

CmpE 473 Internet Programming
Fall 2007
Assignment 4—Due: 26/12/2007

You will do this assignment in groups of four. One person in the group will be the project leader for each assignment in the course. The project leader should be different in each assignment.

Scenario: You will develop a Web search engine for searching Boğaziçi University's Web pages. As usual, the users of the Web search engine will enter keywords to be searched. In addition, the users will have the option to state whether they want the search results to be ranked using a connection-based ranker or a content-based ranker. Intuitively, the choice for ranker will affect the rank of the pages.

Technical Details: To do this, you will start with an existing search engine and extend it to do connection-based and content-based ranking. Nutch (<http://lucene.apache.org/nutch/>) is an open source software for searching the Web. Nutch comes with its own crawler, indexer and searcher. It can be served through a JSP container such as Tomcat. When you setup Nutch correctly and feed in `boun.edu.tr` domain as input, it can crawl the pages under this domain, index the pages, and give you an application to search these pages.

You will extend the ranker in the Nutch to capture the above ranker types. For content-based ranker, use the content of the Web pages, such as the frequency of terms, the font sizes, and so on to decide which words are more relevant to a given page. For the connection-based ranker, either use PageRank or design your own algorithm to compute the ranks.

Conduct the following steps and report their results.

1. Have your search engine rank the results for “Pinar Yolum”, “Ayse Soysal”, and “computer” first using the content-based ranker and next for the connection-based ranker.
2. Execute the same queries on the Search page of `www.boun.edu.tr`.
3. Consider the first 10 results. For each search, count the number of pages that appear both in your first 10 pages and also in the first 10 pages of the `www.boun.edu.tr` search engine.
4. Consider the first 30 results. For each search, count the number of pages that appear both in your first 30 pages and also in the first 30 pages of the `www.boun.edu.tr` search engine.
5. Discuss which ranker is performing better and try to explain why there are differences.

You should be able to demonstrate that (1) you can access the search engine and search keywords in the `boun.edu.tr` domain, (2) you can perform searches with a content-based and a connection-based ranker and (3) you perform the above experiments to compare your results to that of `www.boun.edu.tr`.

You will give a demo of your system and hand in a detailed report on CD. You should turn in your CD on the due date **before** class.

Watch for the notification of demo dates. **All** members of the group need to be present in the demo. Be prepared to show all aspects of your program, including the different scenarios listed above. You need to explain the design of your system, your algorithms for the rankers

and implementation of the algorithm, and other details that are important for your design and implementation.

The demos will take place in the university so make sure that your demo runs on the PC Lab in the university or that you can bring in the necessary hardware and software for the demo.