

Web Intelligence

Assignment 1

After this assignment 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. [Indexing]: Next week
3. [Ranking]: In two weeks