

Lucene Project II

Specialized Search Engine

- DBLife- Directory of researchers in the database community.
 - <http://dblife.cs.wisc.edu/>
- Index websites of researchers in the database field
- Accept and respond to queries with a ranked list of homepages.
- You can work in groups of 1-3.

Requirements

- Crawl websites of all the researchers, following links, in order to index content.
- Extract text from at least one content type other than text or html. (PDF, PowerPoint, Word, etc.)
- Link homepage to the content of the site.
- Determine how to test for similarity so that you can rank homepages.
- Create a simple UI for inputting queries and returning homepages

Design Deliverable

- Describe the design choices you made in creating your search engine.
- Describe how your system is implemented.
- Talk about some of the problem you ran into.
- Propose how you might evaluate your system.
 - You don't have to actually evaluate

Other Deliverables

- Send me a copy of all the code developed and libraries used.
- Show the results of the queries listed on the project handout.
- If you work in a group please make sure you list all that contributed so everyone can get credit!