

CmpE 473 Internet Programming
Spring 2006
Assignment 4—Due: 23/05/2006

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 to search *academic* or *administrative* Web pages. For example, if the search query is *Ayşe Soysal* and the user has chosen to search academic Web pages, then pages related to Ayşe Soysal's teaching or research interests should appear (e.g., <http://www.math.boun.edu.tr/instructors/soysal/soysal.htm>). However, if the user searches the same query within administrative pages, then the search engine will show pages related to her position as a rector (e.g., <http://www.boun.edu.tr/government/rector.html>).

Technical Details: To do this, you will start with an existing search engine and extend it to do categorical search. 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 indexer in the Nutch to capture categorical information. Design your algorithm so that you can tell apart which pages are academic and which pages are administrative. For example, you can add markers to each page to denote whether they are academic or administrative. Or, you can associate the index terms with the categories.

Bonus: Add a ranker to rank the pages that are returned by the indexer, so that you can show the more relevant pages before others. You can implement PageRank or create your own algorithm for the ranker.

You should be able to demonstrate that (1) you can access the search engine and search keywords in the `boun.edu.tr` domain without selecting any category, (2) you can perform searches within academic and administrative pages based on user selection and (3) you rank the returned pages based on some importance criteria (bonus).

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 algorithm for managing categories 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.