Web Intelligence
Mini Project 1 – part 2

After this mini-project you will not be needing Google anymore – you will instead build your own complete search engine, including what is needed for crawling, indexing, and ranking!

1. [Crawling]: Implement a simple crawler which will store some, but not a too big number of pages. About 1.000 pages will be sufficient – but you decide. Start with a seed of your own choice and decide if you want to target pages by filtering with respect to a particular type of content or location. **Store the html and its URL** (we will be parsing it next time). Be sure to implement *politeness*. Integrate near duplicate analysis module from last exercise into your crawler.

   For the exam: please also make sure you have learned and understood the architecture for your crawler; how did you prioritize the crawl; did you cut any corners?

   You might find some inspiration on the web (e.g., on the "Chilkat Software" page http://www.example-code.com/csharp/spider.asp, you will also find other examples just by googling, both advanced as well as basic crawler examples, for python or other programming languages). But remember, you must understand what is going on and not just copy code-stumps! **[2 weeks before last week]**

2. [Indexing]: Implement an inverted index for the (content on the) pages that you stored away from your crawl last week. Include as much preprocessing/normalization of terms that you find important and time allows.

   For the exam: What did you do and where did you cut corners. What are the implications?

   It would be a good idea to debug on the simple example that you constructed in this week exercise session! **[1 week before last week]**

3. [Ranking (content-based)]: Last time, you implemented the boolean inverted index. Expand this index to handle the tf-idf vector-space model and use this index in an implementation of a standard retrieval of top-k matches to a query.

   Next, experiment with at least one of the contender-pruning methods and evaluate results at different levels of pruning. **[last week]**

   For the exam: What did you do and where did you cut corners. What are the implications?

4. [Ranking (link-based) and wrapping it out]:

Last time you have implemented the content based ranking of the keyword based search on top of you crawled example. Update the ranker with the page rank or if you have time also with hits. In traditional version of page rank, the score is query independent. Update therefore your indexes so that in your search, you will be able to apply page rank together with your content based rank. [**this week**]

For the exam: What did you do and where did you cut corners. What are the implications? Have you experience any memory or efficiency problems? Why?