Das Toolkit

annif

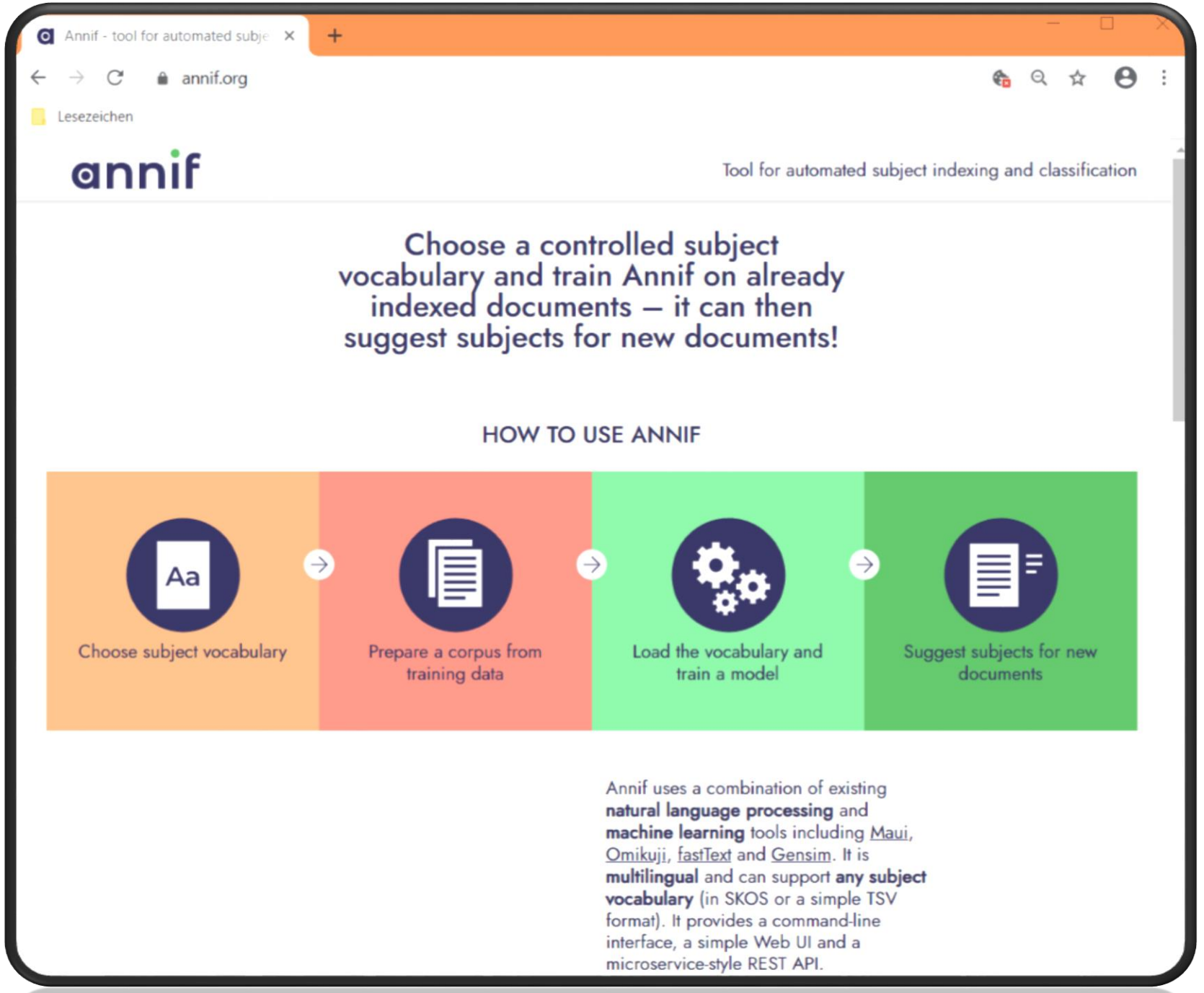
Claudia Grote | Deutsche Nationalbibliothek | 04.12.2020

annif

DEUTSCHE NATIONAL BIBLIOTHEK

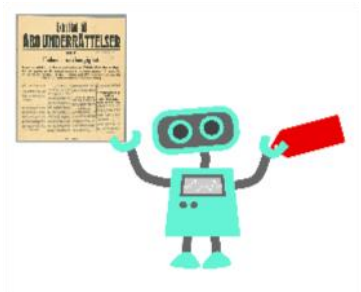
- Überblick
- Anwendung
- Aufbau
- Textvorverarbeitung
- Verfahren
- Web-Kommunikation
- Weiterführendes

Annif-Website



Quelle: https://annif.org

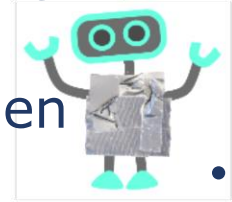
annif



DEUTSCHE NATIONAL BIBLIOTHEK

Software zur automatischen Schlagwort-Indexierung und Klassifikation

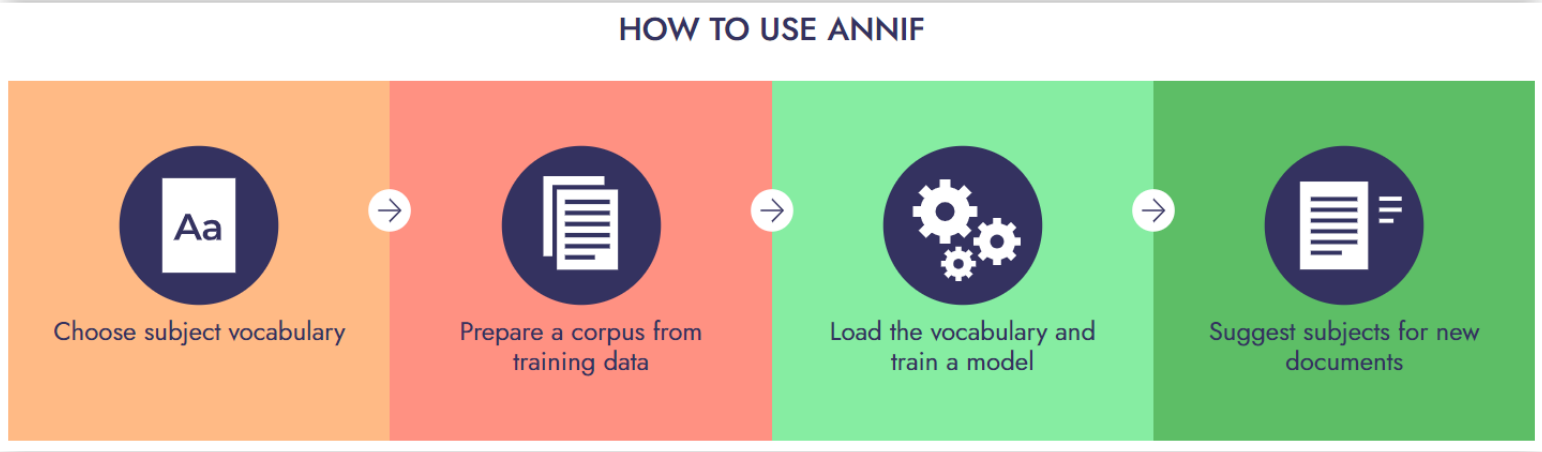
- Open-Source-Projekt (Github)
- an der Nationalbibliothek von Finnland entstanden
- Hauptentwickler: Osma Suominen



2017 Prototyp von Osma Suominen
2018 Neuentwicklung auf solider Grundlage
+ Vorstellung auf der SWIB18
2019 Tutorial auf der SWIB19
Juli 2020 Version 0.49
weitere Tutorials und Workshops
produktive Services basieren auf Annif

- verwendet existierende Verfahren
  - zur Verarbeitung natürlicher Sprache
  - zum maschinellen Lernen
- ist multilingual
  - Einsatz des Natural Language Toolkit, NLTK
- kann jedes Schlagwort-Vokabular verwenden
  - in SKOS oder einfachem TSV
- ist über Kommandozeile, Web UI und Rest API bedienbar
- ist in Python implementiert
- ist modular erweiterbar

Bilderquelle: https://annif.org



Zum praktischen Einstieg bei Annif:
sehr gut strukturiertes Tutorial auf der Annif-Github-Seite

Link zu https://github.com/NatLibFi/Annif-tutorial

Quelle: https://annif.org

1. Schritt: Auswahl eines Vokabulars für annif

DEUTSCHE NATIONAL BIBLIOTHEK

tsv

```
<http://d-nb.info/gnd/040254372>        Hörstörung
<http://d-nb.info/gnd/990692604>        Önologe
<http://d-nb.info/gnd/041526775>        Erdmännchen
<http://d-nb.info/gnd/943860660>        Exordium
<http://d-nb.info/gnd/96914167X>        Spornammer
<http://d-nb.info/gnd/040592529>        Technischer Fortschritt
<http://d-nb.info/gnd/041869028>        Unipolarmaschine
```

GND

```
<http://d-nb.info/ddc/700>      700
<http://d-nb.info/ddc/710>      710
<http://d-nb.info/ddc/720>      720
<http://d-nb.info/ddc/730>      730
<http://d-nb.info/ddc/740>      740
<http://d-nb.info/ddc/741.5>    741.5
<http://d-nb.info/ddc/750>      750
```

DNB-Sachgruppen

SKOS/RDF

```
    skos:prefLabel "Erdkröte"@de .

<http://d-nb.info/gnd/041526775> a skos:Concept ;
    skos:altLabel "Erdhündchen"@de,
        "Scharrtier"@de,
        "Suricata suricatta"@de,
        "Surikate"@de ;
    skos:prefLabel "Erdmännchen"@de .

<http://d-nb.info/gnd/041526783> a skos:Concept ;
    skos:altLabel "Erdmagnetisches Feld Pulsation"
```

GND in SKOS als ttl-Datei

2. Schritt: Trainings-/Testkorpus für annif

Textdateien
(Volltexte oder TOC-Texte)

Korea 151
ein Land
zwischen K-
Pop und
Kimchi in 151
Momentauf-
nahmen
Geschichte

1175430633.txt

Datei entweder mit GND-Schlagwörtern, DNB-Sachgruppe oder DDC-Kurznotation (aus der intellektuellen Erschließung), je mit URI

<http://d-nb.info/gnd/040324664> Korea
<http://d-nb.info/gnd/040739724> Landeskunde

bzw.

<http://d-nb.info/ddc/910> 910

bzw.

<http://d-nb.info/ddc/915.19> 915.19

1175430633.tsv

1175430633.tsv	1176212354.tsv	1177163918.tsv	1178451437.tsv
1175430633.txt	1176212354.txt	1177163918.txt	1178451437.txt
1175431508.tsv	1176212621.tsv	1177164086.tsv	1178451674.tsv
1175431508.txt	1176212621.txt	1177164086.txt	1178451674.txt
1175431702.tsv	117621263X.tsv	1177164507.tsv	1178451887.tsv
1175431702.txt	117621263X.txt	1177164507.txt	1178451887.txt

im Dateisystem: zusammengehörige txt- und tsv-Dateien haben den selben Dateinamen

3.+4. Schritt: Training und Test/Evaluation mit annif

DEUTSCHE NATIONAL BIBLIOTHEK

Konfiguration eines Annif-Projekts (projects.cfg)

z.B. für ein einfaches Verfahren (tfidf):

```
[gnd0-tfidf-de]
name=gnd0-tfidf-de
language=de
backend=tfidf
analyzer=snowball(german)
limit=50
vocab=gnd0
```

Laden des Vokabulars und Trainieren eines Modells

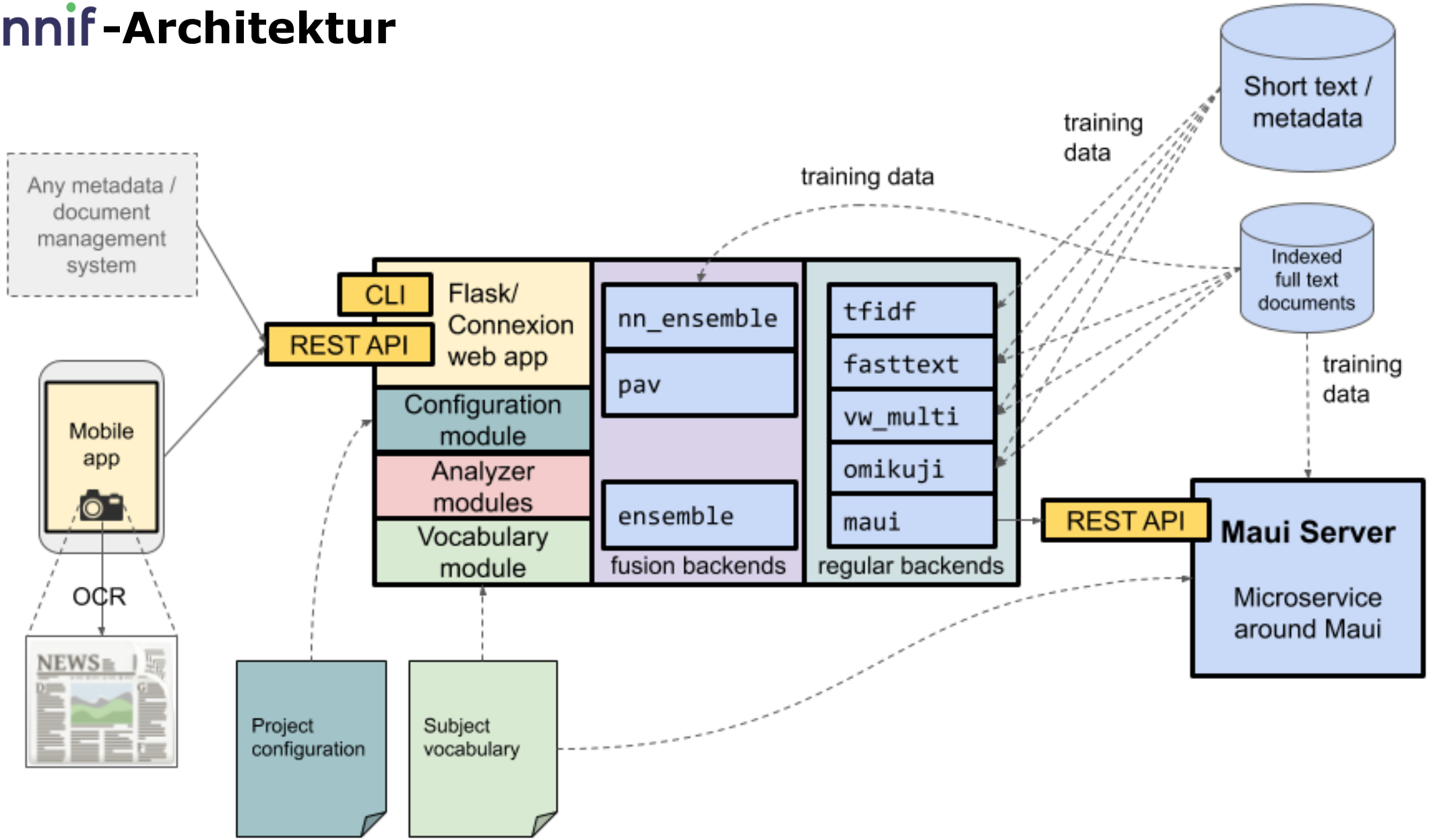
```
annif loadvoc gnd0-tfidf-de <Pfad>/gnd0.tsv
annif train gnd0-tfidf-de <Pfad_zum_Trainingskorpus>
```

Anfragen des trainierten Modells mit unbekannten Texten

`cat text.txt | annif suggest gnd0-tfidf-de` (Einzeltext)

`annif eval gnd0-tfidf-de <Pfad_zum_Testkorpus>` (Testkorpus)

annif-Architektur



Quelle: https://github.com/NatLibFi/Annif/wiki/Architecture

annif-Analyser-Module

Textvorverarbeitung/Textnormalisierung

- NLTK (natural language toolkit)
  - Tokeniser
  - Snowball-Stemmer
- *Simple Analyser* führt nur eine Kleinsetzung nach der NLTK-Tokenisierung durch
- Lemmatisierung nur für Finnisch (*Voikko Analyser*)

annif-Backends: lexikalische und assoziative Verfahren

DEUTSCHE NATIONAL BIBLIOTHEK

Lexikalische Ansätze

suchen Übereinstimmungen zwischen relevanten Termen im Text und Termen im kontrollierten Vokabular

„Die **Ästhetische Theorie** ist ein posthum erschienenes Werk des Philosophen **Theodor W. Adorno**."

⇨ <http://d-nb.info/gnd/4122759-1> Ästhetische Theorie

⇨ <http://d-nb.info/gnd/118500775> Adorno, Theodor W.

Probleme z.B.:

- semantische Disambiguierung bei Polysemen/Homonymen
- semantische Ähnlichkeit bei Synonymen, die im Vokabular fehlen
- fehlende Kandidaten zum Matching, z.B. fehlende „Match-Einstiege"
- falsch positives Matching bei fehlendem korrekten Begriff

annif-Backends: lexikalische und assoziative Verfahren

DEUTSCHE NATIONAL BIBLIOTHEK

Assoziative Ansätze

modellieren Korrelationen der Begriffe des kontrollierten Vokabulars (bzw. der Klassen eines Klassifikationsschemas) mit Termen im Text auf der Basis von Beispielen (Trainingsdaten)

Probleme z.B.:

- semantische Ähnlichkeit von Synonymen
- Modellüberanpassung
- Modellübergeneralisierung
- fehlende Trainingsdaten
- Modellanpassung bei Änderungen am Vokabular

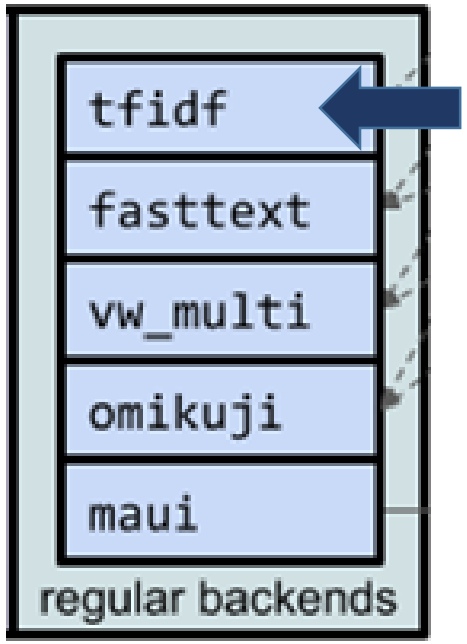
annif-Backends: Einzelverfahren und Ensemble-Verfahren

Die Verfahren in den modularen Backends können einzeln eingesetzt oder kombiniert werden.

Durch Kombination von Verfahren können Schwächen der Einzelverfahren ausgeglichen werden.

Insbesondere können sich lexikalische und assoziative Verfahren ergänzen.

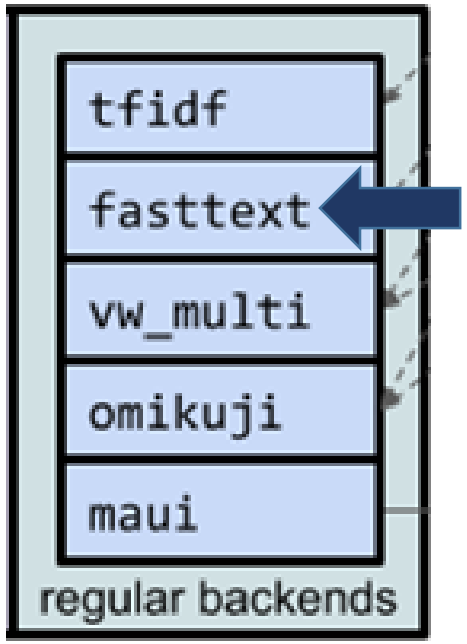
annif-Backends: Die 5 Einzelverfahren



TF-IDF

- Bag-of-words-Verfahren (Reihenfolge der Wörter wird ignoriert)
- nutzt Topic-Modelling-Bibliothek Gensim
- modelliert ein Schlagwort mittels relevanter Terme in den damit beschlagworteten Texten
- Relevanz eines Terms in Bezug auf ein Schlagwort ergibt sich durch seine Vorkommenshäufigkeit in den Texten mit diesem Schlagwort kombiniert mit dem inversen Anteil (wird größer, je kleiner der Anteil der Dokumente ist) an Dokumenten, die diesen Term enthalten
- Vorhersage durch das Modell geschieht durch Abgleich der Termhäufigkeiten mit den gelernten Relevanzgewichten
- Baseline-Verfahren

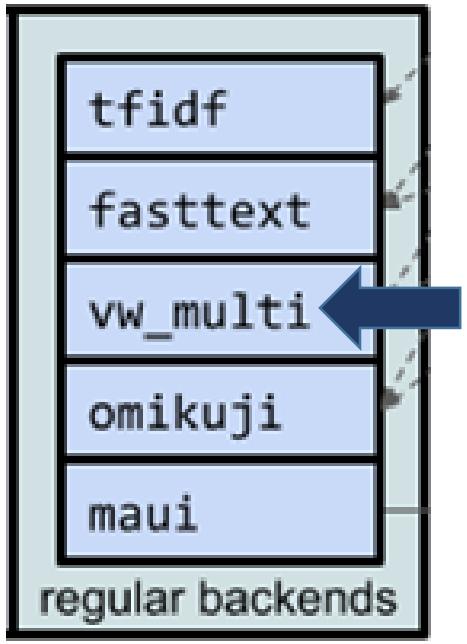
annif-Backends: Die 5 Einzelverfahren



fastText

- Klassifikationsverfahren von Facebook AI Research
- nutzt vektorbasiertes Zeichen- bzw. Wort-Embeddings
- Wörter werden mit ihrem direkten Wort-Kontext repräsentiert
- trainiert ein Klassifikationsmodell als vorwärtsgerichtetes Neuronales Netzwerk
- ressourcenteure Berechnungen
- profitiert von paralleler Verarbeitung

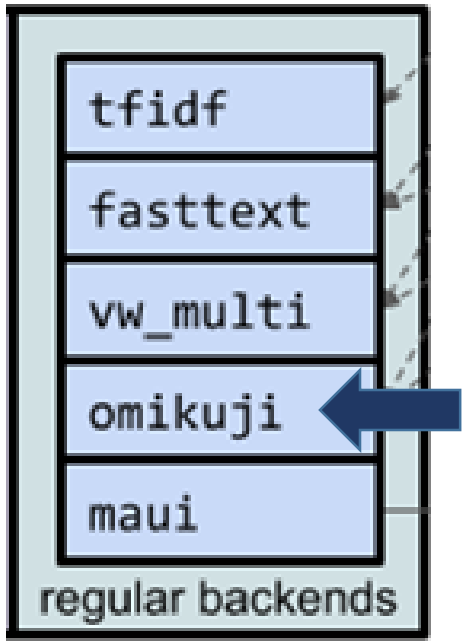
annif-Backends: Die 5 Einzelverfahren



vw_multi

- „Vowpal Wabbit”-Machine-Learning-System
- entwickelt von Yahoo! (jetzt Microsoft Research)
- multiclass- und multilabel-Klassifikation
- viele Algorithmen zur Modellanpassung (Fehlerminimierung)
- viel Tuning
- empfohlen für weniger als 1000 Klassen
- unterstützt Online-Lernen (Weiterlernen nach Training durch Feedback)

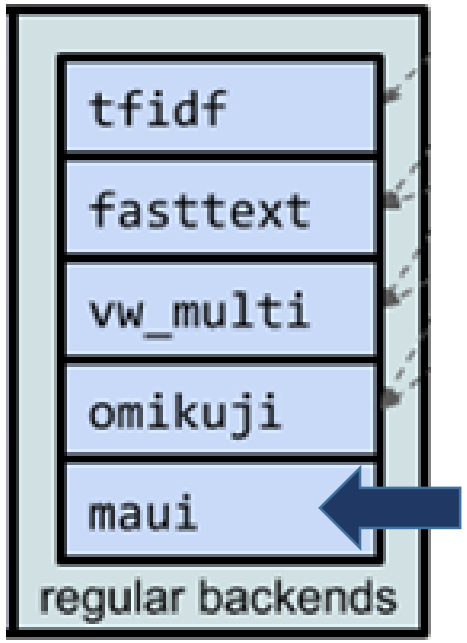
annif-Backends: Die 5 Einzelverfahren



Omikuji 御神籤

- effiziente Implementierung baumbasierter Algorithmen für *extreme multi-label*-Klassifikation
- implementiert die Algorithmen *Parabel* und *Bonsai*
- zeigt mit Default-Konfiguration schon sehr gute Ergebnisse
- Training sehr ressourcenteuer
- profitiert stark von paralleler Verarbeitung

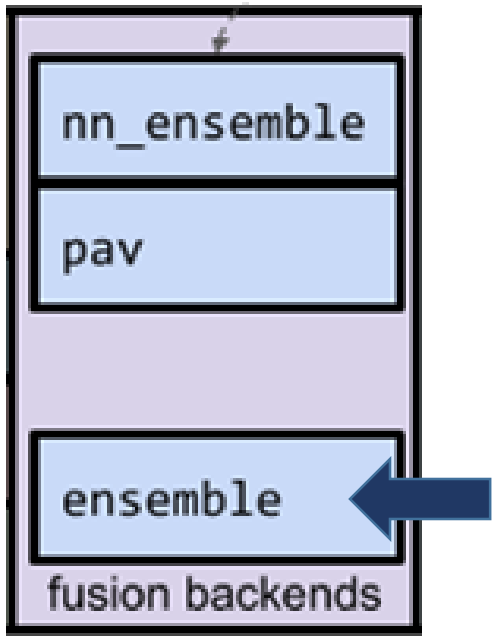
annif-Backends: Die 5 Einzelverfahren



Maui

- lexikalisches Verfahren, gleicht relevante Terme im Text mit Termen im kontrolliertem Vokabular ab
- lernende Komponente modelliert, welche Terme aus einem Text als Schlagwörter genutzt werden und welche nicht
- basiert auf KEA (*key extraction algorithm*) zum Topic Modelling
- durch KEA extrahierte Token-n-Gramme werden v.a. auf der Basis von Textstatistik und taxonomischer Kohärenz gerankt
- verwendet Vokabular in SKOS/RDF
- seit 2015 keine Weiterentwicklung der Software

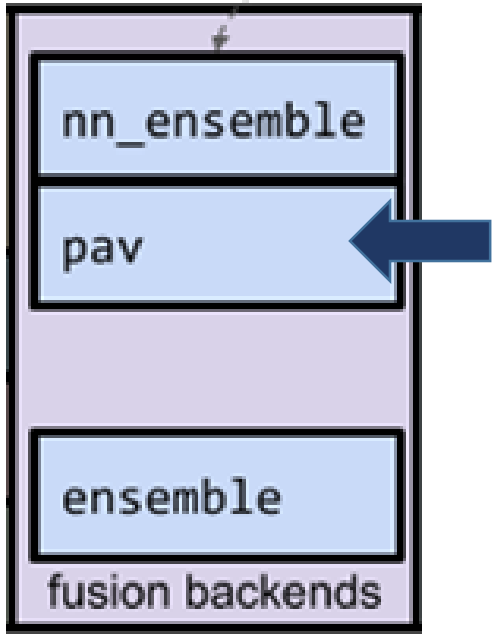
annif-Backends: Die 3 Ensemble-Verfahren



Ensemble

- einfaches Ensemble-Verfahren
- konfigurierbare Gewichtung der Ergebnisse der zu kombinierenden Einzelverfahren
- Mittelung der Werte für jedes Ergebnis

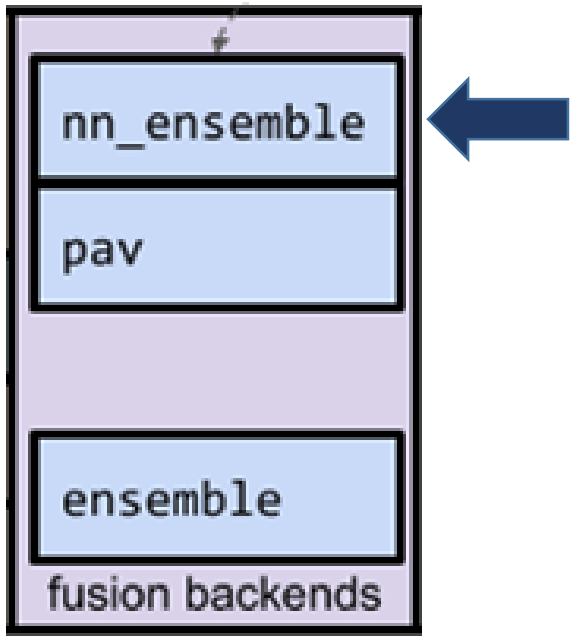
annif-Backends: Die 3 Ensemble-Verfahren



PAV

- trainierbares dynamisches Ensemble-Verfahren
- Verwendung des PAV-(*pool-adjacent-violaters*-) Algorithmus aus der Python-Bibliothek scikit-learn
- konfigurierbare Gewichtung der Ergebnisse der Einzelverfahren
- Neugewichtung durch ein trainiertes Regressionsmodell

annif-Backends: Die 3 Ensemble-Verfahren



nn_ensemble

- trainierbares dynamisches Ensemble-Verfahren
- Verwendung eines Neuronalen Netzwerks mittels Keras und Tensorflow 2
- Neugewichtung der Ergebnisse durch das trainierte Neuronale Netzwerk
- unterstützt Online-Lernen nach dem Training durch Feedback

Web UI (eval)

annif

Web UI

DEUTSCHE NATIONAL BIBLIOTHEK

Welcome!

See the Swagger documentation for an interactive REST API specification.

Die Ästhetische Theorie ist ein posthum erschienenes Werk des Philosophen und Soziologen Theodor W. Adorno. Sie enthält Adornos Philosophie der Kunst als eine gattungsübergreifende Theorie der künstlerischen Moderne mit den Leitmotiven der Negativität und der Versöhnung sowie den ästhetischen Grundkategorien des Schönen und des Erhabenen. Als seine letzte große Arbeit zählt sie zu seinen philosophischen Hauptwerken. Obwohl als Torso 1970 aus dem Nachlass herausgegeben, stellt sie eine Summa seiner ästhetischen Überlegungen und Einsichten dar.

Adorno schöpft in der Ästhetischen Theorie aus seiner lebenslangen – auch als Komponist aktiven – Beschäftigung mit der Kunst und den Künsten. Ausgehend von den Besonderheiten moderner Kunst entfaltet Adorno eine umfassende kategoriale Analyse der Kunst, ihres nicht-diskursiven Wahrheitsgehaltes bei gleichzeitigem Rätselcharakter und ihres utopischen Kerns: der Versöhnung von Allgemeinem und Besonderem, von Natur und Geist, von Mimesis und Konstruktion. Er versteht Kunst als die „gesellschaftliche Antithesis zur Gesellschaft“ (ÄT 19)[1] und „Statthalter einer besseren Praxis“ (ÄT 26).

Inhaltsverzeichnis
Stellenwert im Werk Adornos und formale Struktur
Dem Germanisten Gerhard Kaiser zufolge werden in der Ästhetischen Theorie alle Motive von Adornos Denken „enggeführt“.[2] Für Günter Figal ist die Ästhetische Theorie als Hauptwerk und philosophisches Vermächtnis Adornos anzusehen. Konsequenter als in seinen anderen Schriften setze Adorno hier „seine Leitbegriffe als eine Vielzahl von Zentren ein, um die sich seine Reflexionen bilden“, und die in der Konstellation zueinander ein Ganzes ergäben.[3]

PROJECT (VOCABULARY AND LANGUAGE)

gnd0-omikuji-bonsai-de-49-1

MAX # OF SUGGESTIONS

10 15 20

Get suggestions

SUGGESTED SUBJECTS

- Theodor W. Adorno
- Ästhetik
- Kritische Theorie
- Theodor W. Adorno Ästhetische Theorie
- Werk
- Ästhetische Wahrnehmung
- Künste
- Subjektivität
- Utopie
- Schreiben nach Auschwitz

annif-Rest-API

DEUTSCHE NATIONAL BIBLIOTHEK

```
curl -X POST --header 'Content-Type: application/x-www-form-urlencoded' --header
'Accept: application/problem+json,
-d
'text=Die+Ästhetische+Theorie+ist+ein+posthum+erschienenes+Werk+des+Philosophen
+und+Soziologen+Theodor+W.+Adorno.&limit=5'
'http://localhost:5000/v1/projects/gnd0-omikuji-bonsai-de-49-1/suggest'
{
  "results": [
    {"label": "Ästhetik", "score": 0.4656588731962585, "uri": http://d-
nb.info/gnd/040006263},
    {"label": "Theodor W. Adorno", "score": 0.3091989943504333, "uri":
http://d-nb.info/gnd/118500775},
    {"label": "Philosophie", "score": 0.0702155306935310 4, "uri": http://d-
nb.info/gnd/040457915},
    {"label": "Kulturtheorie", "score": 0.045728418976068 5, "uri": http://d-
nb.info/gnd/041206274},
    {"label": "Theodor W. Adorno Ästhetische Theorie", "score":
0.032760385423898 7, "uri": http://d-nb.info/gnd/04122759X}}]
}
```

Weiterführendes

- Annif-Hauptseite mit Demo: https://annif.org
- Github-Projekt: https://github.com/NatLibFi/Annif
- Installationsanleitung: https://github.com/NatLibFi/Annif/#basic-install
- Annif-Tutorial: https://github.com/NatLibFi/Annif-tutorial
- Annif-Wiki: https://github.com/NatLibFi/Annif/wiki/Getting-started
- Annif-Forum: https://groups.google.com/forum/#!forum/annif-users
- API-Dokumentation: https://readthedocs.org/projects/annif
- Suominen, O., 2019. Annif: DIY automated subject indexing using multiple algorithms. LIBER Quarterly, 29(1), pp.1–25. DOI: https://doi.org/10.18352/lq.10285
- Toepfer, M., & Seifert, C. (2018). Fusion architectures for automatic subject indexing under concept drift. International Journal on Digital Libraries, 1–21. DOI: https://doi.org/10.1007/s00799-018-0240-3. Preprint: https://research.utwente.nl/files/80439235/Toepfer2018_ijdl_subject_indexing_under_concept_drift_preprint.pdf.

EMa – Erschließungsmaschine

Danke für die Aufmerksamkeit.

c.grote@dnb.de