “departments” which only requires one query in her case.

For the contributions we both did our work separately, we only consulted each other a few times. For my project, all the contributions are mine.

My teammates queries:

-department ai

-research unit ai

-artificial intelligence departments

-researcher ai

-researcher artificial intelligence ai

-work ai research artificial intelligence

-ai conduct research Concordia

-research topic ai concordia

-current research artificial intelligence ai

While comparing each other’s queries, we found that we had some very different results when running the same query on each other’s machine. This is most likely because our indices a different. Her final index is twice as large as mine, so she gets many pages hat re never returned in my searches. We also used a different crawling pattern. We both used webSphynx but she used depth first search and I used breadth first search.