The Project

DM79 Fall 2007

Project

Make a search engine.

- Goal: Index domain `.dk` .
- Language: Any appropriate (Java, C, C++, C#, Python, . . .).
- Cooperation and teamwork necessary.
- No detailed project description. Specification part of the task.
- Deadline (proposal):

Monday, January 7

Subtasks

1. Crawling (40-50% of work)

   Implement a crawler. Make web page describing crawler. Save all documents. Make URL/doc database. Make edge list of web graph.

2. Indexing (20-25% of work)

   From document collection, build inverted file list and lexicon.

3. PageRank (10-15% of work)

   Implement PageRank. Implement an efficient external sorting algorithm.

4. Querying (10-20% of work)

   Make search web interface. Implement search algorithms. Add further data structures to support advanced queries (pattern matching).

Tasks

- Make design choices.
- Define time schedule (milestones).
- Experiment.
- Implement production code.
- Coordinate with relevant other sub-groups:
  - Definition of interfaces.
  - Time schedule.
- Document the process (your deliberations, experiences, design choices, implementation, coordination with other groups)

Process should be based on formal meetings in group.

Documentation

- You should document each meeting (around one page per meeting, or whatever feels appropriate).
- As proof of progress, you should hand in minutes at the first lecture following the 20th of each month (September, October, November, December).
- Your report will have the collection of such meeting minutes as an appendix.
- The report should describe the final solution in a structured and clear fashion, and should describe the design choices made and the reasoning behind these.

Minutes

Minutes should include:

1. What has actually been done since last meeting.
2. Evaluation of this.
3. Deliberations of the meeting.
4. Decisions of the meeting.
5. Plan for actions until next meeting.