DM823 String Algorithms (5 ECTS)

Rolf Fagerberg

Fall 2016

(all semester)

DM823 - Motivation

String data:

- “Once upon a time, there was a little mermaid. . . ”
- “CGTAATCCTTTTAG. . . ”
- “011000101010111101010. . . ”

DM823 - Motivation

String data:

- "Once upon a time, there was a little mermaid..."
- "CGTAATCCTTTTAG..."
- "0110001010101111010 10..."

are both common and fundamental:

- Text files
- WWW
- Database fields
- DNA sequences
- Integers: 934523324423
- Multidimentional data: (255,255,23), 213.43.43.134

In a sense, strings encompasses all known data types.

DM823 - Contents

We would like to:

- Compare strings to each other (string distances).
- Find substrings inside other strings (pattern matching), possibly with wildcards/regular expressions.
- Find a string inside a set of strings (string dictionaries).
- Sort a set of strings (string sorting).
- Compress strings (file compression).

DM823 - Contents

We would like to:

- Compare strings to each other (string distances).
- Find substrings inside other strings (pattern matching), possibly with wildcards/regular expressions.
- Find a string inside a set of strings (string dictionaries).
- Sort a set of strings (string sorting).
- Compress strings (file compression).

Subjects will be chosen based on

- Fundamentality of problem.
- Usefulness of algorithm.
- Algorithmic beauty.

DM823 - Contents

Detailed contents:

- Pattern matching algorithms: Knuth-Morris-Pratt (KMP), Boyer-Moore (BM), Galil variant (GBM)
- Searching in a sorted set of strings.
- Suffix arrays, LCP arrays, linear time algorithms for their construction.
- Tries, ternary search trees.
- Suffix trees, linear time algorithm for their construction.
- String sorting algorithms.
- The Burrows-Wheeler transform.
- String (file) Compressions algorithms: bzip2, LZ78, LZ77.
- String distance measures based on dynamic programming (LCS, edit distance, variants thereof) and Hirschbergs linear space algorithm for this.

DM823 - Formalities

Prerequisites: The contents of DM507 Algorithms and Data Structures should be known.

Materials:
Handouts (notes, articles).

Evaluation: Oral exam (7-point scale). Mandatory project (must be passed in order to attend the oral exam). Can be practical (implementation) or theoretical (exposition of research paper).