Lucene
Outline

1. Introduction
2. Live Demo
3. Conclusion
1. Introduction

- information retrieval library (Java)
- is an Apache project, originally created by Doug Cutting in 1999
  - Apache project since 2001
  - top level project since 2005
- very popular
  - ports for many languages exist
  - used by companies like Wikipedia, Twitter, ...
Index internals

- lucene splits the index data into separate files (each is a distinct index)
- segment file holds metadata about index directory
- index updates
  - commit() writes new index file
  - only one writer at any time
  - multiple readers allowed, never locked
- optimize() merges and defragments files
- big index for
  - boolean queries
  - vector space model
  - phrase search
2. Live Demo

...
3. Conclusion

- easy to use and flexible
  - querying and indexing is completely customizable
  - good api design
  - can be embedded into own application
  - incremental indexing
- highly scalable querying
- feature rich, but steep learning curve
  - many algorithms integrated (e.g. stemming)
  - capable of CJK characters
- actively maintained (v4 released in october)
- language independent via Solr
Bibliography

- Lucene in Action (outdated)
- Wikipedia
  - http://de.wikipedia.org/wiki/Lucene
- official home page
  - http://lucene.apache.org/core/
- heise article
The missing handout

- **indexing**
  - acquire content via apache nutch, apache tika, ..
  - build documents with (stored) (analyzed) fields
  - analyze documents using a self defined tokenizer chain (stemming, stop words, ngram, ..)

- **querying**
  - build query via QueryBuilder (translates lucene's query syntax into boolean/phrase/.. queries) or manually
  - sort results by score or user fields
  - documents and fields can be boosted (by index and query time)