The Project

DM79 Fall 2004

Project

Make a search engine (or two).

- Goal: Index domain `.dk` .
- Groups for each part of search engine.
- Language: Any appropriate (Java, C, C++, Python, . . .).
- Cooperation and teamwork necessary.
- No detailed project description. Specification part of the task.
- Deadline (for final code):

Thursday, December 9

Groups

1. Crawling (5–6 persons)

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

2. Indexing (2 persons)

   From document collection, build inverted file list and lexicon.

3. PageRank (2 persons)

   Implement PageRank. Implement an efficient external sorting algorithm.

4. Querying (2–3 persons)

   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 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 (and sometimes with other groups).

Documentation

- You should document each meeting (around one page per meeting, or whatever feels appropriate).
- Your entire report will consist of the collection of such meeting minutes (plus an overview of the final solution).
- You should hand in minutes at the first lecture following the 20th of each month (September, October, November, December).
- Hand in minutes of all meetings held since last hand-in. Hand-ins should be in ps or pdf format.
- Last hand-in will be

  December 20.

  It should also include an overview of the final solution achieved.

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.