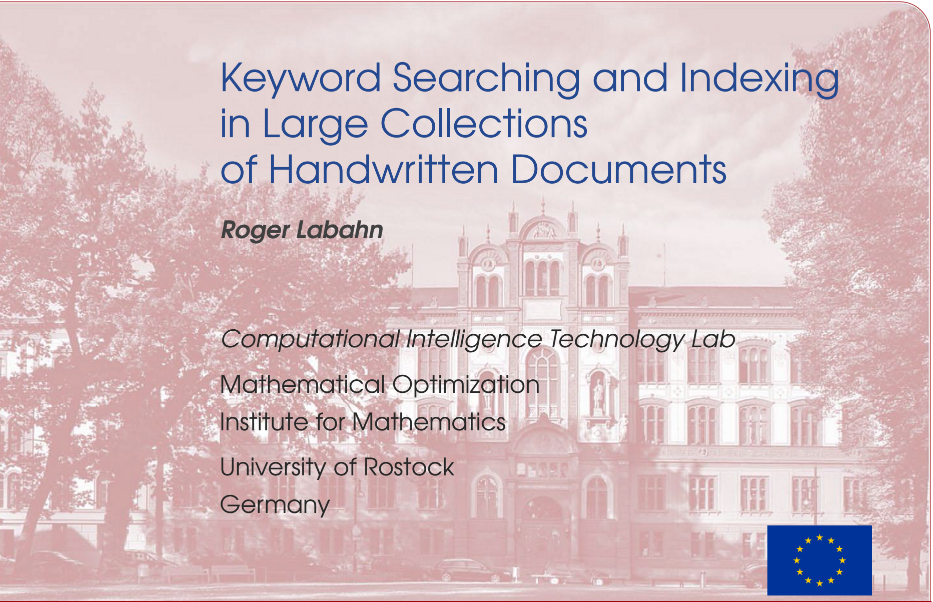
Universität Rostock

Traditio et Innovatio

CITlab
Computational Intelligence Technology

READ



Keyword Searching and Indexing in Large Collections of Handwritten Documents

Roger Labahn

Computational Intelligence Technology Lab

Mathematical Optimization
Institute for Mathematics

University of Rostock
Germany

Prologue

Foundation

Application

Epilogue

Motivation

Question

Do we really need a good transcription for searching and investigating?

Ambitions

- understand different concept & its applications
- learn technological terminology
- know about configuration & behaviour, features & bugs, . . .

Note !

- tools for continuous work – NOT just engines for execution
- adapt for specific challenges – understand & interpret outcomes

Prologue

Foundation

Application

Epilogue

Workflow – Classic

Text Recognition Process

- Recognition Engine ⇝ *character confidence scores* per position
- Neural Network / HMM outputs: estimate character probabilities

Raw Reading Result

- method: choose most likely character per position
- *free* reading: without considering document context
  e.g.: language, time, writer, …

Post – OCR Correction

- find & correct errors: use external sources from document context
  e.g.: language models, dictionaries, transcripts, …
- ⇝ strings: text transcription

Workflow – Recent

Text Recognition Process

- Recognition Engine ⇝ *character confidence scores* per position
- Neural Network / HMM outputs: estimate character probabilities

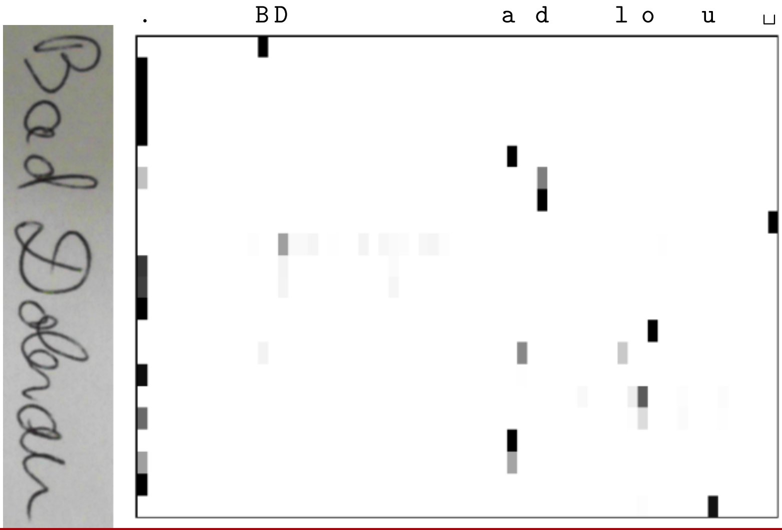
Confidence Matrix – ConfMat

- idea: evaluate entire recognition information
- application: Store this text recognition result!

Decoding

- query strings: use external sources from document context
- find optimal match / representation: query ⟺ ConfMat
- ⇝ ranked alternatives

Confidence Matrix



Measuring Similarity: String vs. String

Definition (Levenshtein Distance)

$$\mathtt{dist}(\text{string1}, \text{string2}) := \mathtt{count}(\text{insertions}, \text{deletions}, \text{substitutions})$$

Algorithm Dynamic Programming

- extremely efficient – very fast
- finds optimal (shortest / cheapest) path through weight (distance / cost) matrix
- weights: distance $\leftrightsquigarrow$ cost
- counting $\rightsquigarrow$ adding costs (weights)

Example: $\text{dist}(\text{WIEN}, \text{WEIN}) = 2$

insertions & deletions

		W	E	I	N
	0	1	2	3	4
W	1	**0**	**1**	2	3
I	2	1	2	**1**	2
E	3	2	1	**2**	3
N	4	3	2	3	**2**

W		I	E	N
W	E	I		N

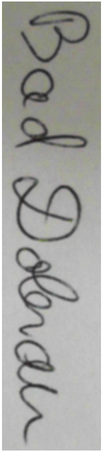
with substitutions

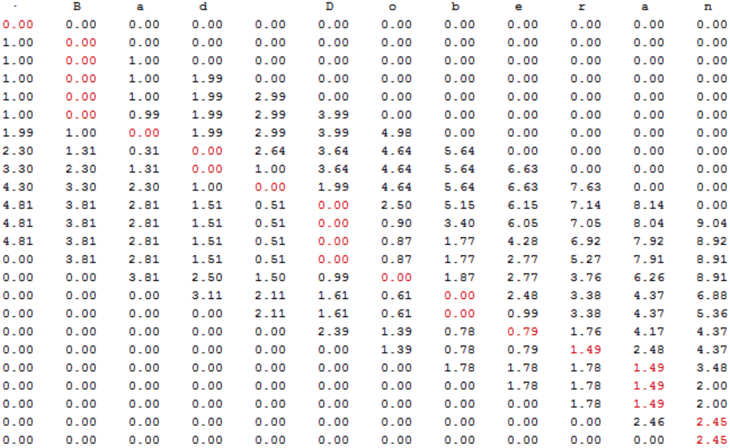
		W	E	I	N
	0	1	2	3	4
W	1	**0**	1	2	3
I	2	1	**1**	1	2
E	3	2	1	**2**	2
N	4	3	2	2	**2**

W	I	E	N
W	E	I	N

▸ Demo

Measuring Similarity: String vs. ConfMat





-	B	a	d		D	o	b	e	r	a	n
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	1.00	1.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	1.00	1.99	2.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.99	1.99	2.99	3.99	0.00	0.00	0.00	0.00	0.00	0.00
1.99	1.00	0.00	1.99	2.99	3.99	4.98	0.00	0.00	0.00	0.00	0.00
2.30	1.31	0.31	0.00	2.64	3.64	4.64	5.64	0.00	0.00	0.00	0.00
3.30	2.30	1.31	0.00	1.00	3.64	4.64	5.64	6.63	0.00	0.00	0.00
4.30	3.30	2.30	1.00	0.00	1.99	4.64	5.64	6.63	7.63	0.00	0.00
4.81	3.81	2.81	1.51	0.51	0.00	2.50	5.15	6.15	7.14	8.14	0.00
4.81	3.81	2.81	1.51	0.51	0.00	0.90	3.40	6.05	7.05	8.04	9.04
4.81	3.81	2.81	1.51	0.51	0.00	0.87	1.77	4.28	6.92	7.92	8.92
0.00	3.81	2.81	1.51	0.51	0.00	0.87	1.77	2.77	5.27	7.91	8.91
0.00	0.00	3.81	2.50	1.50	0.99	0.00	1.87	2.77	3.76	6.26	8.91
0.00	0.00	0.00	3.11	2.11	1.61	0.61	0.00	2.48	3.38	4.37	6.88
0.00	0.00	0.00	0.00	2.11	1.61	0.61	0.00	0.99	3.38	4.37	5.36
0.00	0.00	0.00	0.00	0.00	2.39	1.39	0.78	0.79	1.76	4.17	4.37
0.00	0.00	0.00	0.00	0.00	0.00	1.39	0.78	0.79	1.49	2.48	4.37
0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.78	1.78	1.78	1.49	3.48
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.78	1.78	1.49	2.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.78	1.49	2.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.46	2.45
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.45

Probability / Confidence $\leftrightsquigarrow$ Distance / Cost

Confidence

1.00

⋮

0.00

$\exp(-\text{dist})$

$\Longleftarrow$

$\Longrightarrow$

$-\ln(\text{prob})$

Distance

0.0

⋮

⋮

⋮

∞

Asymptotic Problem

- confidence mapping / scaling of arbitrarily large distances
- What distances practically correspond to probability Zero?

Prologue

Foundation

Application

Epilogue

Setup

Text Recognition Engine

Text ⇝ ConfMats per line

Decoding Engine

Query ⇝ Alternatives ranked by confidence / distance

Measurement Issues

- Confidences are NO PROBABILITIES!
- Confidences / Distances have NO ABSOLUTE meaning!
- Measurements are essentially INCOMPARABLE across different queries!
- Thresholds require MANUAL CONFIGURATION & TUNING!

Keyword Search

String Search

INPUT Query String

Decoding Everywhere: distance to query

OUTPUT rank “reasonable” hits – skip “irrelevant” answers

Big Data Issues

- Response time: inacceptable ! ?
- Preprocessing: index ⇝ database

▸ UPVLC

Transcription

Reading Text

INPUT Language Model – Dictionary incl. word frequencies

Decoding distances: all dictionary entries everywhere

OUTPUT “reasonable” text alternatives: close to ConfMat & Language Model

Issue

- both poor Text Recognition AND Language Model

Investigating

Challenges

- demanding Language Models: fuzzy / dynamic / incomplete
- complex queries:
  specific combinations – character classes – restricted vocabulary

Regular Expression Decoding

- *Regular Expression*: Computer Science & Programming
- unsupported features: named classes, . . .
- additional feature: dictionary classes

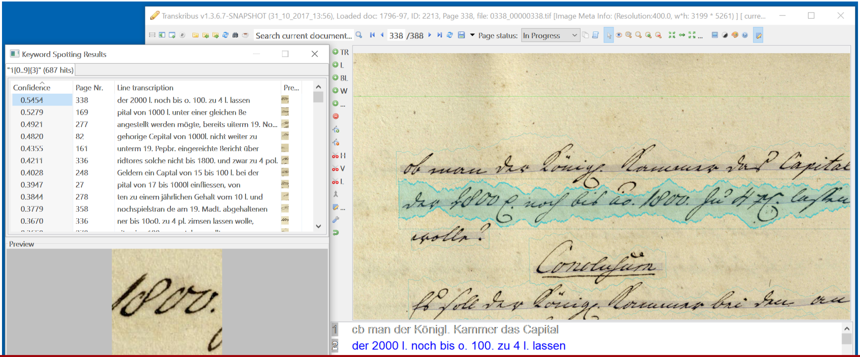
Investigation with Regular Expressions

Regular Expression Example

four-digit year: `1YYY`

`1[0-9][0-9][0-9]`

`1[0-9]{3}`



Transkribus v1.3.6.7-SNAPSHOT (31_10_2017_13:56), Loaded doc: 1796-97, ID: 2213, Page 338, file: 0338_00000338.tif [Image Meta Info: (Resolution:400.0, w*h: 3199 * 5261)] [curre...

Search current document... 338 /388 Page status: In Progress

Keyword Spotting Results

"1[0..9]{3}" (687 hits)

Confidence	Page Nr.	Line transcription	Pre...
0.5454	338	der 2000 l. noch bis o. 100. zu 4 l. lassen	
0.5279	169	pital von 1000 l. unter einer gleichen Be	
0.4921	277	angestellt werden mögte, bereits uiterm 19. No...	
0.4820	82	gehorige Cepital von 1000l. nicht weiter zu	
0.4355	161	unterm 19. Pepbr. eingereichte Bericht über	
0.4211	336	ridtores solche nicht bis 1800. und zwar zu 4 pol.	
0.4028	248	Geldern ein Captal von 15 bis 100 l. bei der	
0.3947	27	pital von 17 bis 1000l einfliessen, von	
0.3844	278	ten zu einem jährlichen Gehalt vom 10 l. und	
0.3779	358	nochspielstran de am 19. Madt. abgehaltenen	
0.3670	336	ner bis 10o0. zu 4 pl. zimsen lassen wolle,	

Preview

1 cb man der Königl. Kammer das Capital

2 der 2000 l. noch bis o. 100. zu 4 l. lassen

Regular Expression Syntax

Regular Expression Example

complete date: `TT.MM.YYYY`

```
.*(?<KW>[0-3][0-9]\.[0-1][0-9]\.[1-2][0-9]{3}).*
```

Note

- match: against the entire line
- score: designated `KW`-group

```
.*(?<KW><query>).*
```

Performance

Transkribus KWS Expert Mode

- Regular Expression Decoding: `.*(?<KW>[0-9]{4}).*`
- KWS demo collection: 388 pages from StAZh 1796-97

» ≈ 40 s

Note

- strongly depends upon hardware
- coming up next:
  massive parallelization on CPU & GPU

Performance

Office laptop single core

- Searching 1 keyword in 10.500 lines (433 BENTHAM pages):

» 2 – 3 s average

- Reading 1 page against 11.650 words dictionary:

» 8 – 9 s average

Prologue

Foundation

Application

Epilogue

Questions

Your wishes – expectations – requirements . . .

- . . . realistic query type / elaborateness / complexity ??
- . . . realistic data corpus size ??
- . . . realistic query response time ??

THANKS …

CITlab Group

CITlab
Computational Intelligence Technology

MoU Partner SME

PLANET artificial intelligence GmbH

EU Horizon2020 Grant

READ Recognition and Enrichment of Archival Documents

… for your kind attention!