Text Indexing

Lecture 00: Course Overview

Florian Kurpicz
Organizational Matters

Lectures

- Monday 14:00–15:30 (50.34, 236)
- lecture only

Project (mandatory)

topics will be handed out 08.11.2021
coding project and small presentation
20 % of the final grade

Oral Exam

20 minutes
80 % of the final grade
pizza marks content not relevant for exam

Office Hours (Room 210)

Monday 15:30–16:00 (lecture period)
by appointment (otherwise)
Organizational Matters

**Lectures**
- Monday 14:00–15:30 (50.34, 236)
- lecture only

**Project (mandatory)**
- topics will be handed out 08.11.2021
- coding project and small presentation
- 20% of the final grade
## Organizational Matters

### Lectures
- Monday 14:00–15:30 (50.34, 236)
- lecture only

### Project (mandatory)
- topics will be handed out 08.11.2021
- coding project and small presentation
- 20 % of the final grade

### Oral Exam
- 20 minutes
- 80 % of the final grade
- pizza marks content not relevant for exam
# Organizational Matters

## Lectures
- Monday 14:00–15:30 (50.34, 236)
- Lecture only

## Project (mandatory)
- Topics will be handed out 08.11.2021
- Coding project and small presentation
- 20% of the final grade

## Oral Exam
- 20 minutes
- 80% of the final grade
- Pizza marks content not relevant for exam

## Office Hours (Room 210)
- Monday 15:30–16:00 (lecture period)
- By appointment (otherwise)
## Materials

### Slides
- Published shortly before the lecture  
  ([https://algo2.iti.kit.edu/4326.php](https://algo2.iti.kit.edu/4326.php))

### Videos
- Will be published (with ≥ 1 week delay)

---

**Additional Material**

- **References to literature** included
  - **Books**

---

**3/9** 2022-10-23 Florian Kurpicz | Text Indexing | 00 Course Overview

Institute for Theoretical Informatics, Algorithm Engineering
## Materials

### Slides
- published shortly before the lecture
  ([https://algo2.iti.kit.edu/4326.php](https://algo2.iti.kit.edu/4326.php))

### Videos
- will be published (with $\geq 1$ week delay)

### Additional Material
- references to literature included
- books
- most likely no script
Content

**Fundamentals**
- tries
- suffix tree
- suffix array
- longest common prefix array
- Burrows-Wheeler transform (BWT)
- wavelet tree (+ bit vector rank/select)
- FM-index

**Compressed Indices**
- compressing the BWT and wavelet trees
- Lempel-Ziv 77/78 compression
- LZ compression vs. BWT compression
- compressed suffix trees and suffix arrays
- r-index

**Additional Topics**
- parallel construction
- different query types
From the Suffix Tree to the $r$-Index
From the Suffix Tree to the $r$-Index

- **Suffix Tree**: 1973
- **Suffix Array**: 1993
- **LCP Array**: 1993
- **BWT**: 1994
- **Wavelet Tree**: 2000
- **FM-Index**: 2000
- **r-Index**: 2018

**Memory Requirements**

- Bit-Vektoren mit Rank/Select-Anfragen
- (Patricia-)Tries
- Bit-Vektoren mit Rank/Select-Anfragen
- Succincte Datenstrukturen

**Compression**

- abccaaca
- 01110001
- 01001
- 101
- 1
- 10
- 11
- 0

- $a$: 0
- $b$: 4
- $c$: 5

**LCE-Anfragen**
From the Suffix Tree to the $r$-Index

- **Suffix Tree** (1973)
- **Suffix Array** (1993)
- **LCP Array** (1993)
- **BWT** (1994)
- **Wavelet Tree** (2000)
- **FM-Index** (2000)
- **r-Index** (2018)

**Memory Requirements**

- [ ] a
- [ ] b
- [ ] c
- [ ] a
- [ ] b
- [ ] c
- [ ] a
- [ ] b
- [ ] c

**Compression**

- abccaaca
- 01110001
- 01001
- 101
- 10
- 11
- 0
- a: 0
- b: 4
- c: 5

**Strings**

- babab
- aabac
- babab
- babab
- aabac
From the Suffix Tree to the $r$-Index

- **Suffix Tree**: 1973
- **Suffix Array**: 1993
- **LCP Array**: 1993
- **BWT**: 1994
- **Wavelet Tree**: 2000
- **FM-Index**: 2000
- **r-Index**: 2018

Memory Requirements

---

5/9 2022-10-23 Florian Kurpicz | Text Indexing | 00 Course Overview Institute for Theoretical Informatics, Algorithm Engineering
Motivation for Text Indices

- collection of text
- scanning not feasible
Motivation for Text Indices

- collection of text
- scanning not feasible
- inverted index (word based)
Motivation for Text Indices

- collection of text
- scanning not feasible
- inverted index (word based)
Motivation for Text Indices

- collection of text
- scanning not feasible
- inverted index (word based)
Motivation for Text Indices

- collection of text
- scanning not feasible
- inverted index (word based)
- phrase search
Motivation for Text Indices

- collection of text
- scanning not feasible
- inverted index (word based)
- phrase search
- counting queries
Motivation for Text Indices

- collection of text
- scanning not feasible
- inverted index (word based)
- phrase search
- counting queries
- what if there are no “words”
Why Texts?

Text is Everywhere

- Text-based Information
  - Wikipedia
  - dblp
  - books
  - news articles
  - code

- Very Important in Bioinformatics
  - DNA
  - proteins

---

Growth of DNA Sequencing

[Ste+15]
Bibliography

