Werkstattbericht zur Nutzung von Annif an der ZBW

Moritz Fürneisen, Christopher Bartz, Anna Kasprzik
ZBW – Leibniz-Informationszentrum Wirtschaft
DNB Annif Workshop, 3./4. Dezember 2020

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Ausgangslage

- Vorhanden
  - Legacy Code zur Verschlagwortung größerer Datenmengen

- Ziele
  - automatisiert generierte Konzeptvorschläge an Referent_innen ausliefern
  - Publikationen, die nicht intellektuell verschlagwortet werden können, automatisiert verarbeiten

Auswahl Technologie: Indexierung

- Bestehende Software
  - auf die Daten der ZBW optimiert
  - Codebasis auf Forschung ausgelegt, ungeeignet für Produktiveinsatz

- Alternative Annif
  - klar strukturierter und getester Code
  - REST-Schnittstelle für eine einfache Integration
  - erste Experimente zeigten vergleichbare F_1-Werte
  - aktive und offene Entwicklung auf Github
  - schnelle Antworten über Mailingliste

Metadatenfluss Indexierung

- SOLR-Index erlaubt es Änderungen abzufragen
- verarbeitet werden geänderte Metadatensätze ohne Konzepte
- Ablage von automatisiert generierten Vorschlägen in Key-Value-Store

https://de.wikipedia.org/wiki/Schl%C3%BCssel-Werte-Datenbank

Anforderungen an Frontend

- Konzeptvorschläge anzeigen
- automatisiert erstellte Indexate übernehmen
- neue Konzepte hinzufügen
- gute Verzahnung mit intellektueller Erschließung:
  - Endresultat im GBV bzw. K10plus nutzbar

Digitaler Assistent 3 (DA3)

- befand sich 2020 in Testphase an der ZBW
  - Export in Verbund bereits geklärt

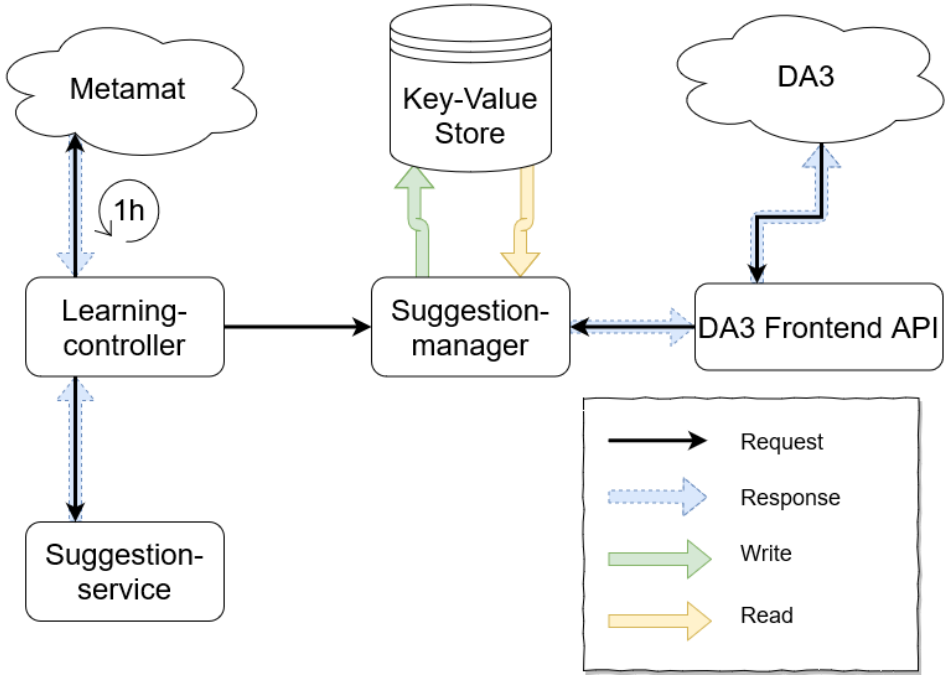
- Tool zur Unterstützung der Indexierung durch die Anzeige von Vorschlägen
- Entwicklung durch externen Dienstleister Eurospider
- nutzt Indexate aus anderen Institutionen
  - zieht u.a. auch Konkordanzen heran
- Eurospider bindet unsere automatisiert erzeugten Vorschläge ein

Datenfluss DA3

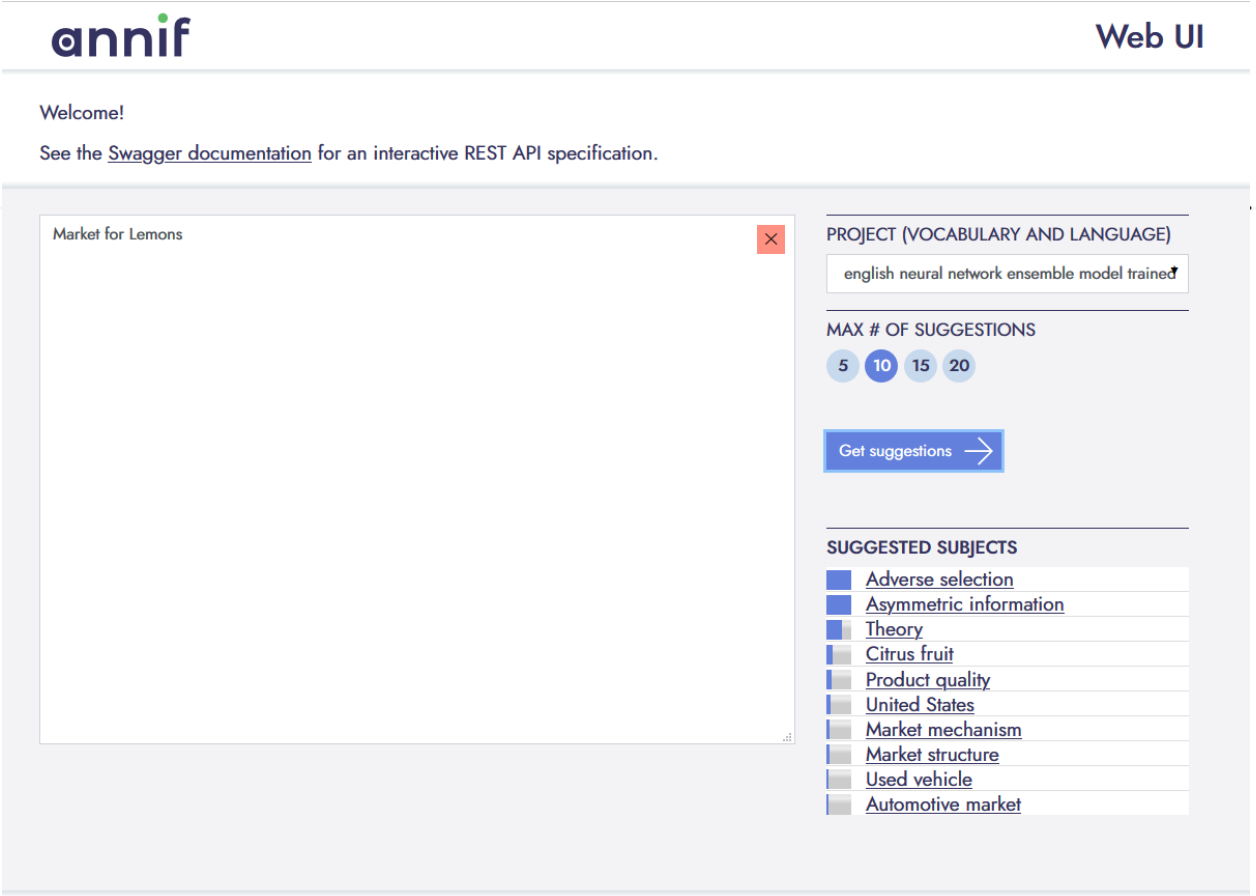
- allgemein:
  - Publikation wird in DA3-Frontend aufgerufen
  - DA3-Backend fragt verschiedene Quellen für Indexate an
    - nutzt eindeutigen Identifier (bei uns: PPN)
  - Quellen liefern Indexate in MARC XML
  - Konzeptvorschläge werden in DA3-Frontend angezeigt

- an der ZBW
  - automatisiert generiertes Indexat wird aus KV-Store gelesen

Gesamtdatenfluss



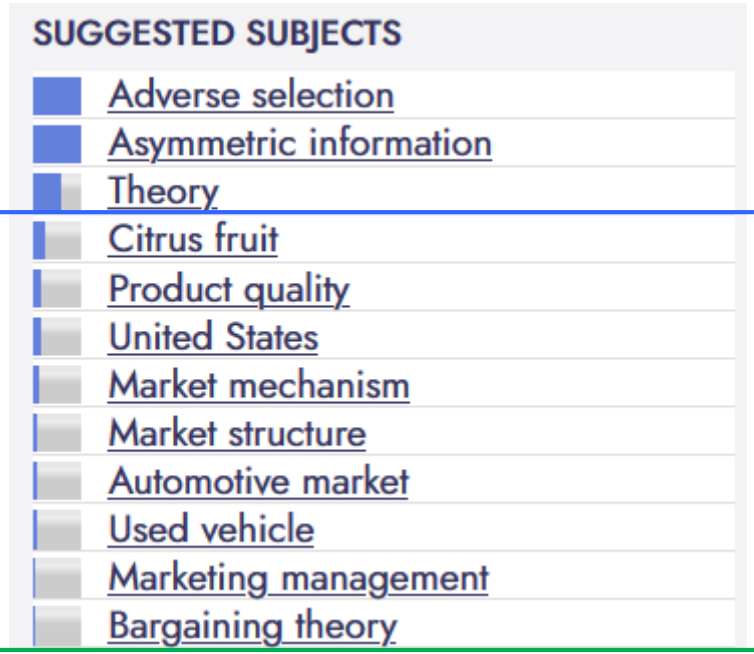
WEB UI


annif
Web UI
Welcome!
See the Swagger documentation for an interactive REST API specification.
Market for Lemons
PROJECT (VOCABULARY AND LANGUAGE)
english neural network ensemble model trained
MAX # OF SUGGESTIONS
5
10
15
20
Get suggestions
SUGGESTED SUBJECTS
Adverse selection
Asymmetric information
Theory
Citrus fruit
Product quality
United States
Market mechanism
Market structure
Used vehicle
Automotive market

Schwellwert

höhere Precision

niedriger Recall



niedrigere Precision

höherer Recall

Suchen nach Schwellwert

- Problem: F_1-Metrik abhängig von Schwellwert
- Suchstrategien für Schwellwert
  - Rastersuche (0,01, 0,02, 0,03, …, 0,98, 0,99)
  - Zufall
- Kandidaten für Schwellwert werden von *hyperopt* erzeugt
  - kümmert sich um Exploration-Exploitation-Trade-Off

https://github.com/hyperopt/hyperopt
https://en.wikipedia.org/wiki/Multi-armed_bandit

Problem: Backendkonfiguration

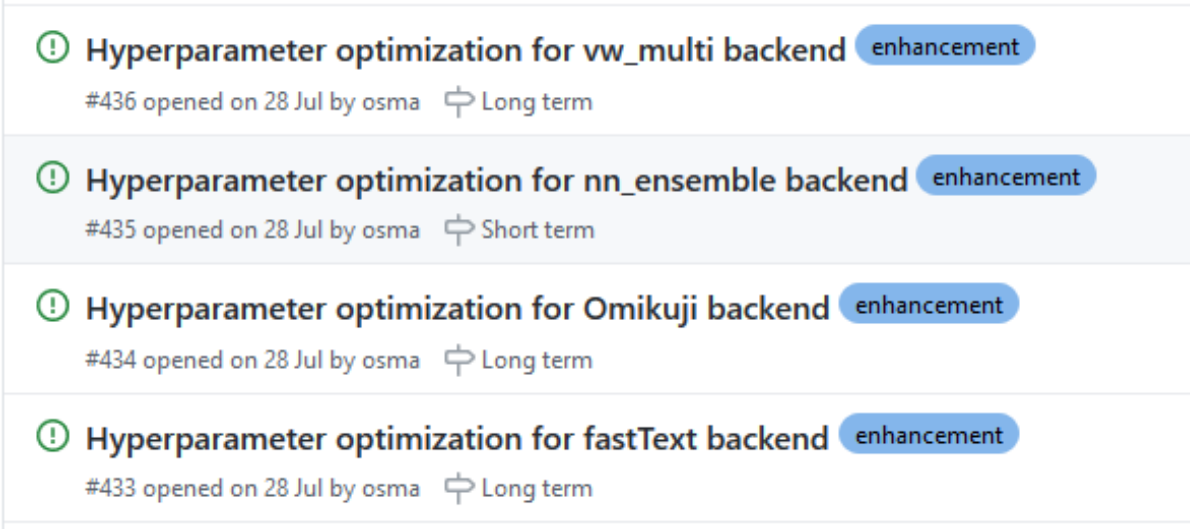
- Verfahren haben verschiedene Parameter
- Konfiguration kann signifikanten Einfluss auf Ergebnisse haben

- manuelle Bestimmung von Parametern erfordert viel Erfahrungswissen
- daher: Optimierung durch automatisierte Suche nach guten Parametern
    1. Parameter erzeugen
    2. Modell trainieren
    3. Modell bewerten

Verfahren Parametersuche

- Verwendung von Annif als Library
  - gleiche API für unterschiedliche Verfahren
  - Umwandlung in geeignetes Datenformat durch Annif
  - Kenntnisse interner Annifstrukturen notwendig

Parameteroptimierung Ausblick



https://github.com/NatLibFi/Annif/issues

Schwellwert ein Modellparameter?

- Problem:
  - Durch ändern des Schwellwerts, können sich Vorschläge drastisch ändern
    - Extrembeispiel: Alle Vorschläge werden verworfen => F1 = 0
  - Parameterkombination wird eventuell aufgrund unpassendem Schwellwert verworfen

Schwellwert nachträglich ermitteln

- mögliche Lösung:
  - bewerte Modelle mittels normalized Discounted Cumulative Gain(nDCG)
  - für Konfiguration mit bestem nDCG-Wert suche geeigneten Schwellwert
- mögliche Alternative:
  - für jede Konfiguration suche Schwellwert mit bestem F_1-Score

https://en.wikipedia.org/wiki/Discounted_cumulative_gain

Deployment von Modellen

- derzeit manuell:
  1. serialisiertes Modell kopieren
  2. Annifkonfiguration anpassen
  3. Annifinstanz neustarten
  4. Konfiguration Learning Controller anpassen
  5. Learning Controller neustarten

Pläne

- Visualisierung von Metrikwerten
- mehrere Modelle parallel betreiben
  - Modellauswahl anhand von Metrikwerten aus A/B-Tests
- Automatisierte Verschlagwortung für Discoverysystem nutzbar machen
- Parametersuche über Web-Interface
- Automatisiertes Deployment von Modellen