Web Search, Télécom ParisTech

Putting it all together

Pierre Senellart (pierre.senellart@telecom-paristech.fr)

23 March 2012

The purpose of this lab session is to put together the crawler, inverted index construction, PageRank computation, and information extractor that you worked on during the first three labs. It mainly consists in the items that were highlighted in the sections "To go further" of each lab assignment. This is also the occasion for finishing and polishing your implementations.

You can either start with your own implementation or with the reference implementation available on the course website, depending on how far you went. If you start with the reference implementation, take the time to study how it actually works. If you start with your own implementation, please check that the results are correct:

- The conjunctive query "president france 2007" should return documents such as "Nicolas Sarkozy" and "Jacques Chirac" in good positions.
- The two pages with the highest PageRank in the Simple English Wikipedia should be "United States" and "France", in that order.

There is no fixed list of assignment for this lab session, but focus on connecting the systems produced in the first four labs: How to use PageRank to improve the results of the inverted index? How to build the inverted index from a crawled subset of the Web, and use PageRank on top of this? How to integrate the information retrieval system with the information extracted from the articles? You are also welcome to try new things and show initiative!