Assignment 3: M3 Results & Query List

Overall, from the first search, every search was fast enough but many of the queries lacked accuracy.

ACM:

Results were not accurate when we first searched. It got better but still the url didn't contain much information about 'ACM'. After adding the filter_url function that strips out the most common url in the domain, the search result improved significantly. (0.009 sec)

Cristina Lopes:

Results were not accurate when we first searched. Only found one url but it wasn't related with the name 'cristina lopes'.

After changing the codes, results were much better, the urls contained good enough information about 'cristina lopes', and the filter function filters out the main page of cristina lopes in ics.uci.edu. (0.003 sec)

Machine Learning:

Results were not accurate when we first searched. Most of the urls have the same score, so after they were sorted, the results were not ideal. That's when we implemented the filter_url functions. What it does is it will find the most common domain and url substring that appeared the most number of times. The results were quite good after adding the filter function. No more duplicate pages and results were very similar to the result searched on ics.uci.edu. (0.2 sec)

ICS46:

Results were good from the beginning and are still good. Returned relevant course links primarily for Prof. Thornton.

However, if you search for "ics 46" instead of "ics46" the performance will decrease. That is due to the fact that we separate the entire index into multiple smaller indexes, and while we search for a number, we need to read from the index file for the number again. That needs more time than just searching for "ics46" as one single word. The performance is still acceptable even though there's a drop.(0.001 sec)

Fall Quarter:

Previous search is completely random and showed no any information about Fall quarter

The urls mostly had course information instead of information about the "Fall Quarter". The reason for mismatching results to the query might result in the query itself since many websites have the phrase "Fall quarter" but the information in it is about something else. (0.02 sec)

Informatics:

Results were not as accurate. Because the term "informatics" will be stemmed to "informat" with the porter stemmer. After we added a white list to the stemmer, the result improved.

The performance of the search was not acceptable as well, because many websites contains the key word "informat" and that causes problem when you try to calculate tf-idf for every url. But it improved after we added 'informatics' in the stemmer white list. (0.283sec)

Graduation:

Initially, the search result for "graduation" is performing poorly, the result are somewhat irrelevant to graduation, but there are many url under the search result we want: "https://www.ics.uci.edu/grad/". And that's why we implemented the filter_url function, which pick out this url and rank it first among all the url we get.(Time 0.04 sec)

Artificial Intelligence:

Results were good with a majority coming from links for courses in AI and were almost identical to results for "AI". (0.02 sec)

ICS Majors:

The filter_url function filters out the wrong url, but the normal search result without filtering still produces a

good result.
Even though sometimes the filter_url fails to produce the correct result, our normal searching function will make up for it. In this case, even though the top url is not acceptable, the first three links still contain the desired result. (0.03 sec)

Richard Pattis:

Professor Pattis name appeared in many pages under his personal main page in ics.uci.edu. And this is when filter_url shows its strength. It gather all the urls collected and filter the “ics.uci.edu/~pattis/” out from all the webpages and rank it first.(0.03 sec)

Data Science:

This is where the filter function did not produce the desired result, but the rest of the links still makes up for that.
(0.06 sec)

Capstone Project:

The ideal result, main page for Capstone project can be found within the first two links, and the performance of search is good.(0.02 sec)

Course Listing:

The filter function filtered out one of the professor’s main pages. That is due to the fact that many of his pages contain the phrase “course listing”. But still the search engine was able to provide links to both undergraduate course listing page and graduate course listing in the first 5 links.

The rest of the links are irrelevant because they contain “course” and “listing” but not as a phrase. Another reason why there are many pages from Professor Lathrop is because the stemmed “listing” is the same as the stemmed “list”. List is a data structure and is frequently used in ICS courses, so many pages contains “course” and “listing”. That’s why the filtering function filtered them out. Even though this might create confusion for users, we can delete the filter function because this problem is derived from the fact that when we stemmed the text, the meaning of it might change. We are actually sacrificing accuracy for performance, and we do get good results (good performance) in the first few links. (0.04 sec)

Machine and Learning:

In the old search system, the results were not useful or relevant and some duplicate links, similar to “Machine Learning” previously. After changing the code and adding the filter function, the results and search times are good and results are identical to “Machine Learning" (0.2 sec)

Scholarship and Fellowship

This query produces frustrating results at the beginning when we only use term frequency instead of cosine scoring. As we moved along, the result improved, right now it produces results about scholarship information in the top 4 links. (0.0082 sec)

Scholarship

This query was used to test the performance change between searching for one word and two words. We found there is an increase in the time it takes for search engines to search for two words rather than one, but the time increase is linear, therefore acceptable. (0.0010 sec)

Internship

Another example which filtering function does its job. The top ranking web page is the career page for students in ICS schools. And we also set up a timer for the filter function, since it does its job in linear time, and the parameter passed in is already processed, the speed is incredibly fast. (0.0023 sec)

ICS 46 project 5

This query suffers a huge performance drop at the beginning of our search engine implementation. That is because the search engine loads the index for number terms again and again, which is an inefficient use of system resources. We adjusted our loading function to only load the index for numbers once, it greatly improved the performance and provided good results at the same time (project 5 spec page in the first 2 links). (0.214)

CS project courses

We expect to see project courses appear in the search results. The accuracy is high, it provides links to the project course of Professor Lathrop. But the performance was low. After tracing the root of the problem, we found that the search engine spends the most amount of time loading the index. When there are more than one word in the query, the search engine will load the partial index repeatedly, which creates a problem and wastes a lot of time. So we adjust the index structure to have more indexes with less size. It greatly improved the loading time and the performance.(0.138 sec)

Enrollment

This is a vague search query, but the search engine still produces results that include all kinds of information. From course enrollment info to undergraduate enrollment info. (0.00039 sec)

CS majors

While the user enters this query, they want to see a list of cs majors. The search engine provides them with undergraduate course listing and undergraduate policies. It is considered to be good results. And the performance is good. (0.03 sec)