Praxis der Forschung: Efficient Document Retrieval on General Sequences

Timo Bingmann, Simon Gog (gog@kit.edu), Peter Sanders
Motivation

- Massive amounts of textual data are publicly available, e.g. WWW data, including source code and DNA databases
- We consider the problem of searching
  - single word or phrase in the data
  - "bag-of-words" queries (like google does)
  - resulting documents are ranked
- Efficient search requires index data structures
- Most search engines (Google, Lucene) are inverted-index based
- We consider a theoretically more attractive solution
  - based on compressed suffix arrays
  - 2d and 3d range search structures
Tools

- SDSL (our own library)
  - C++ template library
  - of highly-optimized
  - succinct data structures

- Apache Lucene
  - industry standard inverted file based search engine
  - written in Java
  - baseline system
Prerequisites

Are you here?

- Algorithm lectures, text-indexing lecture, advanced data structures, ...
- Programming skills in C++11 and C++14, Java
- Unix tools (GDB, valgrind, vtune, perf)
- Low level programming
- Scripting languages (bash, python, R)
Literature


