GuUCLe – The UCL Search Engine

Emmet Cassidy, Jason Cheung, David Kelly & Nicholas Read

Aims

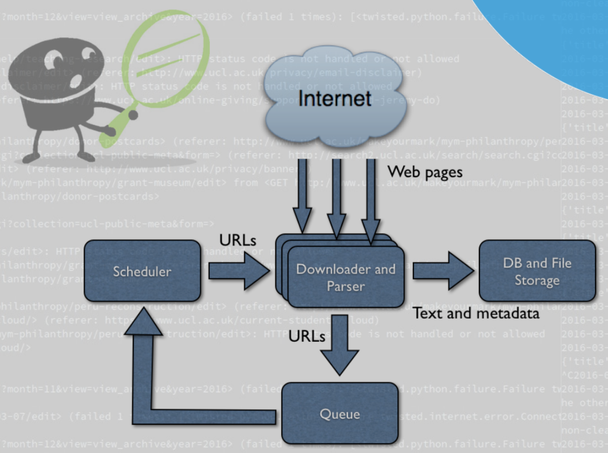
In this project we aim to implement a live search engine that is capable of indexing and searching the ucl.ac.uk website, including all sub-domains. Using open source information retrieval and parsing packages we will implement a PageRank algorithm and evaluate the efficacy of our solution with respect to the existing UCL search engine.

Query Methodology

We intend to implement the extended Boolean search paradigm, adjusting the ranking of each result according to the occurrence frequency of search terms. To improve upon this we may combine term frequency with PageRank in a novel normalized metric to augment ranking of more authoritative results.

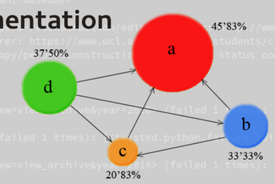
GuUCLe

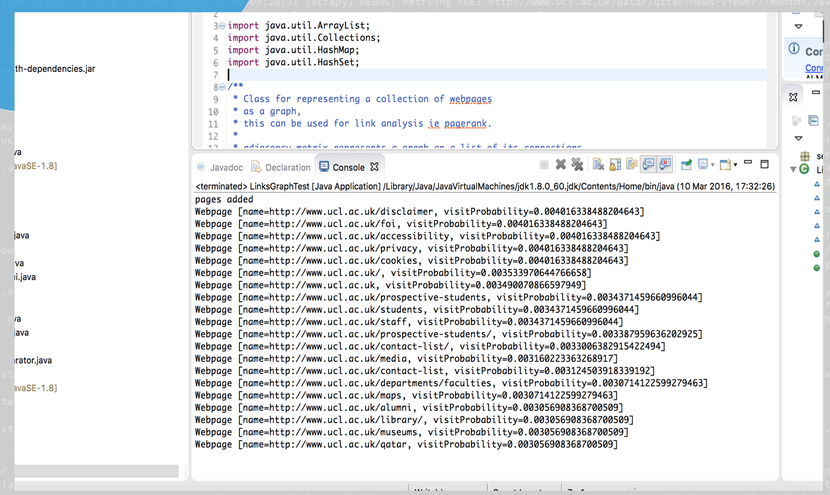
Custom-built webcrawler!



PageRank implementation!

We will write our own efficient implementation of the PageRank algorithm





Evaluation

Compare the results of three search engines... Google, UCL internal and guUCLe!

By implementing various different retrieval methods within our search engine, we can compare and contrast results, thereby allowing us to deduce which type of relevancy ranking system UCL is currently using.

Using our own judgement, we can determine the relevance of each of the top ten returned search results for various queries. We can then calculate the precision, recall and ndcg@k for each of the three search engines, and compare the results.