BIBLIOTHECA ARABICA

ML-basierte Texterkennung arabographischer Handschriftenkataloge

Herausforderungen und Best Practices

Konferenz „KI-Methoden im Akademienprogramm“, Hamburg 2024

Daniel Kinitz (SAW Leipzig)

Sächsische Akademie der Wissenschaften zu Leipzig

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Ablauf

- Bibliotheca Arabica: Kontext
- Herausforderungen

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

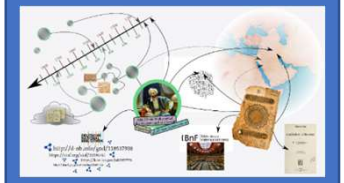
3 Arbeitsbereiche

Makroperspektive

u.a.
Textgenres,
Textpraktiken

Bsp.
Hadith
Commentaries,
Randkommen-
tare

Digitale BA



Mikroperspektive

u.a.
MS-Vermerke,
Bibliotheken

Bsp.
Shirwani

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

3 Arbeitsbereiche

Digitale BA

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

https://khizana.saw-leipzig.de

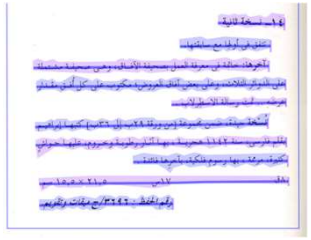
Graph Database (Knowledge Graph)

- Data from > 100 manuscript catalogues (print & online)
- Manuscript notes on owners, readers, etc.
- …

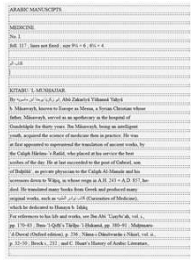
Automatische Datenverarbeitung vs. manuelle Eingabe

BIBLIOTHECA ARABICA

Scan (catalogue, etc.)

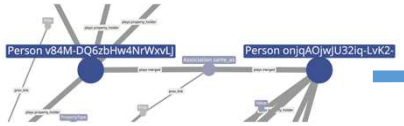


full text

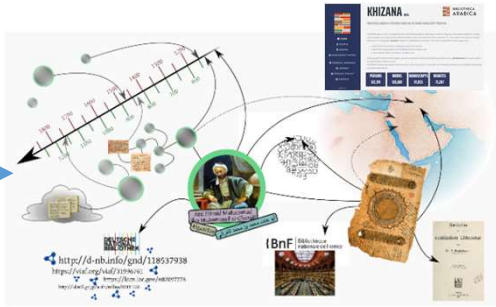


semantic parsing

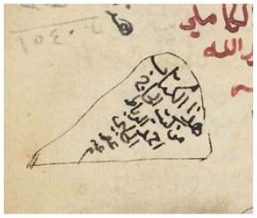
entity resolution (semi-automatic)



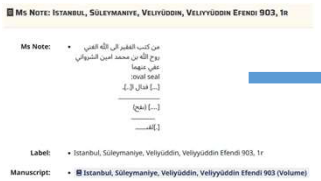
Internal / external links



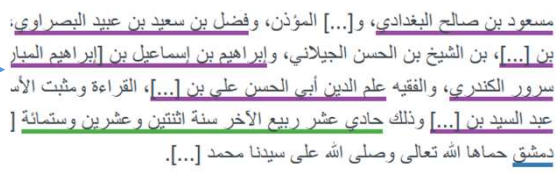
Manuscript (Ms Notes)



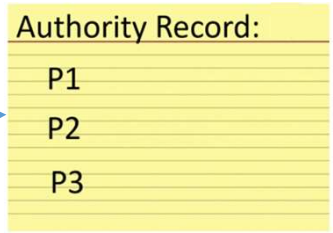
full text



annotation



entity resolution (manual)

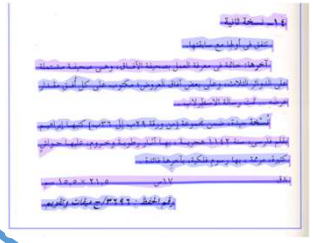


BIBLIOTHECA ARABICA

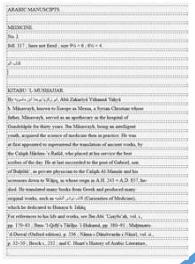
Automatische Datenverarbeitung vs. manuelle Eingabe

BIBLIOTHECA ARABICA

Scan (catalogue, etc.)

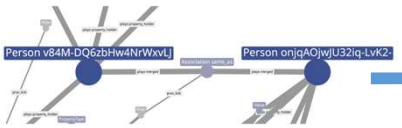


full text

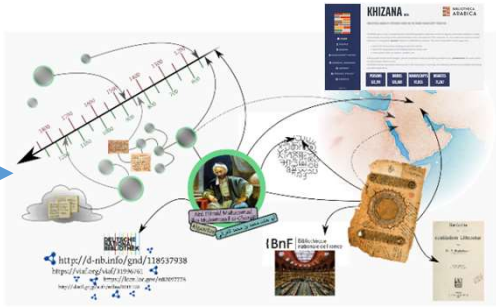


semantic parsing

entity resolution (semi-automatic)



Internal / external links



OCR: WORKFLOW

eScriptorium

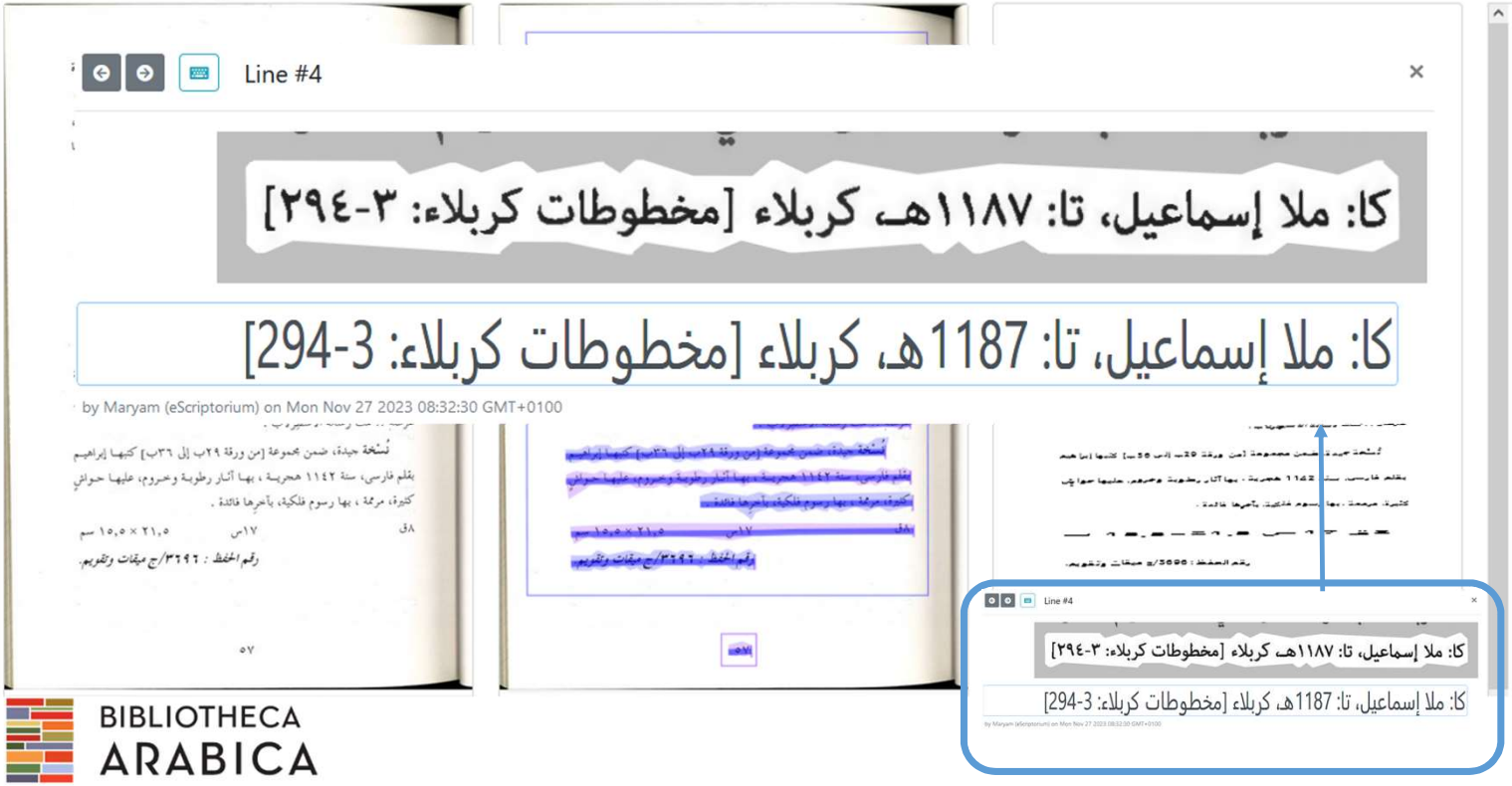
kraken

image

segmentation

training/correction

full text



1 آخرها: إن كان الإرتفاع شرقياً، وإلا فهو الباقى للغروب، مع زيادة
2 نصف التعديل فى الشمال وإسقاطه فى الجيوب، والله تعالى أعلم بحقيقة الحال.
3 نُسْخة جيدة، كُتِبَت بقلم معتاد، فى القرن الثالث عشر الهجرى تقديراً،
4 بها آثار رطوبة وخروم، مرممة، عليها حواشٍ ، بأولها فوائد فلكية وبآخرها
5 نقول باللغة الفارسية .
6 10ق 19س 21×15 سم
7 موضوعها : فلك .
8 رقم الحفظ : 3825/ج ميقات وتقويم.
9 14- نسخة ثانية
10 تتفق فى أولها مع سابقتها.
11 آخرها: خاتمة فى معرفة العمل بصحيفة الآفاق، وهى صحيفة مشتملة
12 على الدوائر الثلاث، وعلى بعض آفاق العروض؛ مكتوب على كل أُفق مقدار
13 عرضه . . تمت رسالة الاسطرلاب .
14 نُسْخة جيدة، ضمن مجموعة [من ورقة 29ب

BIBLIOTHECA ARABICA

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Historische Ausgangslage: arabographisches OCR

- Gedrucktes Arabisch, Persisch, Urdu usw. (RTL, Ligaturen) nicht mit herkömmlichen OCR-Anwendungen gut erkennbar (hoher Aufwand, relative schlechte Ergebnisse)

- Ab Mitte der 2010er Jahre erste frei verfügbare ML-basierte Anwendungen für OCR (kraken, Ben Kiessling), aber: zunächst ungeeignete/schlechte Transkriptionsinterfaces für RTL-Sprachen; → eScriptorium Repo seit 6 Jahren auf gitlab.com

→ Character Accuracy: > 0.99 möglich, word accuracy: > 0.97 möglich

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Herausforderungen & Best Practices

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Herausforderung: Kinderkrankheiten App

Rechte Spalten: Reihenfolge Zeilen vertauscht

Best Practices:

- App selbst fixen (pull request) → teuer
- auf Update warten / Arbeitspakete verschieben → umständlich
- eigenes Script → pragmatisch
- Workarounds, inklus. manuelle Arbeit SHKs

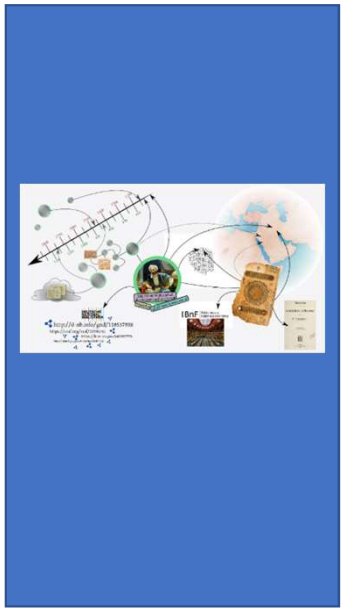
maBH5Znovmi6vInfIP.webp

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Herausforderung: WissZeitVG

Digitale BA



Team

- Informatiker/RSD
- Data Engineer / Digital Humanist
- ~~IT-Admin~~
- SHKs/WHKs

Problem: IT-Admin als wissenschaftlich-technischer Mitarbeiter → nicht mehr im WissZeitVG → „drohende“ Entfristung → Weggang

Folge: Verlagerung Arbeit auf mich, Informatiker, DH-Referent SAW
+ keine enge Zusammenarbeit mehr mit kraken/eScriptorium

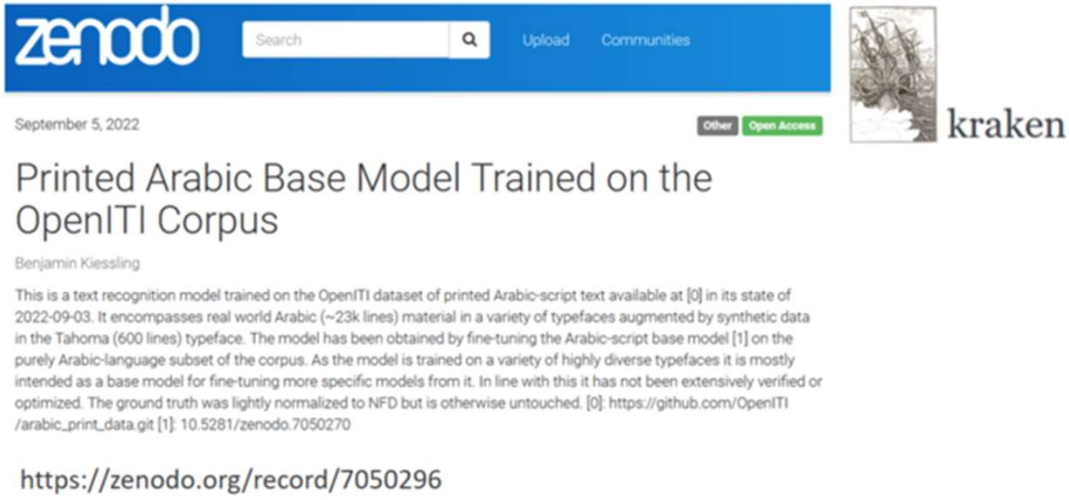
Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Herausforderung: Trainingsdaten generieren → viel Arbeit

mit wenig Manpower …

- bis vor Kurzem Fonts/Typen einzeln trainieren, von Grund auf neu
- Sep. 2022: erstes Gesamtmodel für gedrucktes Arabisch → Finetuning



Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Herausforderung: Korrekturlesen Hunderter Seiten

mit wenig Manpower …

→ 2 Modelle trainieren und diffs vergleichen

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Herausforderung: gemischte Schriften

→ gemischte Schriften (arabisch/persisch): Homographen und Verwechslungen pers. vs. arab. Ziffern

Bsp.: Verbundkatalog irak. Handschriften: Bd. 1-17 arab. Zeichensatz, Bd. 18 arab.+pers.

lat: 0123456789

ar: ٠١٢٣٤٥٦٧٨٩

fa: ۰۱۲۳۴۵۶۷۸۹

→ pers.-arab. Homographen bzw. Verwechslungsmöglichkeiten
arabisches ʿAyn (ع) → erkannt als pers. 6 (۶) → neues Modell
Herausforderung: den genauen Zeichensatz kennen (großen Textkorpus)

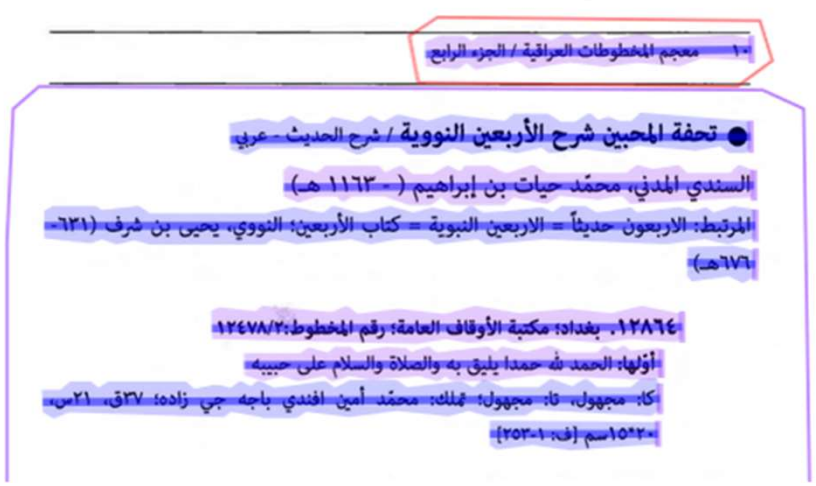
+ pers. Bearbeiter: arabische 1 (١ U+0661) vs. Persische 1 (۱ U+06F1) → Modell trainiert verschiedene, Zeichen, die m.E. nur aus politischen Gründen versch. code points erhalten haben

Sächsische Akademie der Wissenschaften zu Leipzig

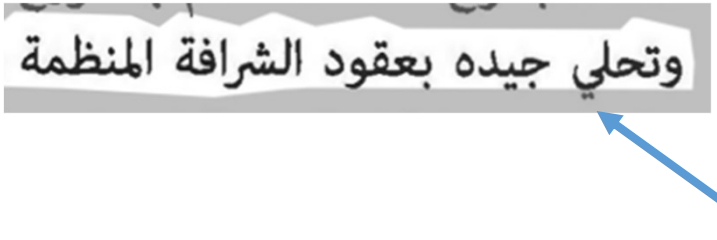
BIBLIOTHECA ARABICA

Herausforderung: Nichttrainierbarkeit der Masken

Trainierbar:
Regionen
Baselines



Bis dato nicht trainierbar:
Masken um Zeilen (Polygone)



Möglicher Workaround:

Statistik: seltener Endbuchstaben ى → Postkorrektur über Liste

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Herausforderung: Dependencies: py bidi

→ (fehlerhafte?) Python-Implementierung für Unicode bidirectional algorithm → Zero width non-joiner (Bindehemmer) gelöscht

می‌باید mit Bindehemmer (korrekt)

میباید ohne Bindehemmer (→ inkorrektes Persisch, ggf. bedeutungsunterscheidend)

می@باید **Workaround**: durch Zeichen ersetzen, das nicht gelöscht wird

https://github.com/MeirKriheli/python-bidi/issues/4

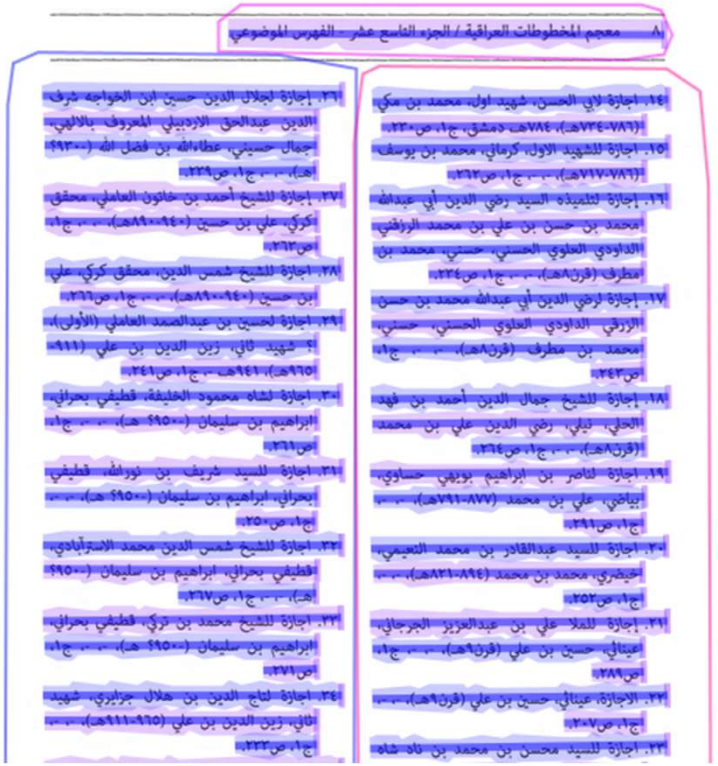
→ **Problem**: Abhängigkeit von Privatinitiativen / Einzelentwicklern

Sächsische Akademie der Wissenschaften zu Leipzig

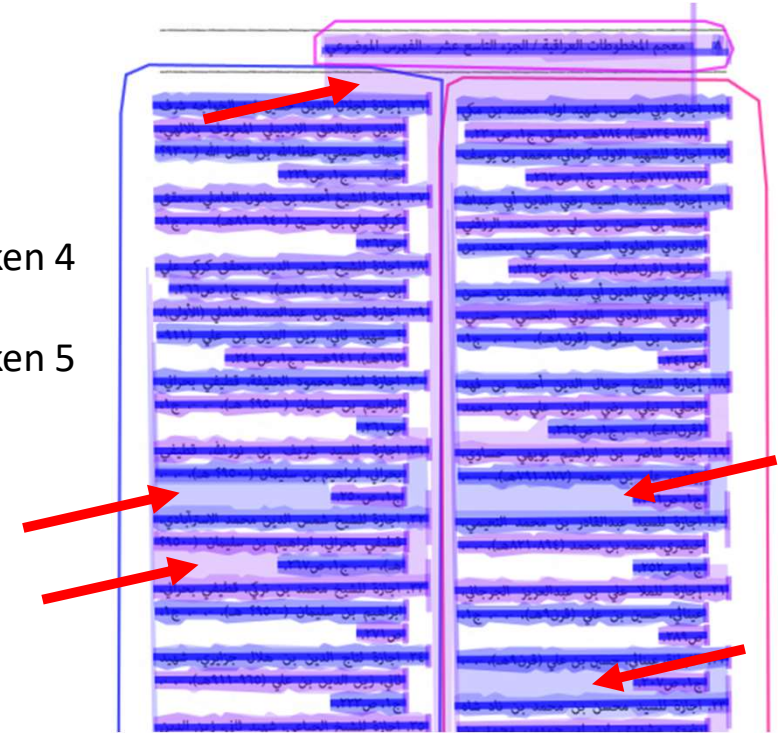
BIBLIOTHECA ARABICA

Herausforderung: Regression

Regression bei Masken (Polygone um Zeilen), mit kraken bis dato nicht trainierbar



kraken 4
→
kraken 5



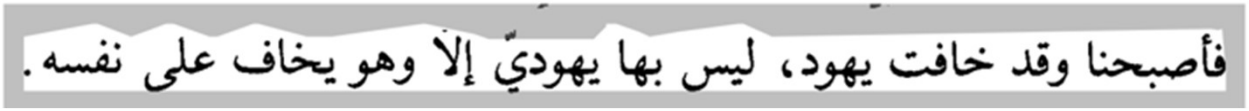
Workaround:
kraken 4: segmentation
kraken 5: recognition

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Herausforderung: manuelle Fehler + Overfitting

Bild



Erkannt فأصبحنا وقد خافت يهود، لوقعتنا بعدو الله فليس بها يهودي إلا وهو يخاف على نفسه.

Kauderwelsch (bestimmter Abschnitt)

?

Ground Truth فأصبحنا وقد خافت يهود، لوقعتنا بعدو الله فليس بها يهودي إلا وهو يخاف على نفسه.

! Grund: Nachnutzung vorhandener Scan-Text-Paare (Volltexte Bücher)

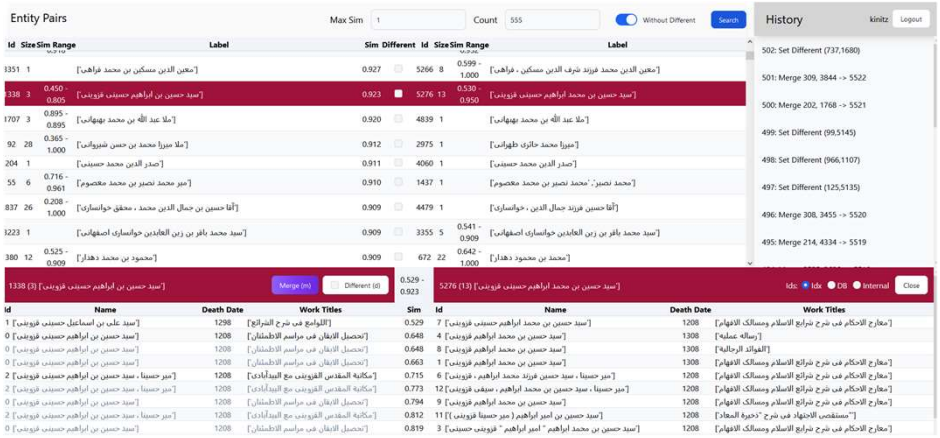
Lösungsansatz: Kenne deine Trainingsdaten/GT!
Kenne die Schwächen deines Modells!
Implementiere ausreichend Checks!

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Mögliche weitere Anwendungsfälle ML

- Handwritten Text Recognition (mehr Trainingsdaten)
- Normierung: identische Personen, Werke in Daten finden (s. Bild)
- Arabische Schrift ←→ latinisierte wiss. Umschrift (nicht eineindeutig)
- Retrieval Augmented Generation: Sprachmodell + eigene Quellen → Befrag den Chatbot über unsere Daten!



Vorschlagssystem für identische Personen (M. Reckziegel/ D. Kinitz)

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

https://gifer.com/de/UrpW

Literatur:

Daniel Kinitz/Thomas Efer (2023): *Towards a Dynamic Knowledge Graph of a Non-Western Book Tradition*
In: Baillot et al. (eds.). Digital Humanities 2023: Book of Abstracts. Graz 2023, pp.216-217.

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Fragen?
Anmerkungen?

Literatur:

Daniel Kinitz/Thomas Efer (2023): *Towards a Dynamic Knowledge Graph of a Non-Western Book Tradition*
In: Baillot et al. (eds.). Digital Humanities 2023: Book of Abstracts. Graz 2023, pp.216-217.

BIBLIOTHECA ARABICA

Leitung Prof. Verena Klemm