DEUTSCHE NATIONAL BIBLIOTHEK

EMa – Erschließungsmaschine

Automatische Vergabe von GND-Schlagwörtern mit Annif - Ergebnisse einer Evaluation im DNB-Projekt EMa

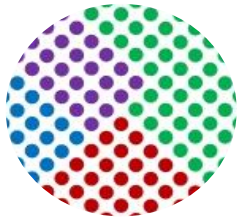
Sandro Uhlmann | Deutsche Nationalbibliothek

EMa – Erschließungsmaschine

DEUTSCHE NATIONAL BIBLIOTHEK

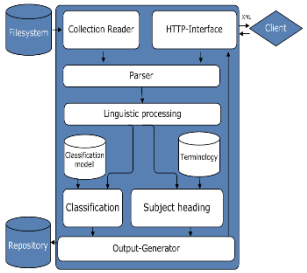
- Projekt EMa: Ausgangslage, Ziele, Evaluation
- Ergebnisse der automatischen Indexierung mit GND-Vokabular
- Ausblick

Maschinelle Inhaltserschließung in der DNB



Maschinelle Klassifizierung von Netz- und ausgewählten Printpublikationen mit DDC-Sachgruppen und DDC-Kurznotationen (Support Vector Machine)

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen anhand der normierten Terminologien GND und LCSH (Text Mining, Textstatistik)



Software:

Averbis Extraction Platform (AEP)

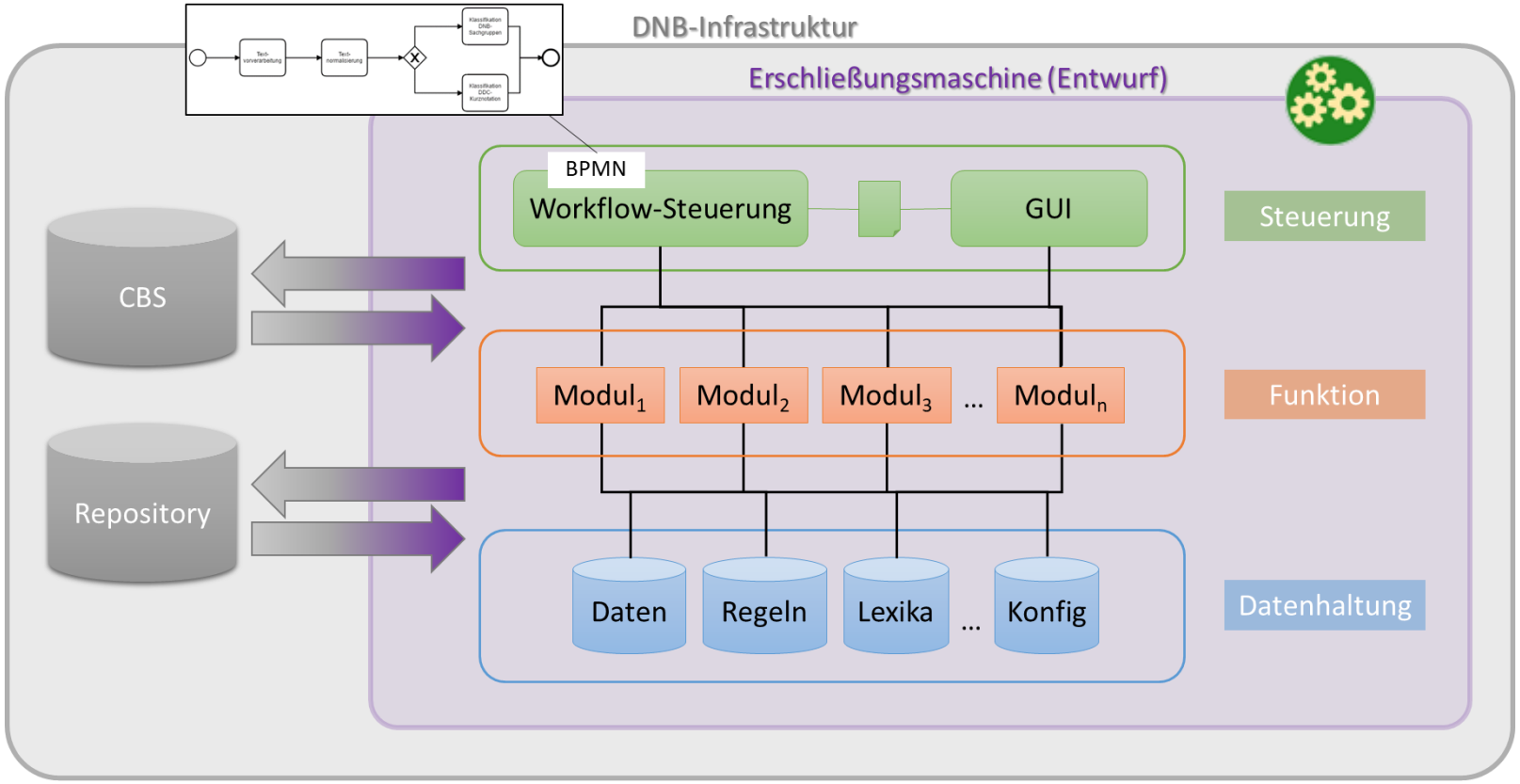
Erschließungsmaschine - Ziele

Ziel des Projektes EMa ist die Ablösung des bislang eingesetzten Averbis-Software (Altsystem) durch ein den fachlichen und technischen Anforderungen entsprechendes, modular aufgebautes System zur maschinellen Inhaltserschließung.

Stichworte (Auswahl):

#Klassifizierung mit unterschiedlichen Klassifikationssystemen
#Indexierung mit kontrolliertem Vokabular
#Sprachenidentifizierung #Modularität
#Erweiterung um neue Funktionen oder Verfahren
#kontinuierliche Verbesserung der Erschließungsergebnisse u.a.

Erschließungsmaschine – Skizze einer modularen Architektur



Projektstruktur und Planung

Phase 1:

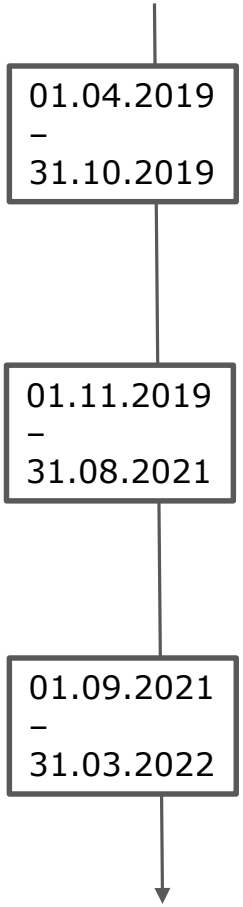
Anforderungssammlung, Marktsichtung

Phase 2:

Evaluation, Anschaffung, Anpassung und Implementierung

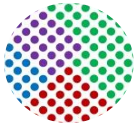
Phase 3:

Produktivnahme und Konsolidierung (sukzessive Ablösung des Altsystems durch das neue Erschließungssystem)



Projektphase 2.1A: Evaluation

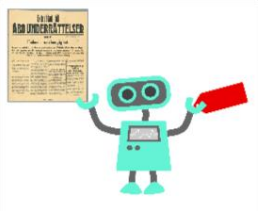
Für eine erste Iteration wurden zunächst nur die Kernfunktionalitäten betrachtet, die für eine Ablösung des Altsystems prioritär sind:



Automatische Klassifizierung (DDC-Sachgruppen oder DDC-Kurznotationen) für deutsch- und englischsprachige Publikationen

Automatische Beschlagwortung mit GND-Vokabular für deutschsprachige und mit GND- und LCSH-Vokabular für englischsprachige Publikationen

Projektphase 2.1A: Evaluation



- annif Open Source Werkzeugkasten zur automatischen Klassifizierung und Indexierung, entwickelt an der Finnischen Nationalbibliothek in Helsinki, Chefentwickler: Osma Suominen
- sprachunabhängig, kombiniert verschiedene Verfahren des Text Mining und des maschinellen Lernens
- Verfahren können einzeln genutzt werden oder in Kombination (Ensembles)

Siehe auch http://annif.org/ und Suominen, O., 2019. Annif: DIY automated subject indexing using multiple algorithms. LIBER Quarterly, 29(1), pp.1–25.

Annif – Testcases GND-Schlagwörter

Testcase GND Top-0

Alle GND-Konzepte*

Anzahl: 1.254.577

Testcase GND Top-1

Alle GND-Konzepte* mit mind. 1 Titeldatenverknüpfung in 041A* im Bestand der DNB

Anzahl: 339.099

Trainingsmaterial, deutsch:

Es existiert kein vollständiges Trainingsset für alle GND-Konzepte, nur 339.099 haben mind. einen Textobjekt, 915.478 haben kein Textobjekt.

*Katalogisierungslevel 1 oder z und aus dem Teilbestand s

Annif – Testcases GND-Schlagwörter

DEUTSCHE NATIONAL BIBLIOTHEK

Testcase GND Top-0

Alle GND-Konzepte*

Anzahl: 1.254.577

Testcase GND Top-1

Alle GND-Konzepte* mit mind. 1 Titeldatenverknüpfung in 041A* im Bestand der DNB

Anzahl: 339.099

Trainingssets

1.164.773 Titel

627.504 TOCs plus Titel

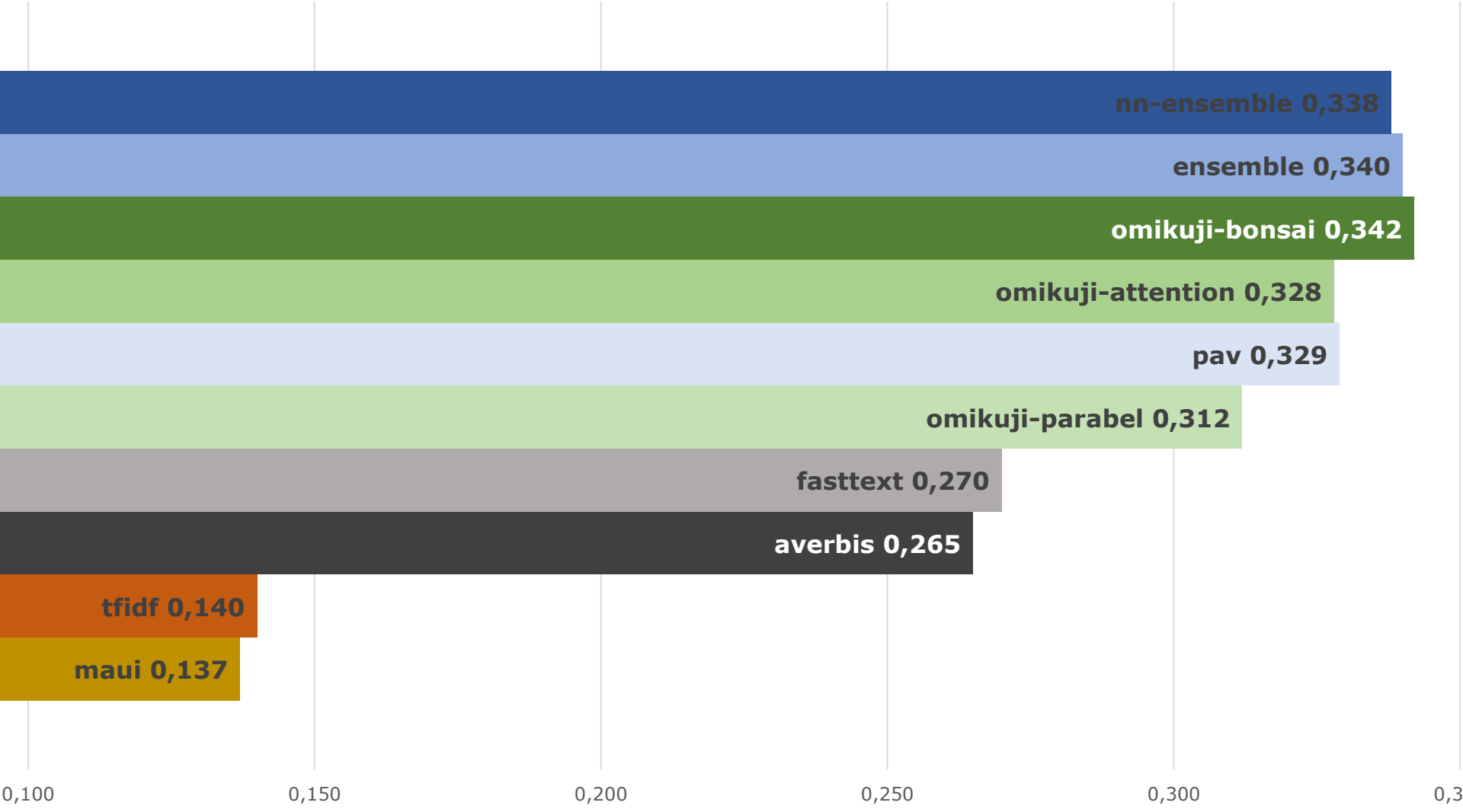
126.213 Volltexte plus Titel

*Katalogisierungslevel 1 oder z und aus dem Teilbestand s

Annif - Ergebnisse GND-Schlagwörter

DEUTSCHE NATIONAL BIBLIOTHEK

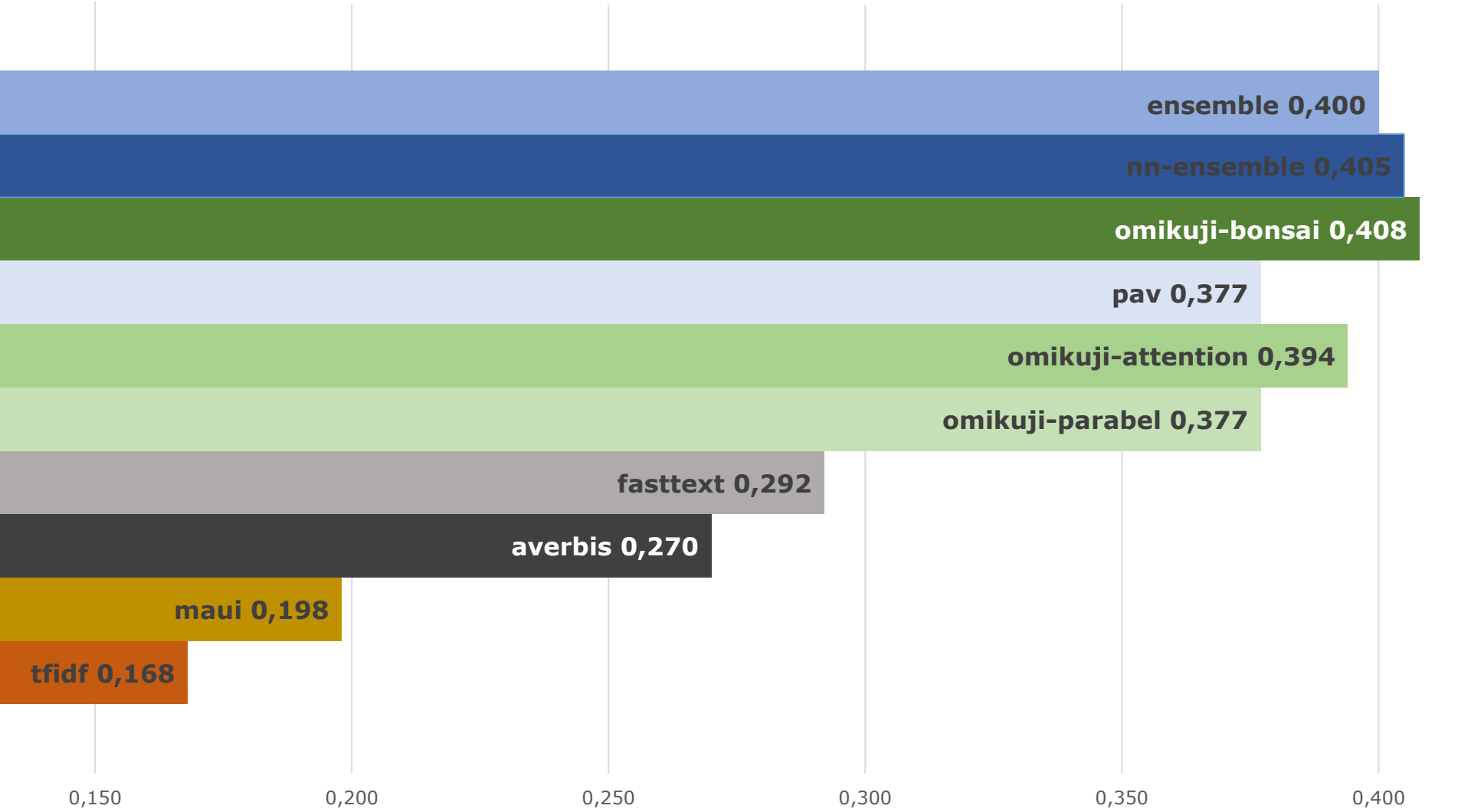
Testset: 1.068 NPs | Gold-Standard: F1-Score (n=5 SW)



Annif - Ergebnisse GND-Schlagwörter

DEUTSCHE NATIONAL BIBLIOTHEK

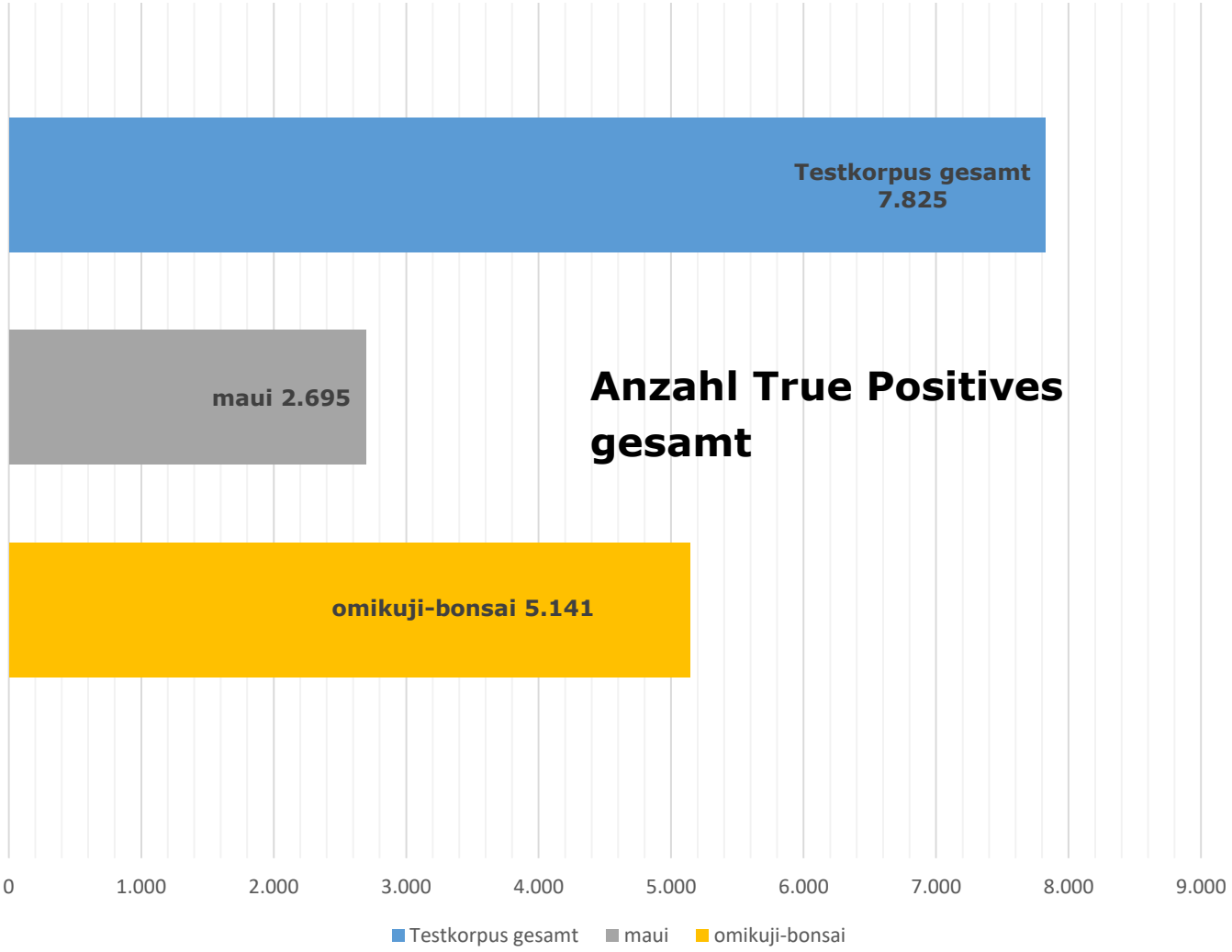
Testset: 937 TOCs | Gold-Standard: F1-Score (n=5 SW)



Annif – das assoziative Verfahren omikuji-bonsai und das lexikalische Verfahren maui im Vergleich

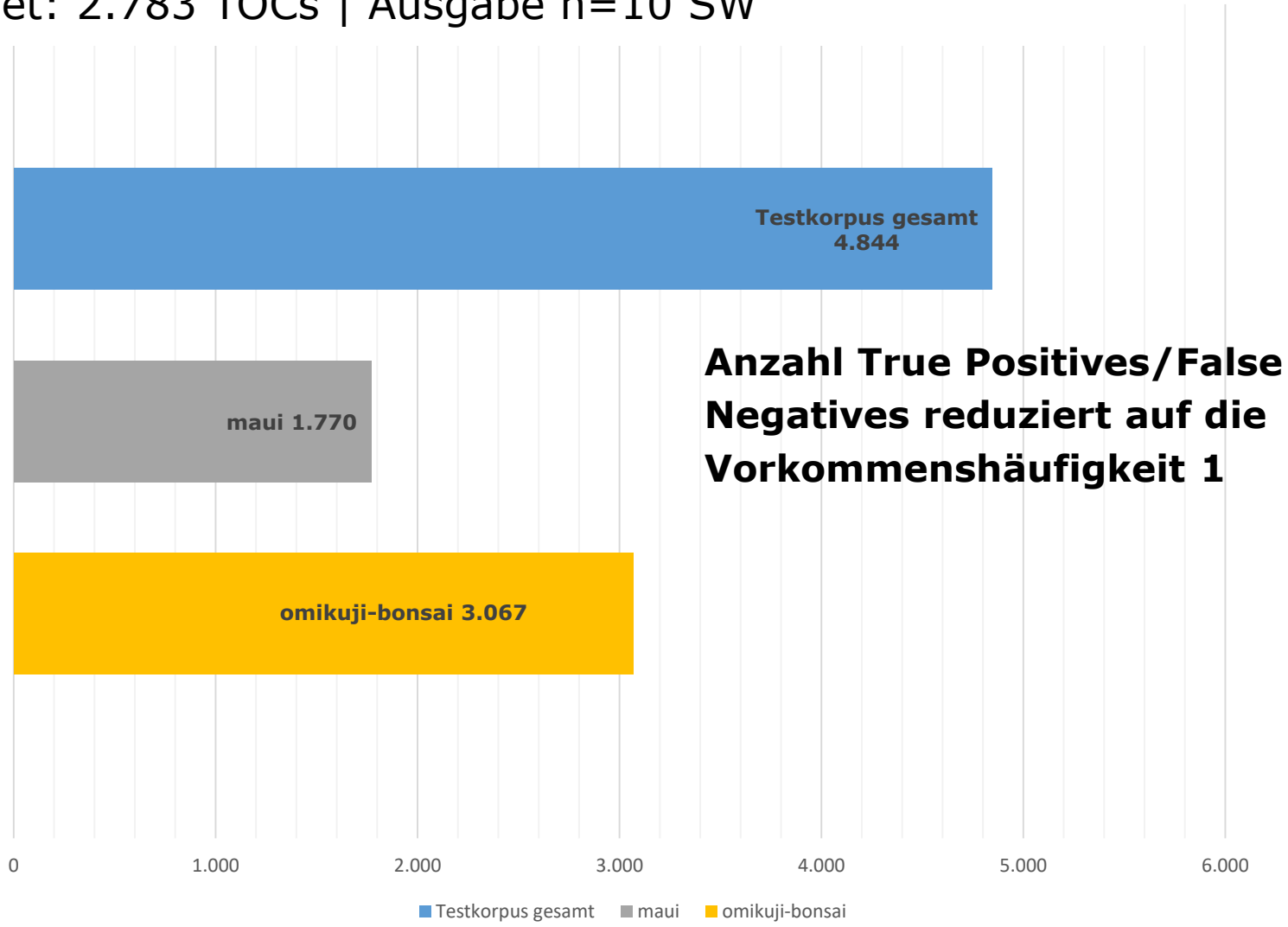
DEUTSCHE NATIONAL BIBLIOTHEK

Testset: 2.783 TOCs | Ausgabe n=10 SW



Annif – Vergleich omikuji-bonsai vs. maui

Testset: 2.783 TOCs | Ausgabe n=10 SW



Anzahl True Positives/False Negatives reduziert auf die Vorkommenshäufigkeit 1

Annif – Vergleich omikuji-bonsai vs. maui

Testset: 2.783 TOCs | Ausgabe n=10 SW

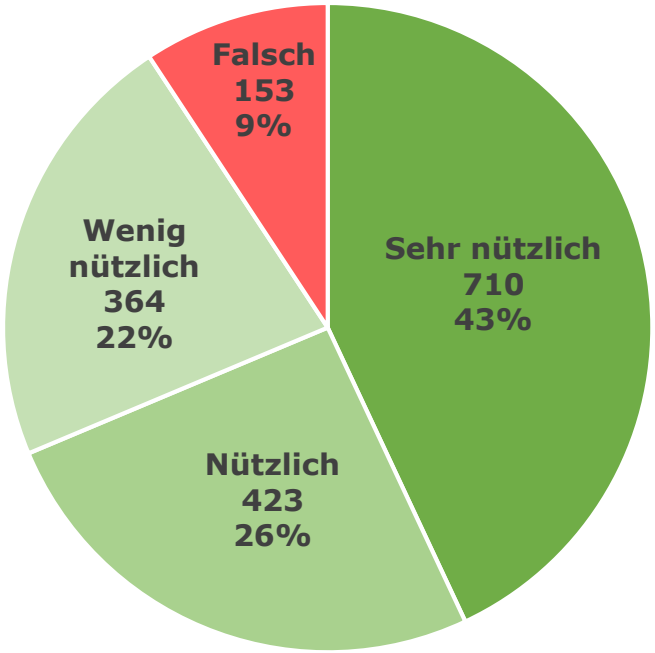
Anzahl True Positives/False Negatives reduziert auf die Vorkommenshäufigkeit 1

Annif – Intellektuelle Bewertung

Einzelbewertung

434 Datensätze (NPs)

1.650 durch Annif vergebene GND-Schlagwörter



Intellektuelle Bewertung durch die Indexierer* der Abteilung Inhaltserschließung

Bewertungsskala:

- Sehr nützlich
- Nützlich
- Wenig nützlich
- Falsch

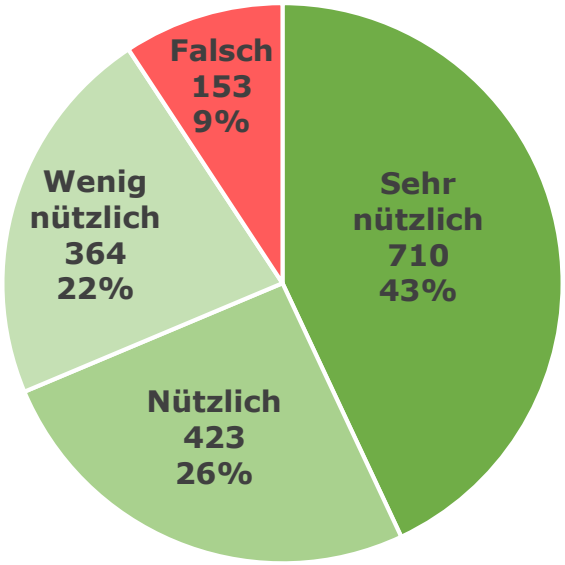
GND-Schlagwörter n = max 5

735 Fehlende Aspekte (1,69 Ø)

Vergleich* Annif vs. Averbis

434 Datensätze (NPs)

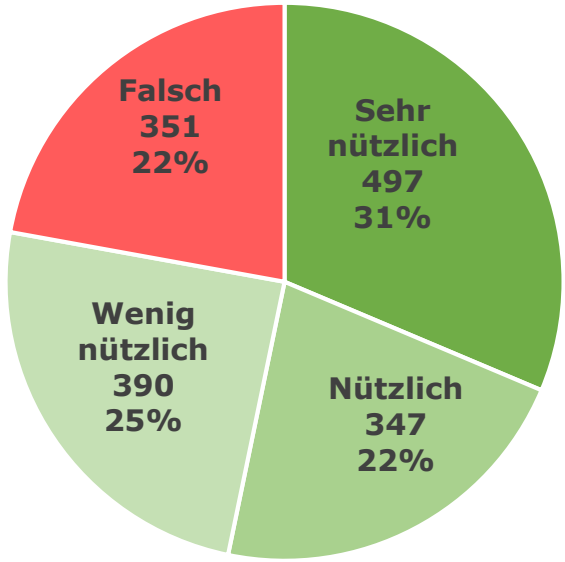
1.650 Annif mSW



735 Fehlende Aspekte (1,69 Ø)

434 Datensätze Reihe O 2020 (NPs)

1.585 Averbis mSW



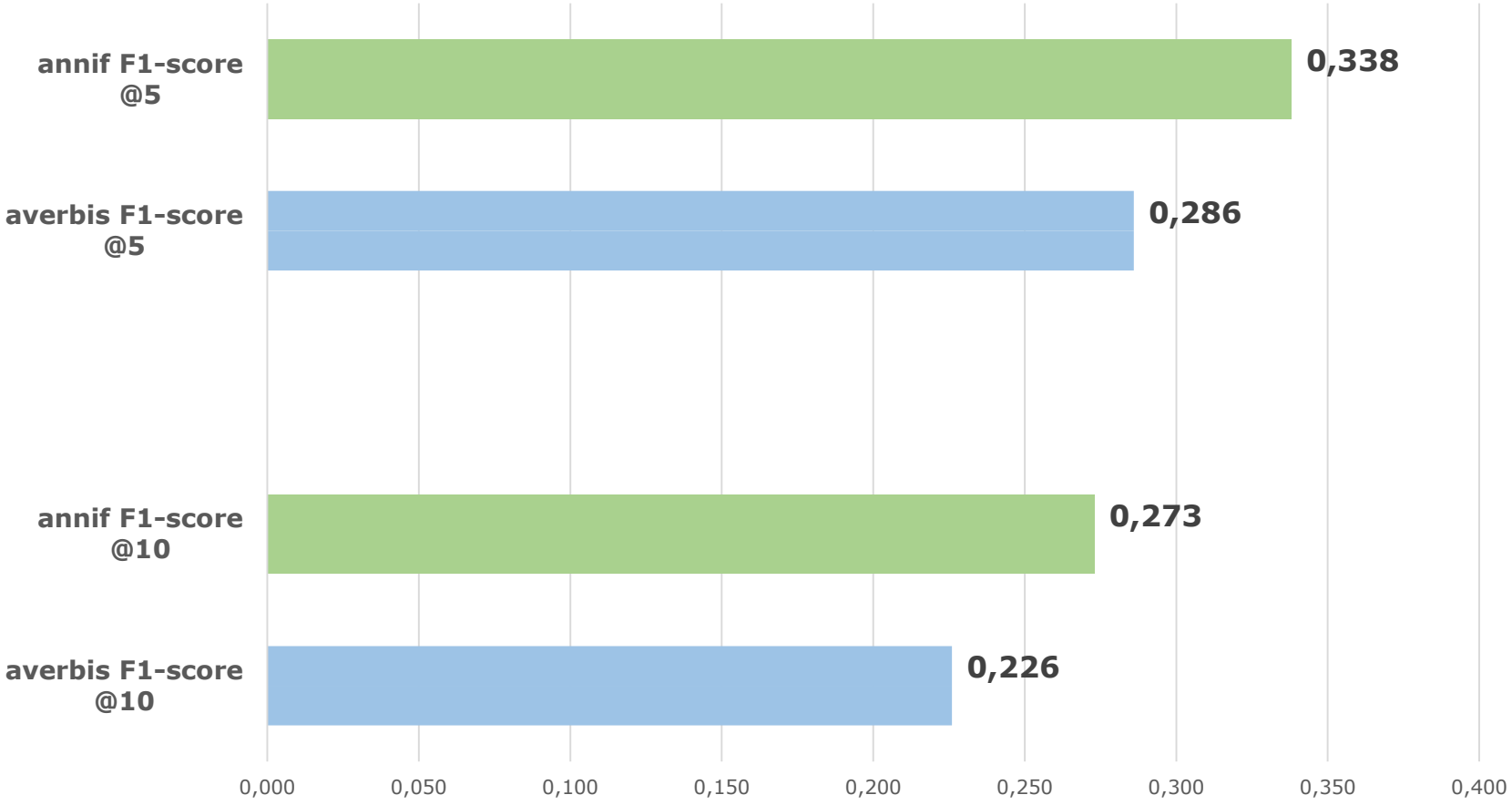
859 Fehlende Aspekte (1,98 Ø)

*kein 1:1 Vergleich, Averbis: gleiche Anzahl Datensätze, zufällig ausgewählt

Vergleich ground truth - annif vs. averbis

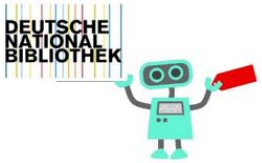
DEUTSCHE NATIONAL BIBLIOTHEK

Testkorpus: test_title-fulltext_gnd_1 (434 Datensätze NPs)



Projektphase 2.1A: Entscheidung

- Annif erfüllt nicht nur die fachlichen Kernfunktionalitäten, sondern ist auch aus technischer Sicht grundsätzlich für einen produktiven Betrieb in der DNB geeignet
- Annif ist für einen Start zum Aufbau der modularen Erschließungsmaschine geeignet und bietet eine Aussicht auf ein Erreichen der Projektziele einer "kontinuierlichen Verbesserung der Erschließungsergebnisse" sowie einer zeitnahen „Ablöse des Averbis-Systems“



DEUTSCHE NATIONAL BIBLIOTHEK

EMa – Erschließungsmaschine

Danke für Ihre Aufmerksamkeit.

s.uhlmann@dnb.de