Search Engines

Exercise 1

Dustin Lange

19.04.2011

Exercise Contents

- Search engine frameworks
  - Apache Lucene / Nutch / Luke / Solr …
- Programming tasks
  - Evaluate stemmers
- Algorithmic tasks
  - Apply PageRank
- Theoretical tasks
  - Compare IR models
- Reading papers
  - Read Google‘s paper on BigTable

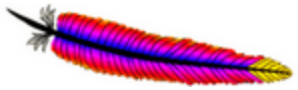
Your Wishes?

Formalities

- Exercise
  - 6 exercises (assignments)
  - Teams of 2 students
  - Solutions
    - Sometimes submissions
    - Some students present solutions
  - Regular attendance
  - Passing is necessary for exam admission
- Exam
  - Probably written
  - Grade determined by exam

Today:
Let‘s build a search engine!

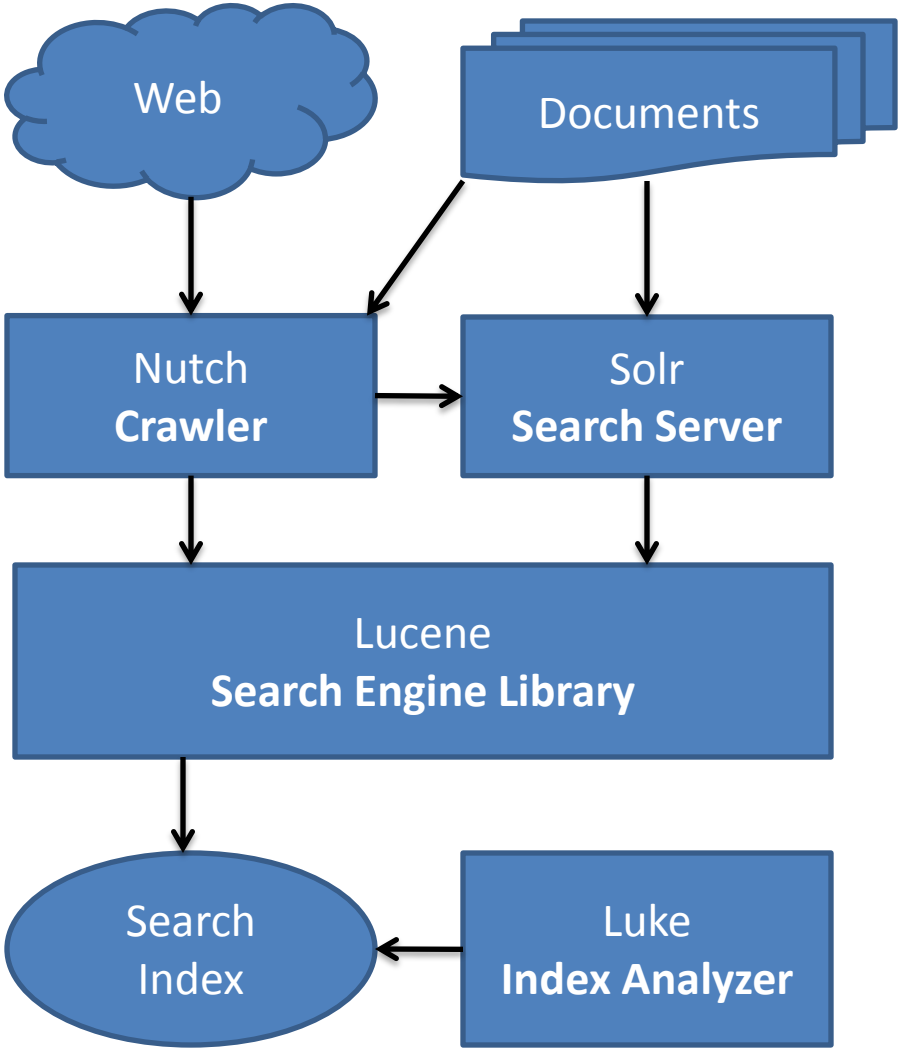
- Popular (http://wiki.apache.org/lucene-java/PoweredBy)
  - AOL, Apple, Disney, Eclipse, IBM, …
- Open source
- Scalable
- Many features (http://lucene.apache.org/java/docs/features.html)
- Many configuation options, much to try out in the exercise



The Lucene Family

- Apache Lucene
  - Apache Lucene(TM) is a high-performance, full-featured **text search engine library** written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.
- Apache Nutch
  - Nutch is open source web-search software. It builds on Lucene and Solr, adding web-specifics, such as a **crawler**, a link-graph database, parsers for HTML and other document formats, etc.
- Apache Solr
  - Solr is a standalone **enterprise search server** with a REST-like API. You put documents in it (called "indexing") via XML, JSON or binary over HTTP. You query it via HTTP GET and receive XML, JSON, or binary results.
- Luke
  - Luke is a handy **development and diagnostic tool**, which accesses already existing Lucene indexes and allows you to display and modify their content in several ways

The Lucene Family



Introduction to Nutch

- See PDF slides *NutchIntroduction.pdf*
  - Also available on SlideShare: http://www.slideshare.net/abial/nutch-webscale-search-engine-toolkit
- Demo

Task 1: Apache Nutch

- Download and install Apache Nutch 1.2
  - You will need: Java >= 1.4, Apache Tomcat >= 5.0 (for GUI, optional), Cygwin (for shell support on Windows)
  - Wiki: http://wiki.apache.org/nutch
  - Tutorial: http://wiki.apache.org/nutch/NutchTutorial
  - Tutorial for Windows: http://wiki.apache.org/nutch/GettingNutchRunningWithWindows Also available for your favorite OS
- Crawl the web pages of our research group
  - Everything that starts with http://www.hpi.uni-potsdam.de/naumann
  - Only HTML pages (no PDF)
  - User agent should start with “SearchEngineExercise”
- Optional: Connect the crawled index to Tomcat and access the created search engine in a web browser

Task 1: Questions

Nutch Architecture

1. How can the Nutch components (NutchIntroduction.pdf, slide 11) be matched to the components of a search engine presented in the lecture (SearchEngines_02_Architecture.pdf, slides 4-5)?

Crawling with Nutch

2. What are the ways to tell Nutch what to crawl and what not?
3. How long did the crawling take?
4. How many results are shown for these queries:
   - duplicate detection
   - search engines
   - Android

Compare with Google

5. Compare your results for “duplicate detection” with Google’s results (when restricted to our group pages)
   - How does the appropriate Google query look like?
   - Are there differences in the results (top ranked results/number of results)? Can you imagine why?

Task 2: Googlewhacking

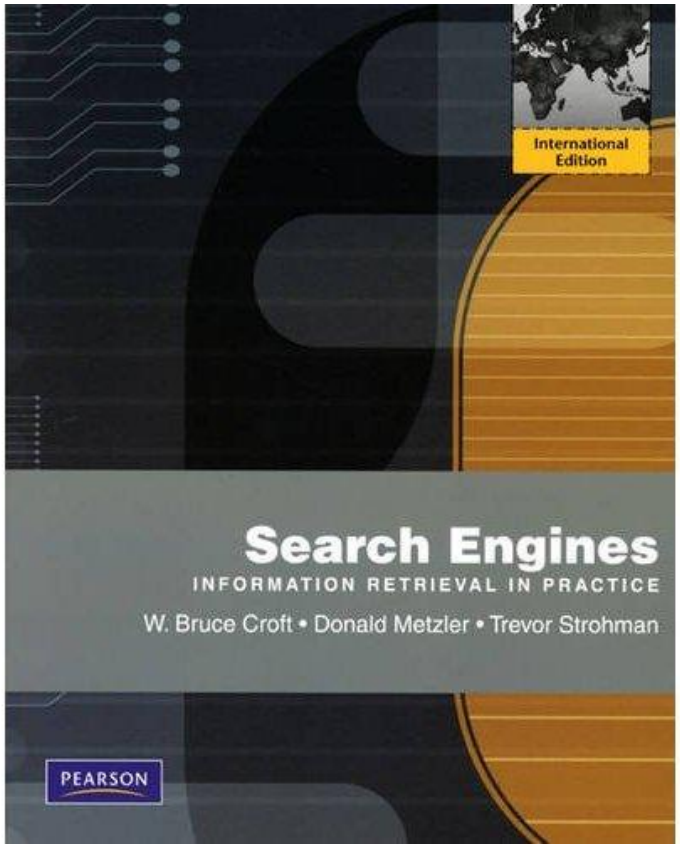
- Find the Goolge search query with the fewest results (>= 1)
- Given first term, add second term
  - hpi
  - potsdam
  - search
- Example:

What happens?

- Rules: Exactly two words, no quotation marks (+ is allowed), only words contained in some dictionary (no fantasy terms)

Task 2: Submission Details

- For each of the three given search terms, propose an additional term to complete the query
- Submit your terms **until today, 23:59,** using: http://goo.gl/9cU4P
- The team with the lowest overall number of results wins a prize (in case of tie: earliest submission wins)
  - Only first submission counts



Next Exercise

- On **28.04.2011**: Be prepared to present your solutions (for task 1)
  - English or German
  - Absent: Send me an e-mail in advance

Thanks for Listening

- Updates
  - See website
  - Mailing list: tbd
- Questions
  - Via e-mail: dustin (dot) lange (at) hpi (…)
  - Office: A-1.6