Workshop

Evaluierung automatischer Verfahren zur Inhaltserschließung

Maximilian Kähler
Deutsche Nationalbibliothek

2026-01-28

Willkommen…

…zum Workshop: **Evaluierung automatischer Verfahren zur Inhaltserschließung**

DEUTSCHE NATIONAL BIBLIOTHEK

Über mich:

- Maximilian Kähler, Deutsche Nationalbibliothek (DNB)
- Ausbildung: Mathematik & Theoretische Physik
- Projektverantwortlicher für das DNB-KI-Projekt zur Weiterentwicklung automatischer Erschließung

Über diesen Workshop:

- Techniken zur Evaluation automatischer Inhaltserschließungsverfahren mit Fokus auf praktische Anwendungen des **CASIMiR**[1]-Pakets.
- Entwickelt im Rahmen des DNB-AI-Projekts, gefördert vom Bundesbeauftragten für Kultur und Medien (BKM) [2]
- Alle Workshop-Materialien sind online verfügbar[3]

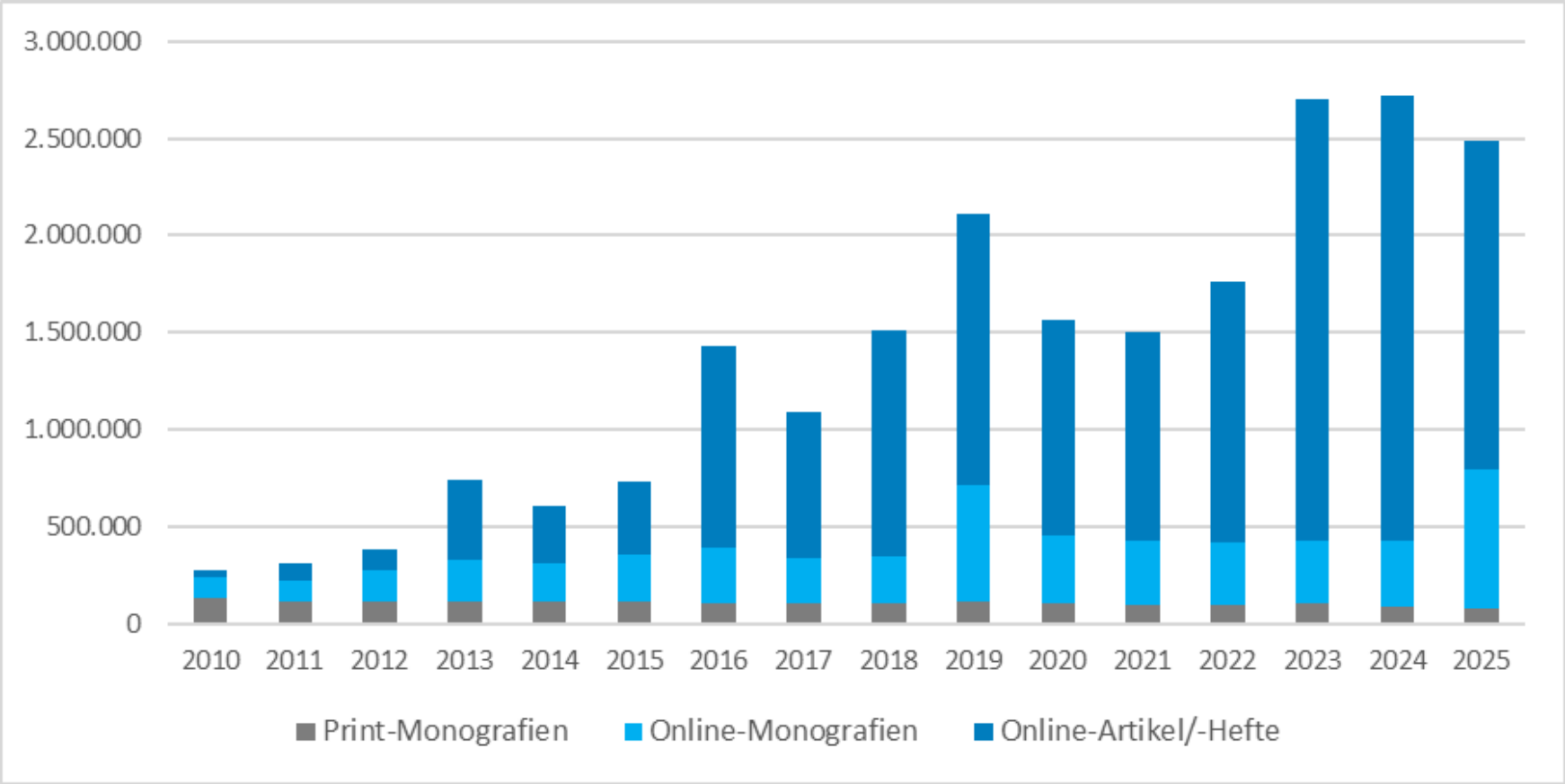
1. https://github.com/deutsche-nationalbibliothek/casimir
2. https://kulturstaatsminister.de/#
3. https://github.com/deutsche-nationalbibliothek/casimir-workshop

Einleitung

Inhaltsverzeichnis

DEUTSCHE NATIONAL BIBLIOTHEK

Warum automatische Sacherschließung?



- wachsendes Publikationsvolumen
- begrenzte personelle Kapazitäten für manuelle Sacherschließung
- Bedarf an konsistenten und skalierbaren Indexierungsverfahren

Automatische Indexierung an der DNB

- Indexierung produktiv seit 2014
- Vollständige Überarbeitung der Software-Architektur im April 2022 [1]
- Kernkomponente ist die Open-Source-Bibliothek annif [2]
- Jährlicher Durchsatz von ca. 170.000 Publikationen pro Jahr
- Unser Zielvokabular ist die GND[3] (Gemeinsame Normdatei), die über 1,4 Mio. potenzielle Konzepte enthält

1. Poley et al. 2025. “Automatic Subject Cataloguing at the German National Library.” https://doi.org/10.53377/lq.19422
2. Suominen 2019. “Annif: DIY Automated Subject Indexing Using Multiple Algorithms.” https://doi.org/10.18352/lq.10285.
3. https://gnd.network/

Warum automatische Sacherschließung schwierig ist

DEUTSCHE NATIONAL BIBLIOTHEK

- sehr große Vokabulare wie die GND oder LCSH (Tausende bis Millionen von Labels)
- sehr spärliche Annotationen (nur wenige korrekte Labels pro Dokument)
- stark ungleichmäßige Label-Verteilungen (einige Labels sind sehr häufig, andere sehr selten)
- unvollständige Ground Truth-Daten (nicht alle korrekten Labels werden bei der manuellen Erschließung zugewiesen)
- hierarchische Beziehungen zwischen Labels (manche Labels sind allgemeiner/spezifischer als andere)
- …

Verbesserung der automatischen Sacherschließung

Drei ergänzende Ansätze:

Bessere Algorithmen für die automatische Indexierung finden und schreiben

Die Komplexität des Problems reduzieren

Besser darin werden, gute Schlagwortextraktion zu diagnostizieren

Methoden für die automatische Indexierung

Überblick

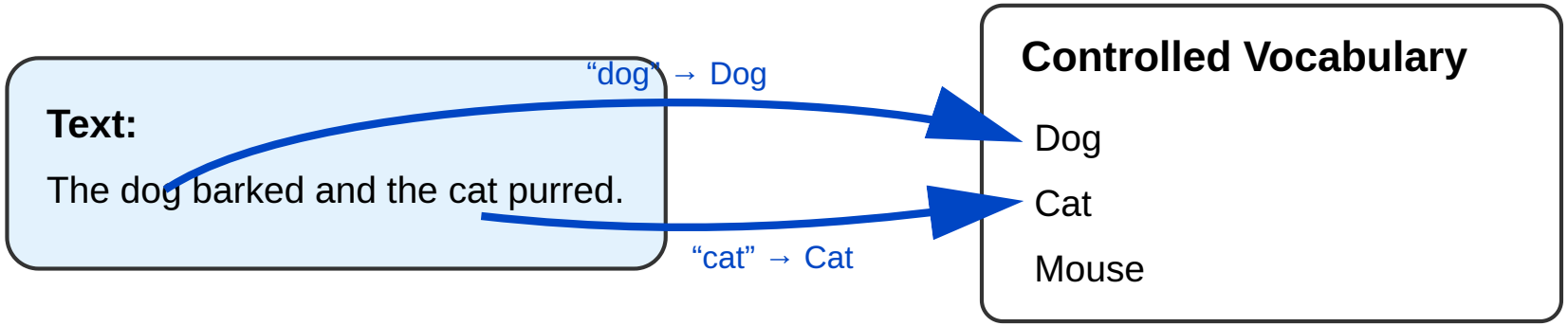
Vor dem Aufkommen von Deep Learning und Transformers:

- Lexikalische Verfahren [1,2]
- Statistische Verfahren für “Extreme Multi-Label Classification (XMLC)”:[3,4]
  - 1vsAll Verfahren [5]
  - Partitioned Label Trees [6,7,8]

2 and 6 are backends of annif

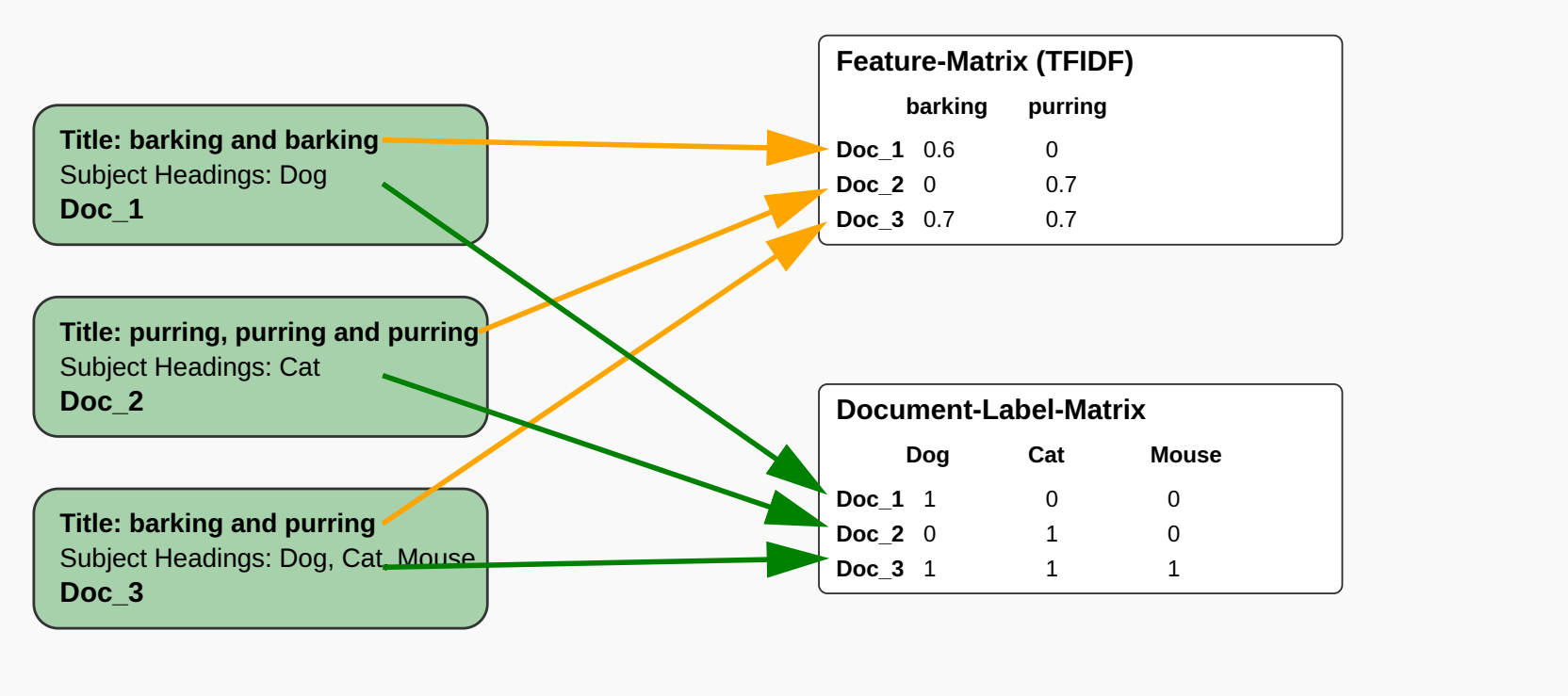
1. Medelyan, O., Frank, E., & Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction.
2. Maui-like Lexical Matching https://github.com/NatLibFi/Annif/wiki/Backend%3A-MLLM
3. Bhatia et al. 2016. “The Extreme Classification Repository: Multi-Label Datasets and Code.” http://manikvarma.org/downloads/XC/XMLRepository.html.
4. Dasgupta et al. 2023. “Review of Extreme Multilabel Classification,” February. https://arxiv.org/abs/2302.05971v2.
5. Schultheis, E. and Babbar, R. (2021) “Speeding-up One-vs-All Training for Extreme Classification via Smart Initialization.” Available at: https://doi.org/10.48550/arxiv.2109.13122
6. Khandagale, S., Xiao, H., & Babbar, R. (2020). Bonsai: diverse and shallow trees for extreme multi-label classification.
7. Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., & Varma, M. (2018). Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising
8. Omikuji Rust-Library: https://github.com/tomtung/omikuji

Lexikalisches Matching



- kann jedes Schlagwort im Vokabular zuordnen
- Heuristiken wie Häufigkeit, Position usw. werden verwendet, um die Schlagwörter nach der Zuordnung zu bewerten

Statistische Methoden



- statistische Ansätze lernen Korrelationen zwischen **Feature-Matrix** und **Dokument-Label-Matrix**
- statistische Ansätze benötigen keine prefLabels: Sie interpretieren die Beschreibung der Labels nicht

Nach der Erfindung der Transformer I

Anpassung alter Methoden mit Transformers:

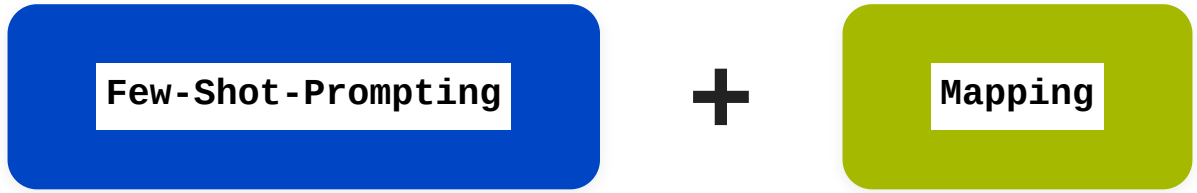
- Lexikalisches Matching → Embedding-basiertes Matching [1]
  - führt den Vergleich im Embedding-Raum durch, nicht lexikalisch
- Statistische Methoden:
  - Partitioned Label Trees → X-Transformer [2]
  - stellen Dokumentmerkmale (auch) mit Transformer-Embeddings dar, statt nur mit TFIDF

1. DNB-Prototype: https://github.com/deutsche-nationalbibliothek/ebm4subjects
2. Zhang et al. (2021). Fast Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification. https://arxiv.org/abs/2110.00685v2

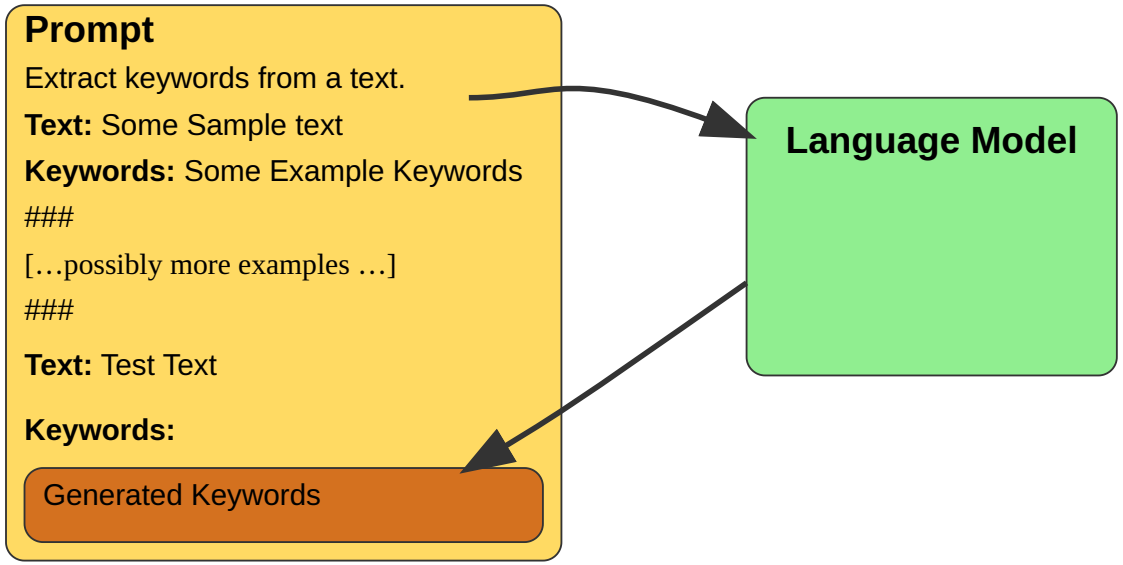
Nach der Erfindung der Transformer II

Nutzung generativer KI-Methoden:

- Prompt von LLMs zur Generierung von Schlagworten für Dokumente
- Verwendung von Few-Shot-Prompting zur Steuerung des Verhaltens von LLMs
- Mapping generierter Schlagwörter in ein kontrolliertes Vokabular



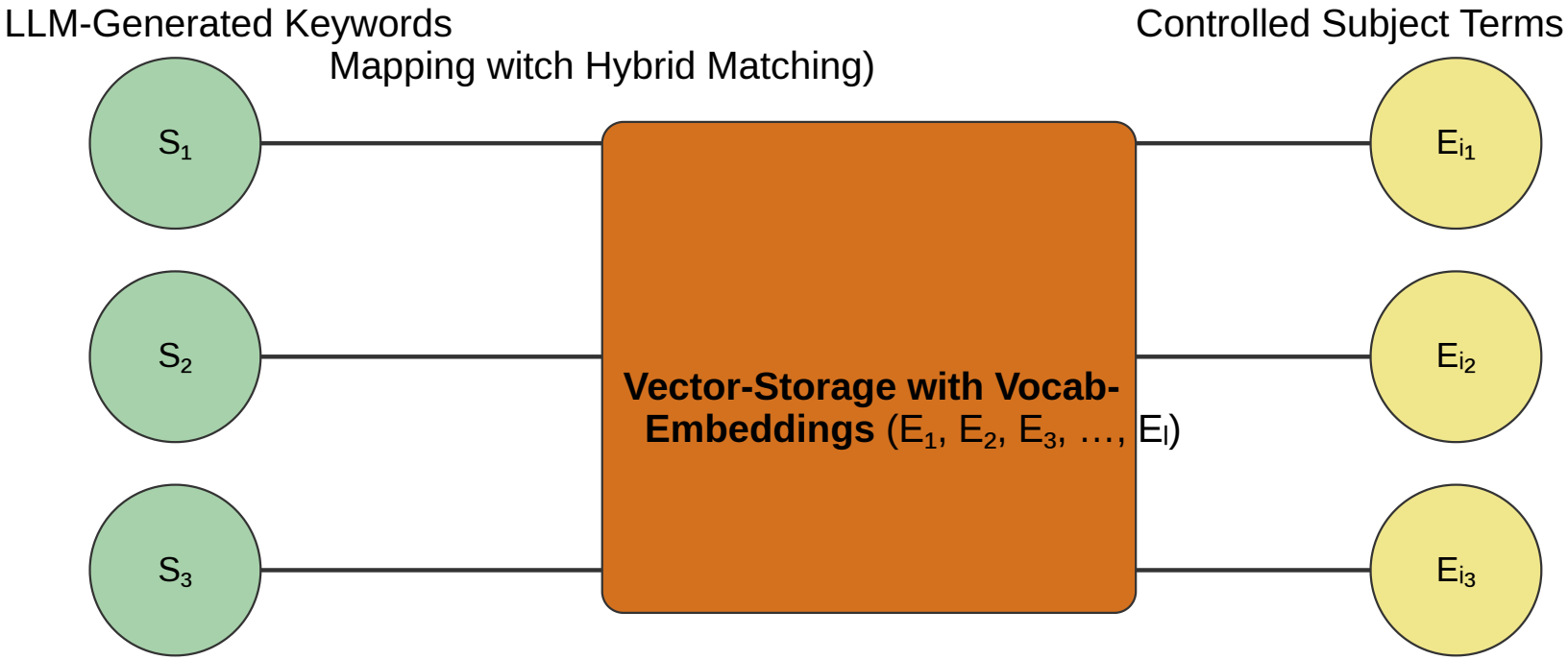
Die Idee des Few-Shot-Promptings



- nutzt Weltwissen in LLMs
- benötigt nur wenige Trainingsbeispiele

Mapping

- LLM hat kein Wissen über kontrolliertes Vokabular
- zusätzlicher Schritt mappt generierte Schlagwörter in das kontrollierte Vokabular
- Hybrid-Suche: lexikalisches Matching + embedding-basiertes Matching durch Vector Storage



Fortgeschrittene LLM-basierte Methoden

Few-Shot-Prompting und Mapping allein ergeben noch keine zufriedenstellenden Ergebnisse.

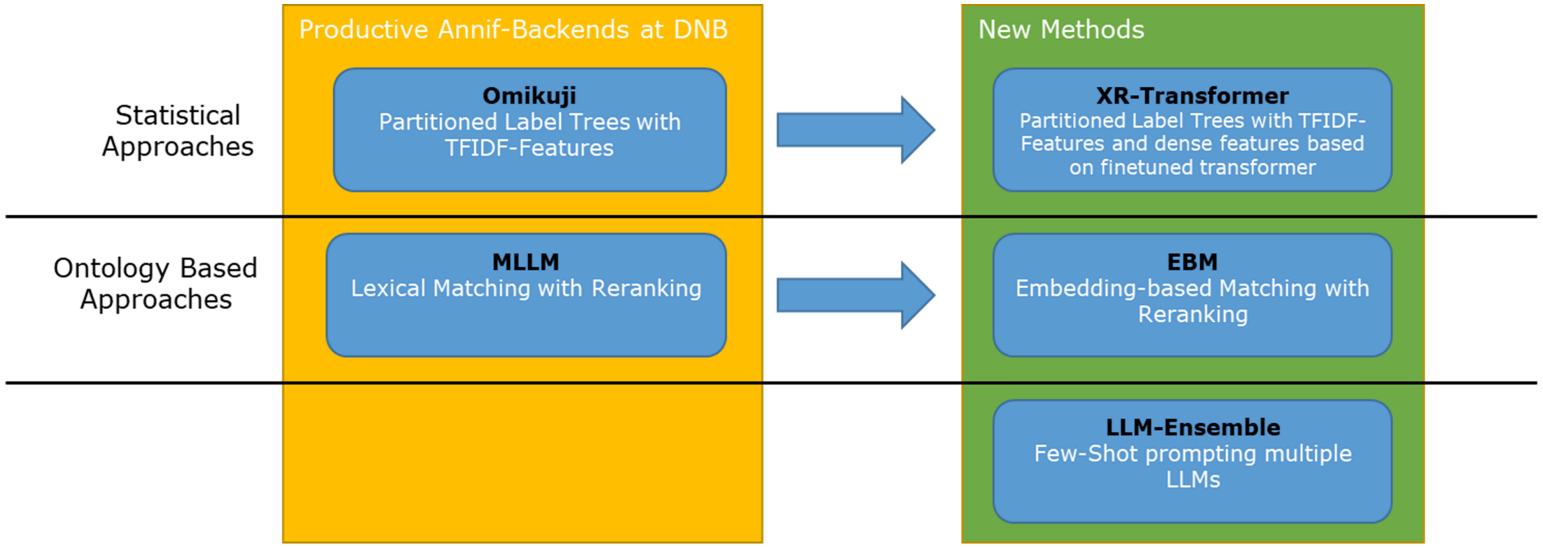
–> Fortgeschrittene Methoden:

- Kombination mehrerer LLM-basierter Methoden zu einem **LLM-Ensemble**[1]
- Verwendung von LLMs zum Sortieren und Filtern von Kandidatenlabels aus verschiedenen Methoden
- Nutzung von Retrieval zur Ergänzung von LLM-Prompts mit relevantem Kontext[2]

1. Kluge, L. and Kähler, M. (2025) “DNB-AI-Project at SemEval-2025 Task 5: An LLM-Ensemble Approach for Automated Subject Indexing,” https://aclanthology.org/2025.semeval-1.148/.
2. Kähler, M., Kluge, L. and Konermann, K. (2025) “DNB-AI-Project at the GermEval-2025 LLMs4Subjects Task: KIFSPrompt - Knowledge-Injected Few-Shot Prompting,” https://aclanthology.org/2025.konvens-2.42/.

Zusammenfassung der Methoden

Wir betrachten heute fünf verschiedene Methoden:



Ihre Aufgabe: finden Sie heraus, welches Verfahren hinter welchem anonymisierten Namen steht.[1]

Artful Accordion

Bold Bassoon

Charming Cello

Dreamy Didgeridoo

Embracing Euphonium

1. Icons created by Freepik - Flaticon

Datasets for this Workshop

Titel und Vokabular

Dokumenttitel:

- Testdatensatz der Deutschen Nationalbibliothek (DNB)
  - 8.415 Dokumenttitel mit intellektueller Schlagworterschließung
  - deutsche wissenschaftliche Literatur aus 18 ausgewählten Fachgruppen in gleichen Anteilen
- Verfügbar im `data/`-Ordner dieses Repositories
- `data/README.md` enthält Beschreibungen aller Datensatzdateien

Vokabular:

- ~200K eindeutige Konzepte, Teilmenge der Gemeinsamen Normdatei (GND)[1]
- jedes Schlagwort hat eine eindeutige Kennung (`label_id`) und einen bevorzugten Schlagworttext (`label_text`)
- die GND enthält Sachschlagwörter, Geografika, Personen, Körperschaften und mehr

1. https://gnd.network/

Maschinelle Vorschäge

- Verfahren sind anonymisiert für eine unvoreingenommene Bewertung
- Maschinelle Vorschläge von 5 verschiedenen Methoden sind unter `data/test-set_predictions/` verfügbar:
  - `artful-accordion.csv`
  - `bold-bassoon.csv`
  - …
- Schlüsselspalten sind `doc_id` für Dokumentkennungen und `label_id` für GND-Sachschlagwörter, welche von CASIMiR-Funktionen erwartet werden
- bis zu 100 vorgeschlagene Dokument-Label-Paare pro Methode und Dokument
- jedes vorgeschlagene Label besitzt auch eine `score` zwischen 0 und 1, welche die Konfidenzwerte der Methode angibt

Workshop-Spiel:

Können Sie erraten, welche Methode hinter welchem Dateinamen steckt?

Übung I: Manuelle Inspektion

Übung: Gehen Sie zu `workbooks/01_inspecting-results-manually.qmd`, um die Datentypen und Datensätze kennenzulernen.

If you don’t speak German use the `_eng` column names in the code examples instead of the `_ger` variants to see AI translated English texts.

Warning: English translations of document titles and GND subject labels are created by generative AI and provided for convenience only. They may contain errors or inaccuracies. Subject-Suggestions are all based on subject indexing methods applied to German texts and German labels.

Übung II: CASIMiR-Beispiel

DEUTSCHE NATIONAL BIBLIOTHEK

Qualitative Method Comparison

label_id	label_text	gold	score_artful-accordion	score_bold-bassoon	score_charming-cello	score_dreamy-didgeridoo	score_embracing-euphonium
1166742806 - Inklusive Schul- und Unterrichtsentwicklung Vom Anspruch zur erfolgreichen Umsetzung							
950251194	Transformation	FALSE	0.05997121	NA	NA	NA	NA
041316657	Anspruch	FALSE	0.31719807	NA	NA	0.1859743	NA
04126892X	Schulentwicklung	TRUE	NA	NA	0.11813986	0.4799480	0.858
041351487	Unterrichtsorganisation	TRUE	NA	NA	NA	NA	NA
1000723437	Inklusive Schule	TRUE	NA	0.2927039	0.04837416	0.7240117	0.066
965002845	Inklusion (Soziologie)	TRUE	NA	0.1999691	0.06500251	NA	0.148
041276612	Schulentwicklungsplanung	FALSE	NA	0.1276198	NA	NA	NA
04053474X	Schule	FALSE	NA	0.1385269	NA	NA	NA
100072185X	Inklusive Pädagogik	FALSE	NA	0.1656229	0.04911679	0.6998804	0.041
041351754	Unterrichtsforschung	FALSE	NA	NA	0.04595719	NA	NA
123322929X	Inklusiver Unterricht	FALSE	NA	NA	NA	0.3014980	NA
040118827	Deutschland	FALSE	NA	NA	NA	NA	0.099

Evaluation automatischer Verfahren

Metriken Grundlagen

Wir arbeiten hauptsächlich im **binären Relevanzszenario**: Vergleich von Vorhersagen mit einem Goldstandard

	Goldstandard ja	**Goldstandard nein**	
Vorhersage ja	True Positives (`tp`)	False Positives (`fp`)	$\mathrm{Prec} = \frac{tp}{tp+fp}$
Vorhersage nein	False Negatives (`fn`)	True Negatives (`tn`)	

$$\mathrm{Rec} = \frac{tp}{tp+fn}$$

Recall

Wie viele der erwarteten Goldstandard-Schlagwörter wurden tatsächlich vorgeschlagen?

Precision

Wie viele der vorgeschlagenen Schlagwörter sind tatsächlich korrekt?

Beispiel: Dokumentenlevel Precision und Recall

Betrachten wir das folgende Beispiel eines Dokumententitels:

> Medialer Habitus und biographische Legende -
> Schriftstellerische Inszenierungspraktiken im Zeitalter der Digitalisierung

label_id	label_text	gold-standard	suggested
041305450	Autorschaft	TRUE	TRUE
040359646	Literatur	FALSE	TRUE
040272230	Inszenierung	FALSE	TRUE
040533093	Schriftsteller	FALSE	TRUE
041230655	Digitalisierung	FALSE	TRUE
041223497	Selbstdarstellung	TRUE	FALSE

$$\text{Prec} = \frac{1}{1+4} = 0.2$$

$$\text{Rec} = \frac{1}{1+1} = 0.5$$

- tp = 1 (Autorschaft)
- fp = 4 (Literatur, Inszenierung, Schriftsteller, Digitalisierung)
- fn = 1 (Selbstdarstellung)
- tn = `#Vocab` - 6 (Alle Labels die nicht Gold-Standard sind und nicht vorgeschlagen wurden)

Metriken Grundlagen: Fortsetzung

- Precision und Recall können auf verschiedenen Aggregationsstufen berechnet werden:
  - pro Dokument
  - pro Schlagwort
  - insgesamt (Mikro-/Makro-averaged)
- Precision und Recall können zu F1-Score kombiniert werden:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Prec} \cdot \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}$$

- Bei der automatischen Indexierung, wenn Ergebnisse nach einem Konfidenzwert sortiert sind, interessieren uns oft die Top-k-Vorschläge:
  - Precision@k, Recall@k, F1@k
  - z.B. wie viele der Top-5 Vorschläge sind korrekt?

Übung II

Übung: Öffnen Sie `workbooks/02_set-retrieval-metrics.qmd` und berechnen Sie die besprochenen Metriken für die fünf Beispielverfahren.

Beispiel: Berechnung von Precision, Recall, und F1-Score für `artful-accordion`

```
compute_set_retrieval_scores(
  predicted = predictions[["artful-accordion"]],
  gold_standard = gold_standard,
  k = 5, # limit to top-5 suggestions
  rename_metrics = TRUE
)
```

```
# A tibble: 4 × 4
  metric  mode    value support
  <chr>   <chr>   <dbl>   <dbl>
1 f1@5    doc-avg 0.277    8415
2 prec@5  doc-avg 0.289    8197
3 rec@5   doc-avg 0.349    8415
4 rprec@5 doc-avg 0.414    8197
```

Übung III: Dimensionen der Evaluation

Datensätze haben oft mehrere **dokumentenbasierte** Dimensionen, die für eine geschichtete Evaluierung verwendet werden können:

- Sachgruppen
- Dokumenttypen
- Sprachen
- Erscheinungsjahre

Alternativ: Schichtung entlang **vokabulargestützter** Dimensionen:

- Entitätstypen (z.B. Themen, Orte, Personen)
- Fachgruppen (des Vokabluars, z.B. GND-Systematik)
- Hierarchieebenen
- Häufige vs. seltene Labels

Praktische Übung: Öffnen Sie `workbooks/03_stratified-set-retrieval-metrics.qmd`, um geschichtete Evaluation mit verschiedenen Dimensionen des bereitgestellten Datensatzes zu erkunden.

Übung III: Beispiel

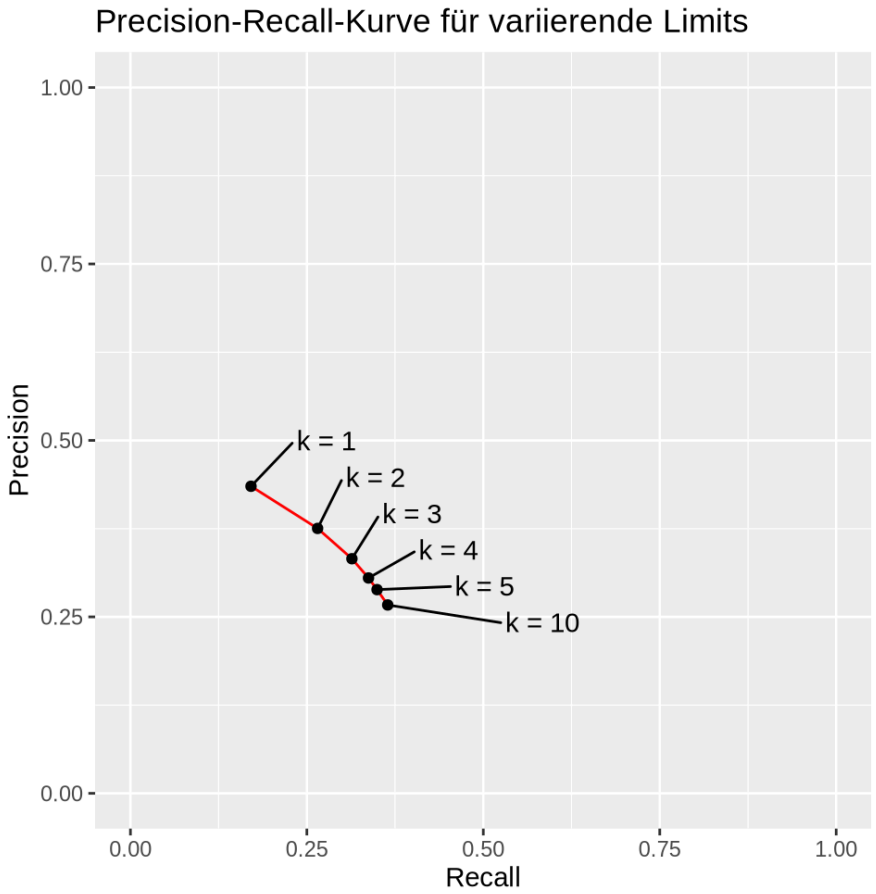
Gegeben die Zuordnungstabelle `subject_groups`, die jedem Dokument eine Sachgruppe zuweist, können wir geschichtete Retrieval-Ergebnisse berechnen:

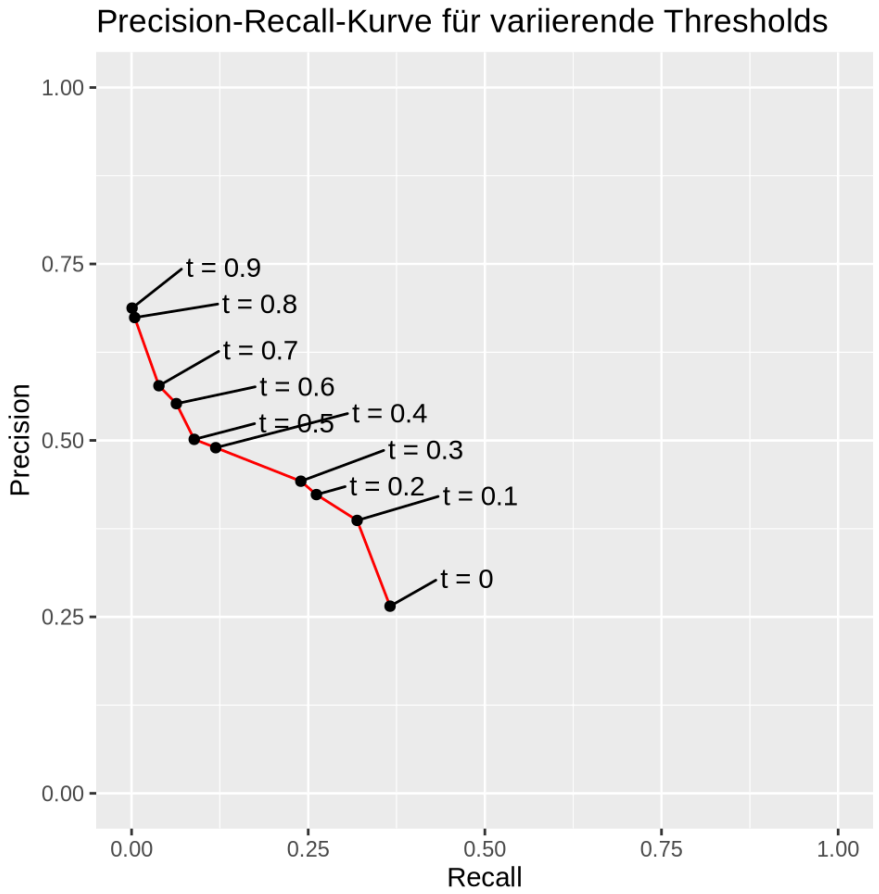
```
res_at_5_by_sg_artful_accordion <- compute_set_retrieval_score
  predicted = predictions[["artful-accordion"]],
  gold_standard = gold_standard,
  doc_groups = subject_groups,
  k = 5
)

kable(head(res_at_5_by_sg_artful_accordion, 3))
```

sg	sg_label_ger	sg_label_eng	metric	mode	value	support
004	Informatik	Computer Science	f1	doc-avg	0.2111046	500
100	Philosophie	Philosophy	f1	doc-avg	0.3383969	500
150	Psychologie	Psychology	f1	doc-avg	0.2642708	500

Precision-Recall Kurven

Bislang: retrieval Metriken für festen Limit `k = 5`





Precision-Recall-Kurven zeigen den Kompromiss zwischen Precision und Recall bei verschiedenen Schwellenwerten.

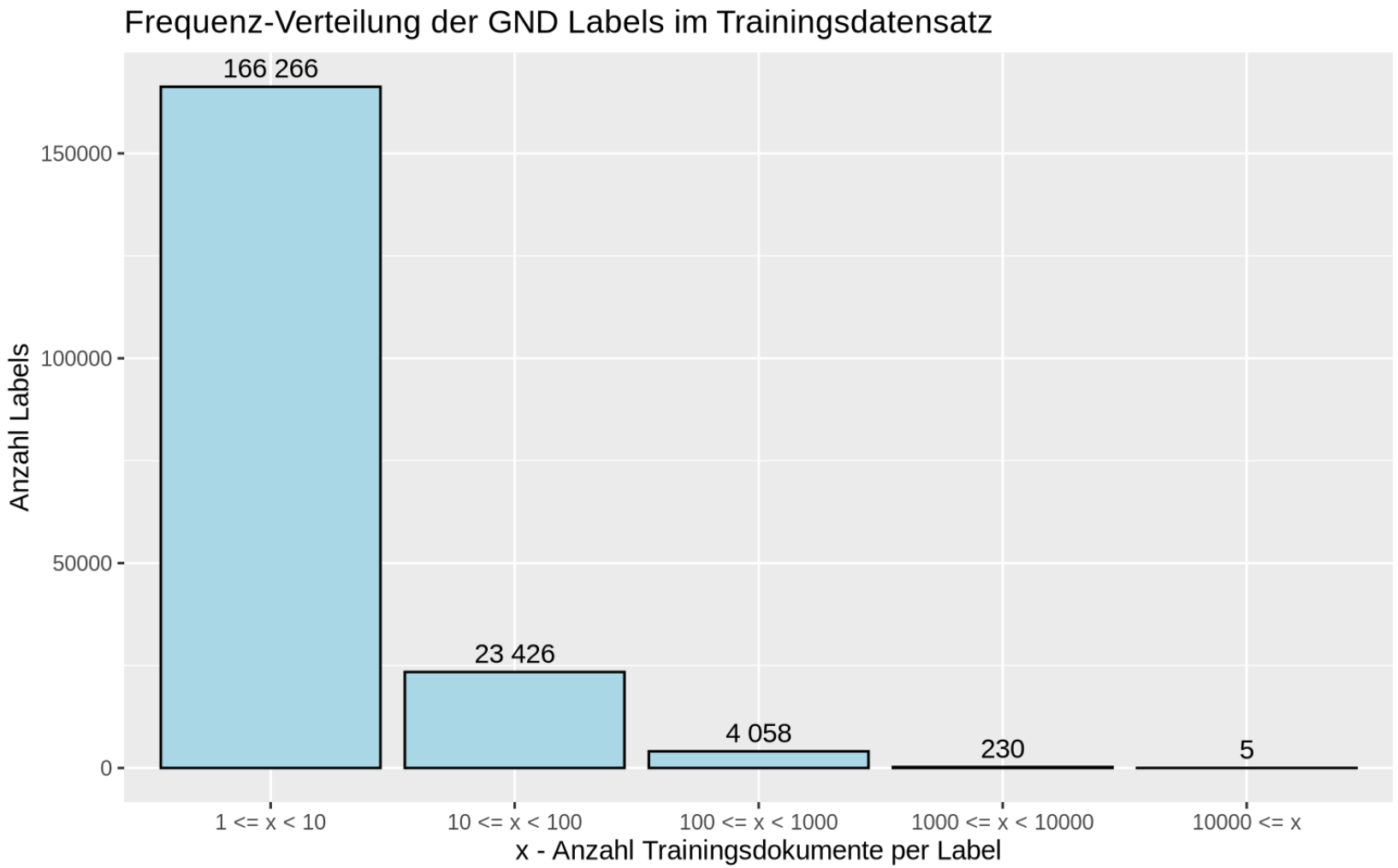
Übung IV: Precision Recall Kurven

Praktische Übung: Gehen Sie zu `workbooks/04_precision-recall-curves.qmd`, um Precision-Recall-Kurven für die bereitgestellten Verfahren zu berechnen.

Beispiel: Berechnung der Precision-Recall-Kurve mit CASIMiR

```
pr_curve <- compute_pr_curve(
  predicted = predictions[["artful-accordion"]],
  gold_standard = gold_standard,
  steps = 10 # number of thresholds to evaluate
)

ggplot(pr_curve$plot_data, aes(x = rec, y = prec_cummax)) +
  geom_point() +
  geom_path() +
  coord_fixed(xlim = c(0, 1), ylim = c(0, 1)) +
  ggtitle("Precision-Recall-Kurve für artful-accordion, berech
```

Analyse des Long Tail



Trainingsdaten: ~ 950T Buchtitel aus der DNB mit intellektuellen GND-Labels

Übung V: Analyse des Long Tail

Übung: Gehen Sie zu `workbooks/05_analysing-the-long-tail.qmd`, um die Leistung von Verfahren zur automatischen Erschließung auf häufige vs. seltene Schlagwörter zu analysieren.

Übung VI: Abschluss der Übung

Übung: Öffnen Sie `workbooks/06_final-assignment.qmd` and lösen Sie das Quiz des Tages?

Ihre Aufgabe: identifizieren Sie, welches Verfahren hinter welchem anonymisierten Namen steht.[1]

Artful Accordion

Bold Bassoon

Charming Cello

Dreamy Didgeridoo

Embracing Euphonium

1. Icons created by Freepik - Flaticon

Auflösung: Methoden-Quiz

Instrument	Verfahren
Artful Accordion	Lexikalisches Matching
Bold Bassoon	LLM-Ensemble
Charming Cello	X-Transformer (Partitioned Label trees mit TFIDF- und Embedding-features
Dreamy Didgeridoo	Embedding-basiertes Matching
Embracing Euphonium	Omikuji (Partitioned Label trees mit TFIDF-features)

Haben Sie Ihr Lieblingsverfahren nicht gefunden?

Erwägen Sie, sie in **Ensembles** zu kombinieren, um die Leistung zu verbessern![1]

1. Orchestra icons created by mia elysia - Flaticon

Fortgeschrittene Themen

Bonus: Conditional label weights

Haben True Positives, False Positives und False Negatives den gleichen Wert?

Wir können verschiedene **Kosten** unterschiedlichen Arten von Fehlern zuordnen:

	Goldstandard ja	**Goldstandard nein**
Vorhersage ja	$C_{\mathrm{tp}} \cdot \mathrm{tp}$	$C_{\mathrm{fp}} \cdot \mathrm{fp}$
Vorhersage nein	$C_{\mathrm{fn}} \cdot \mathrm{fn}$	$C_{\mathrm{tn}} \cdot \mathrm{tn}$

Angenommen, es gelten **labelabhängige Kosten** für True Positives und False Negatives:

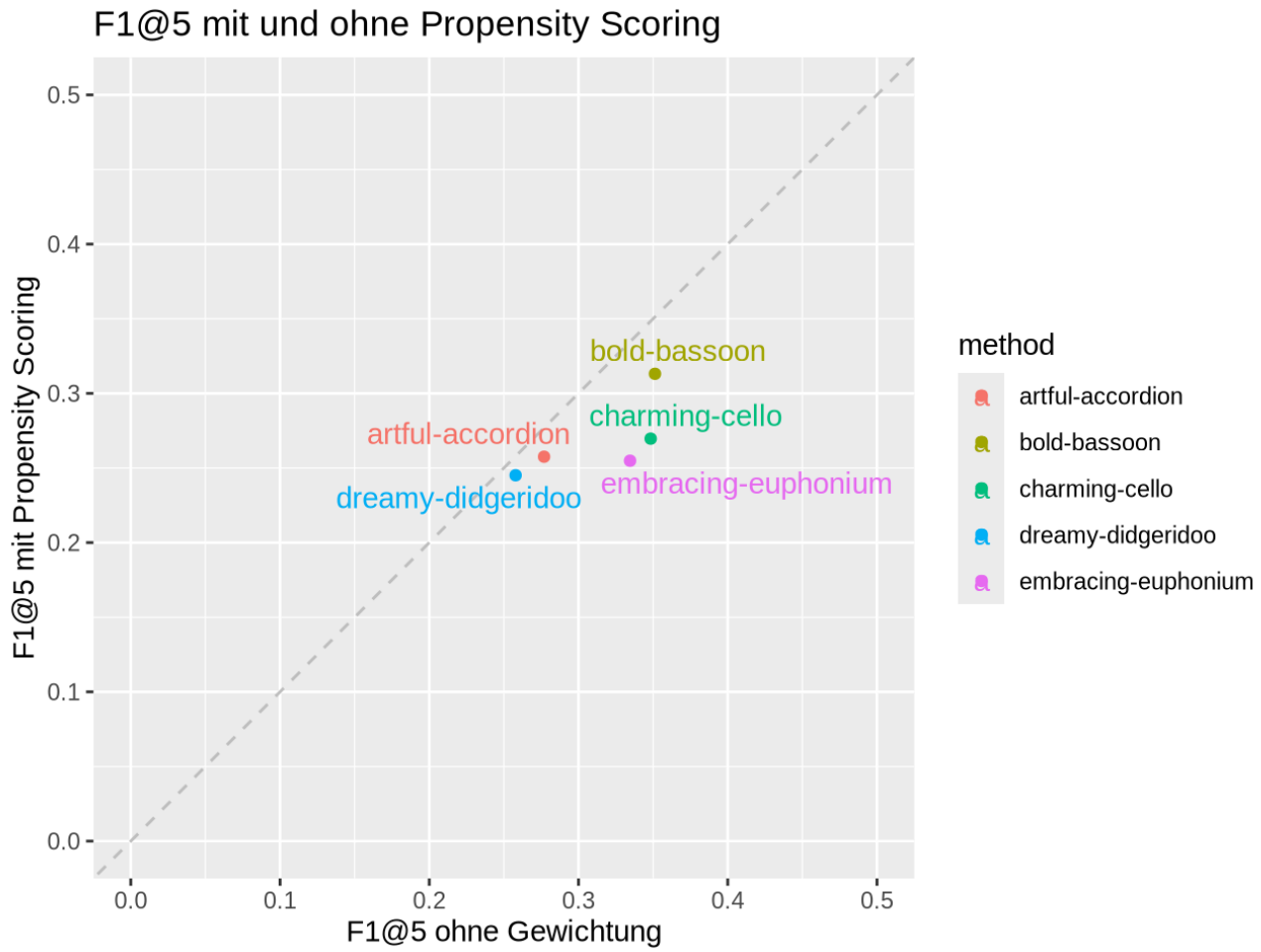
$$C_{\mathrm{tp}} = C_{\mathrm{fn}} = w_\lambda,$$

und konstante Kosten für False Positives:

$$C_{\mathrm{fp}} = const$$

wobei w_λ ein Gewicht ist, welches vom Label λ abhängt und umgekehrt proportional zu dessen Häufigkeit in den Trainingsdaten ist.

Bonus: Conditional label weights II



Bonus: Jenseits von binärer Relevanzbewertung I

Betrachten wir erneut den folgenden Beispieltitel:

> Medialer Habitus und biographische Legende -
> Schriftstellerische Inszenierungspraktiken im Zeitalter der Digitalisierung

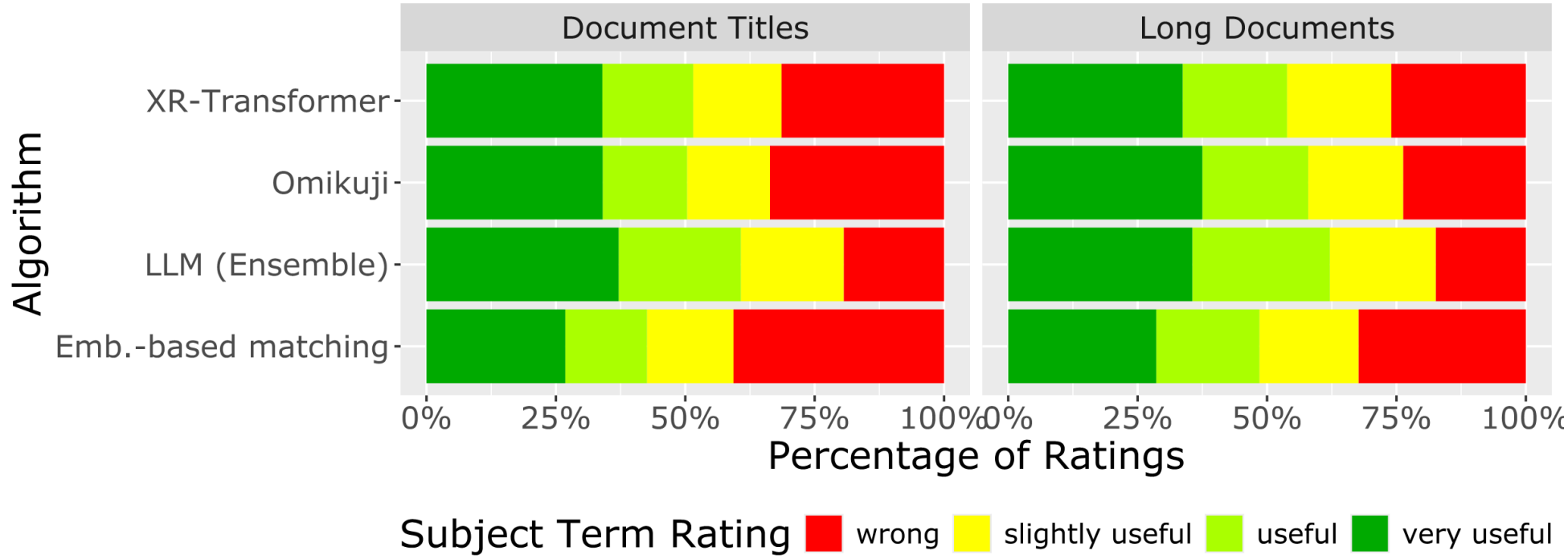
label_id	label_text	gold-standard	suggested
041305450	Autorschaft	TRUE	TRUE
040359646	Literatur	FALSE	TRUE
040272230	Inszenierung	FALSE	TRUE
040533093	Schriftsteller	FALSE	TRUE
041230655	Digitalisierung	FALSE	TRUE
041223497	Selbstdarstellung	TRUE	FALSE

- Unvollständiger Goldstandard: Nicht alle korrekten Schlagwörter werden bei der manuellen Schlagworterschließung zugewiesen
- Falsch Positive können dennoch hilfreiche Vorschläge aus Nutzenden-Perspektive (Retrieval) sein

Bonus: Beyond binary relevance II

Gestufte (ordinale) Relevanzniveaus können verwendet werden, um dies zu modellieren, z.B.:

Ordinales Relevanzniveau
sehr hilfreich
hilfreich
wenig hilfreich
falsch



Generalized Precision und Generalized Recall

Gestufte Relevanz führt zur Idee von verallgemeinerten Metriken: **Generalized Precision** und **Generalized Recall** [1]

Ordinales Relevanzniveau	Metrische Relevanzstufe r
sehr hilfreich	1
hilfreich	2/3
wenig hilfreich	1/3
falsch	0

Generalised Precision

$$\frac{tp + \Delta_{rel}}{tp + fp}$$

Generalised Recall

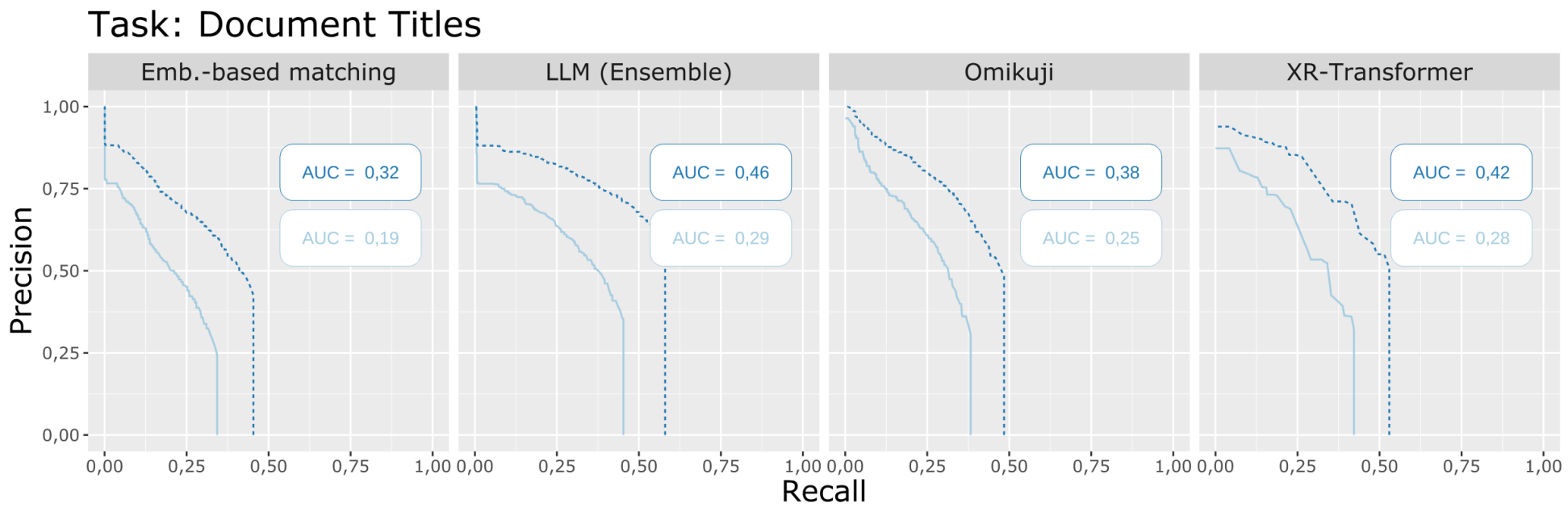
$$\frac{tp + \Delta_{rel}}{tp + fn + \Delta_{rel}}$$

mit $\Delta_{rel} = \sum_{i \in false\ positives} r_i$ und mit metrischen Relevanzbewertungen $0 \leq r_i \leq 1$

1. Kekäläinen, Jaana, and Kalervo Järvelin. 2002. “Using Graded Relevance Assessments in IR Evaluation.” Journal of the American Society for Information Science and Technology 53 (November): 1120–29. https://doi.org/10.1002/asi.10137.

Verallgemeinerte Precision-Recall-Kurven

Verallgemeinerte Precision-Recall-Kurven werden ähnlich zu Precision-Recall-Kurven mit binärer Relevanzbewertung berechnet (auch in CASIMiR implementiert):



Diskussion

Diskussion

Ihre Meinung:

Welche anderen Aspekte gehören zur Evaluation automatischer Erschließung?

DEUTSCHE NATIONAL BIBLIOTHEK

The End

Vielen Dank!

Kontakt

Maximilian Kähler

Deutsche Nationalbibliothek

m.kaehler@dnb.de
https://www.dnb.de
@mfakaehler@openbiblio.social