Informatics 141
Winter 2021

Search Engine Report

Bryan Snyder, Lillian Won, Matthew Olitoquit

List of queries tested on our search engine:

GOOD QUERIES

1) cristina lopes
2) machine learning
3) ACM
4) master of software engineering
5) computational theory
6) computing technology with statistics
7) professor in computer science
8) 2017 important information
9) chicken
10) method and classes and other stuff

BAD QUERIES

11) master of computer science
12) ics computer science informatics statistics
13) donald bren school of information and computer sciences
14) to be or not to be
15) that is the question
16) the importance of computer science with information retrieval
17) computer science computational theory and graphs
18) a i the
19) a b c d e f g
20) 0 and 1

The queries that are bolded performed poorly during the search retrieval. These queries took more than 300ms to return a list of URLs. To make them perform better, we checked if the query contained any stop words. If more than 50% of the query contained stopwords, the query would not be modified. Otherwise, the stop words would be removed from the query. This reduced the speed for most of the queries.

Our query processing uses “document-at-a-time”. Initially, we would look through the entire document (~56,000) to match each document ids. To fix this, we only look through the document ids that were obtained by the query and put the document ids in a set to remove duplicates before sorting them. Both of these methods significantly reduced the query time (by around 50 milliseconds).

Most of the queries listed also had an issue with URL ranking and the number of URLs that would be returned. We first implemented a boolean retrieval query method,

but that limited our results too much and did not rank any of our URLs. For example, the best URL on the list would be located in the middle of the lists, instead of at the very top. We instead applied TF-IDF and added it to the “important texts” found on the HTML page. This would make the important texts’ weight much more.