Fachhochschule Nordwestschweiz
Hochschule für Architektur, Bau und Geomatik

Metadaten Mining von Mobilitätsdaten für MODI

Ausgangslage nationale Mobilitätsdateninfrastruktur (MODI)

- Entscheid **Bundesrat** zur Erstellung einer nationalen MODI (BAV 2025).
- **Ziel**: Aufbau einer zentralen Metadatenbank, keine Speicherung der eigentlichen Daten (BAV 2025).
- **Auftrag** vom Bundesamt für Verkehr (BAV) zur Erstellung einer Metadatensammlung der relevanten öffentlich verfügbaren Daten.
- **Fragestellung:**
  Wie können die in der Schweiz öffentlich zugänglichen **Metadaten** zu Mobilität und Verkehr **harmonisiert** und **kategorisiert** werden, um daraus eine aktuelle **Übersicht** über die Metadaten und ihre **Qualität** für MODI zu gewinnen?

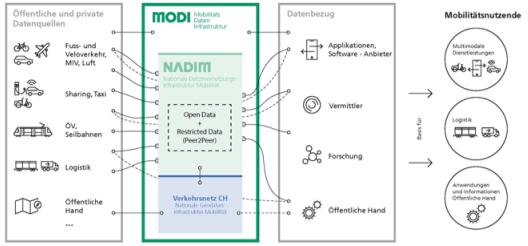
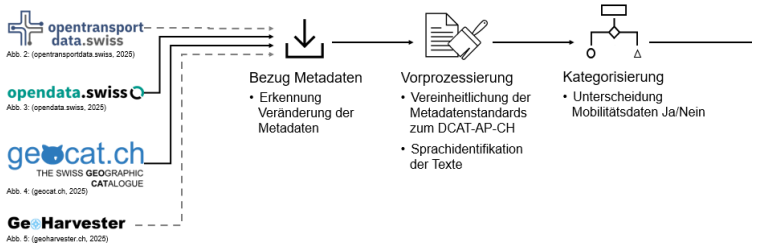
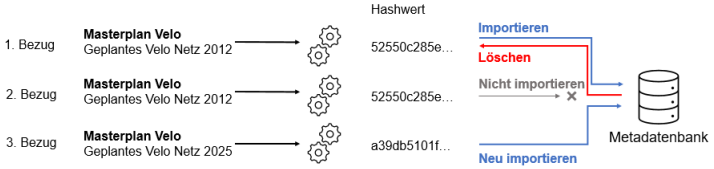


Abb. 1: Die Mobilitätsdateninfrastruktur (MODI) besteht aus zwei Teilinfrastrukturen: Das Verkehrsnetz CH bildet das Schweizer Verkehrssystem einheitlich ab; NADIM ermöglicht die Bereitstellung und den Austausch von Daten zu Mobilitätsangeboten und Infrastruktur. (BAV 2025)

Ablauf Verarbeitung der Metadaten

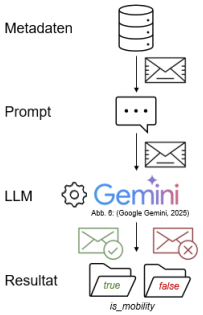


Änderungsdetektion an einem Beispiel gezeigt



Vorteil durch die Änderungsdetektion:
- Verkürzt Dauer von nachfolgenden Prozessen durch Entlastung

Kategorisierung Mobilitätsdaten mit Large Language Model



Kombinieren der Titel und Beschreibungen

chunk_lines = ['dataset_title'] + ['dataset_description']

Integration in den Prompt

«Handelt es sich bei folgendem Inhalt um Verkehrs- oder Mobilitätsdaten?
Antworte nur mit T (True) oder F (False).
{chunk_lines}
Antwort:»

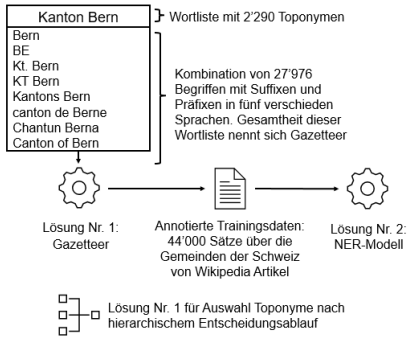
Parameter definieren und Anfrage an die API

temperature = 0 -> keine kreativen Antworten
max_output_tokens = 1 -> Anzahl Zeichen in der Antwort

Einteilung in Mobilitätsdaten und Nicht-Mobilitätsdaten

3639 Mobilitätsdatensätze von 25'055 Metadatensätze
95% korrekt erfasst gemäss Tests mit 150 gelabelten Metadatensätze

Detektion geografischer Abdeckung von Metadaten



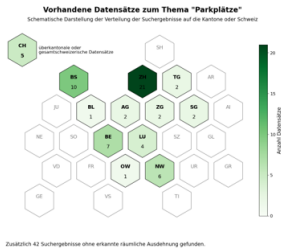


Abb. 7: Beispiel der geografischen Abdeckung vom Themengebiet Parkplätze. Toponym Extraktion mit Gazetteer und Ermitteln der Themenzugehörigkeit mit DBSCAN.

Resultat, Diskussion und Erkenntnisse

Resultat: Metadatenbank

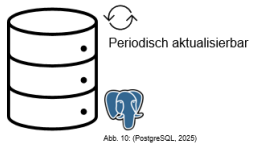


Abb. 10: (PostgreSQL, 2025)

- **25'055 Datensätze** davon sind **3'639 Mobilitätsdaten**
- **85'967** Beschreibungen in 5 Sprachen
- 219'724 Keywords
- PostgreSQL Datenbank

Erkenntnisse
- NLP ist der Way to go für dieses Aufgabe
- LLM sind gute Universal-Werkzeuge, die sich für unterschiedlichste Aufgaben eignen
- Nutzen eines gemeinsamen Metadatenstandards erleichtert Datenintegration

Diskussion
- Korrektur der Daten an Datenpublizierende durch Feedbackloop mitteilen
- Auffindbar sind nur Daten, die Open Government Data sind.

Ausgangslage nationale Mobilitätsdateninfrastruktur (MODI)

- Entscheid **Bundesrat** zur Erstellung einer nationalen MODI (BAV 2025).
- **Ziel**: Aufbau einer zentralen Metadatenbank, keine Speicherung der eigentlichen Daten (BAV 2025).
- **Auftrag** vom Bundesamt für Verkehr (BAV) zur Erstellung einer Metadatensammlung der relevanten öffentlich verfügbaren Daten.

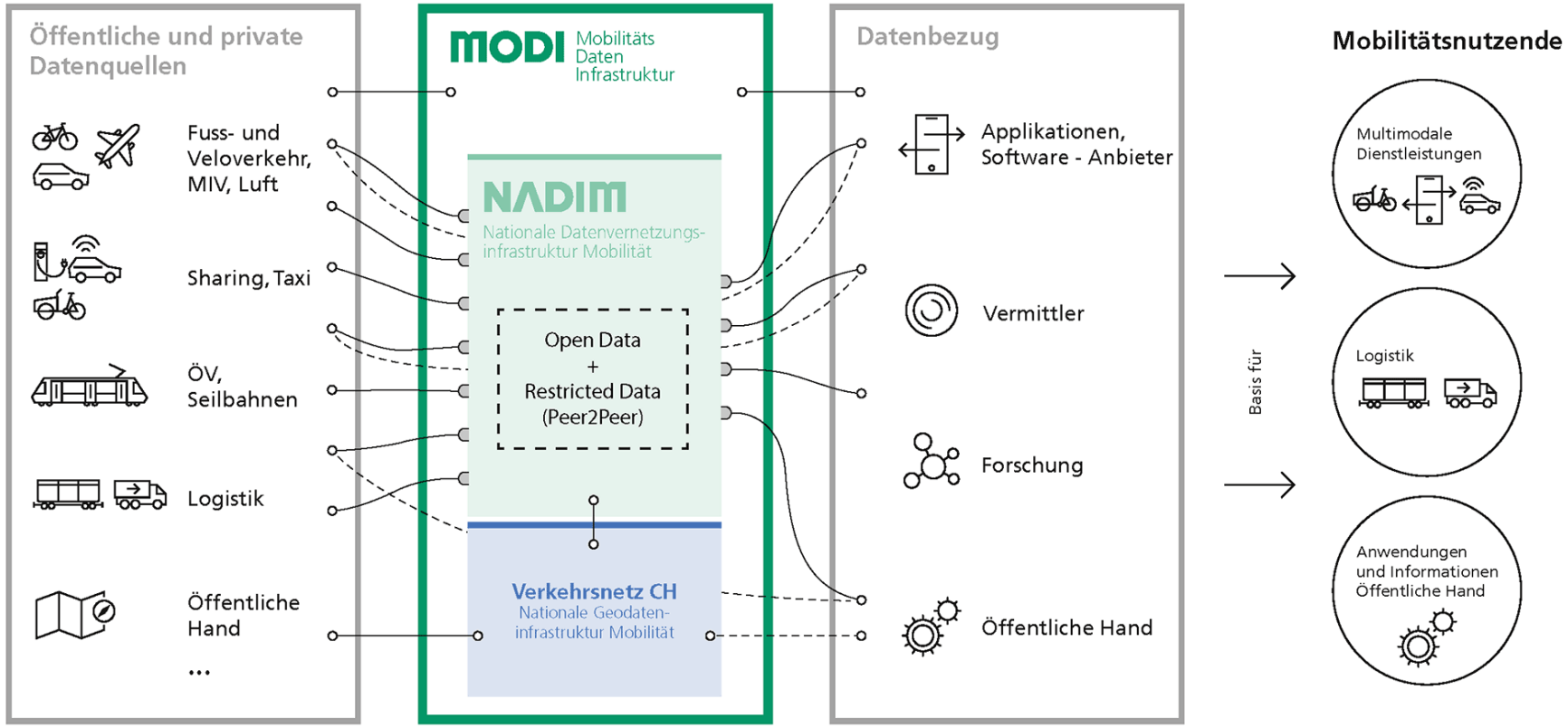


Abb. 1: Die Mobilitätsdateninfrastruktur (MODI) besteht aus zwei Teilinfrastrukturen: Das Verkehrsnetz CH bildet das Schweizer Verkehrssystem einheitlich ab; NADIM ermöglicht die Bereitstellung und den Austausch von Daten zu Mobilitätsangeboten und Infrastruktur. (BAV 2025)

- **Fragestellung:**
  Wie können die in der Schweiz öffentlich zugänglichen **Metadaten** zu Mobilität und Verkehr **harmonisiert** und **kategorisiert** werden, um daraus eine aktuelle **Übersicht** über die Metadaten und ihre **Qualität** für MODI zu gewinnen?

Autoren: Fabian Gross, Silas Haab, Examinator:in: Prof. Dr. Pia Bereuter, Friedrich Striewski, Betreuer: Stefan Eberlein, Experte: Lukas Zurbuchen (BAV)
© IGEO FHNW

Ablauf Verarbeitung der Metadaten

opentransport data.swiss

Abb. 2: (opentransportdata.swiss, 2025)

opendata.swiss

Abb. 3: (opendata.swiss, 2025)

geocat.ch
THE SWISS GEOGRAPHIC CATALOGUE

Abb. 4: (geocat.ch, 2025)

GeoHarvester

Abb. 5: (geoharvester.ch, 2025)

Bezug Metadaten

- Erkennung Veränderung der Metadaten

Vorprozessierung

- Vereinheitlichung der Metadatenstandards zum DCAT-AP-CH
- Sprachidentifikation der Texte

Kategorisierung

- Unterscheidung Mobilitätsdaten Ja/Nein

© IGEO FHNW

Ablauf Verarbeitung der Metadaten

Themencluster

- Detektion Themenzugehörigkeit
- Weiches Clustering mit LDA
- Hartes Clustering mit DBSACN

Geografische Extraktion

- Toponymen Extraktion aus Text
- Gazetteer
- NER-Modell
- Darstellung in Karte

Qualitätsbeurteilung

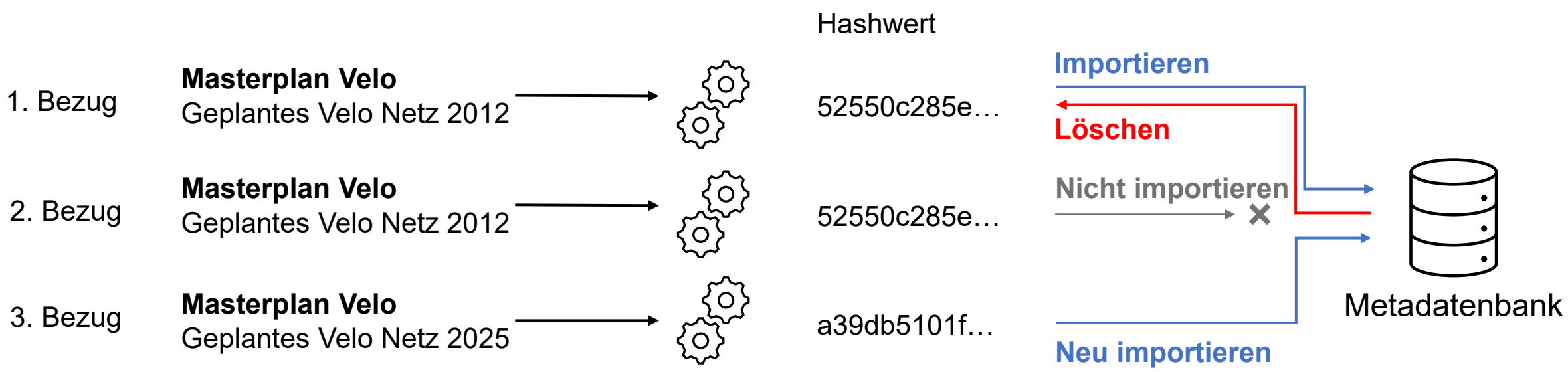
- Länge Beschreibung
- Anzahl Keywords
- Anzahl Formate
- Abrufbarkeit Download Links
- Sprachkonsistenz

Metadatenbank

- Aktualisierbar

© IGEO FHNW

Änderungsdetektion an einem Beispiel gezeigt



Vorteil durch die Änderungsdetektion:

- Verkürzt Dauer von nachfolgenden Prozessen durch Entlastung

© IGEO FHNW

Sprachidentifikation testen mit zwei Ansätzen

Die Python Module «Langid» und «Langdetect» erreichen beide eine Genauigkeit von über 99%. «Langid» erzielt eine bessere Genauigkeit bei kürzeren Texten.

Beschreibungen

- Langid erreicht mit **20 Zeichen** eine Genauigkeit von 95%
- Langdetect benötigt **35 Zeichen** für Genauigkeit von 95%
- Durch Einschränkung auf die im Text vorhandenen Sprachen verbessern sich die Ergebnisse beider Modelle.

Keywords

- Sprachdetektion nur von Keywords ist weniger genau.
- Bei nur einem Keyword liegt die Genauigkeit bei 83%
- Unzureichende Genauigkeit für die Verwendung

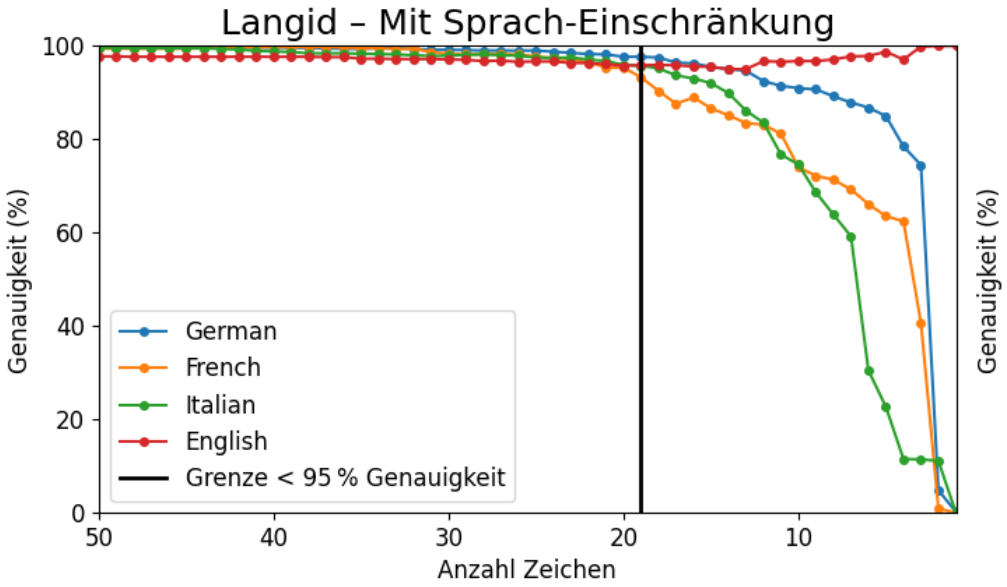


Abb. 8: Zeigt die Genauigkeit der Sprachidentifikation bei abnehmender Textlänge

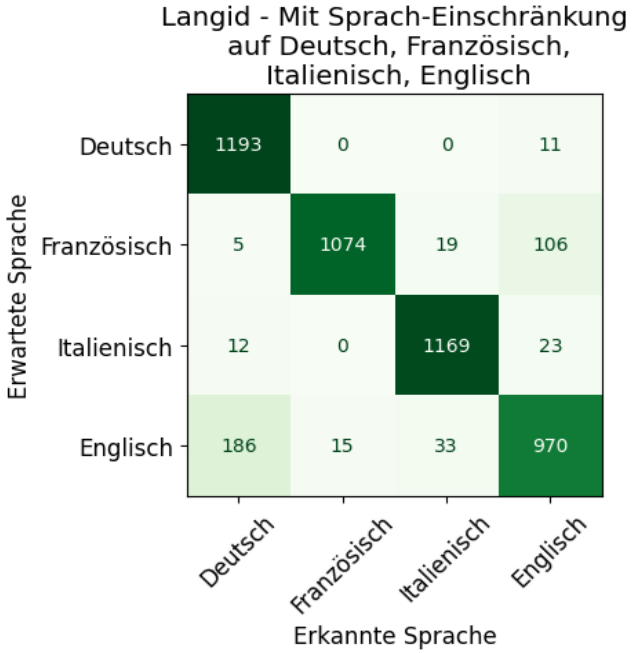


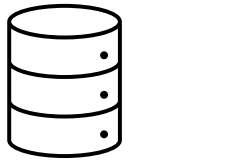
Abb. 9: Zeigt Konfusionsmatrix von der Sprachdetektion von Keywords

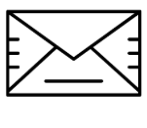
Autoren: Fabian Gross, Silas Haab, Examinator:in: Prof. Dr. Pia Bereuter, Friedrich Striewski, Betreuer: Stefan Eberlein, Experte: Lukas Zurbuchen (BAV)

© IGEO FHNW

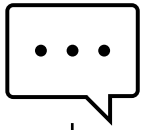
Kategorisierung Mobilitätsdaten mit Large Language Model

Metadaten





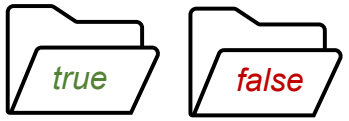
Prompt



LLM

Abb. 6: (Google Gemini, 2025)

Resultat



is_mobility

Kombinieren der Titel und Beschreibungen

chunk_lines = [‘dataset_title’] + [‘dataset_description’]

Integration in den Prompt

«Handelt es sich bei folgendem Inhalt um Verkehrs- oder Mobilitätsdaten?
Antworte nur mit T (True) oder F (False).
{chunk_lines}
Antwort:»

Parameter definieren und Anfrage an die API

temperature = 0 -> keine kreativen Antworten
max_output_tokens = 1 -> Anzahl Zeichen in der Antwort

Einteilung in Mobilitätsdaten und Nicht-Mobilitätsdaten

3639 Mobilitätsdatensätze von 25’055 Metadatensätze
95% korrekt erfasst gemäss Tests mit 150 gelabelten Metadatensätze

Autoren: Fabian Gross, Silas Haab, Examinator:in: Prof. Dr. Pia Bereuter, Friedrich Striewski, Betreuer: Stefan Eberlein, Experte: Lukas Zurbuchen (BAV)

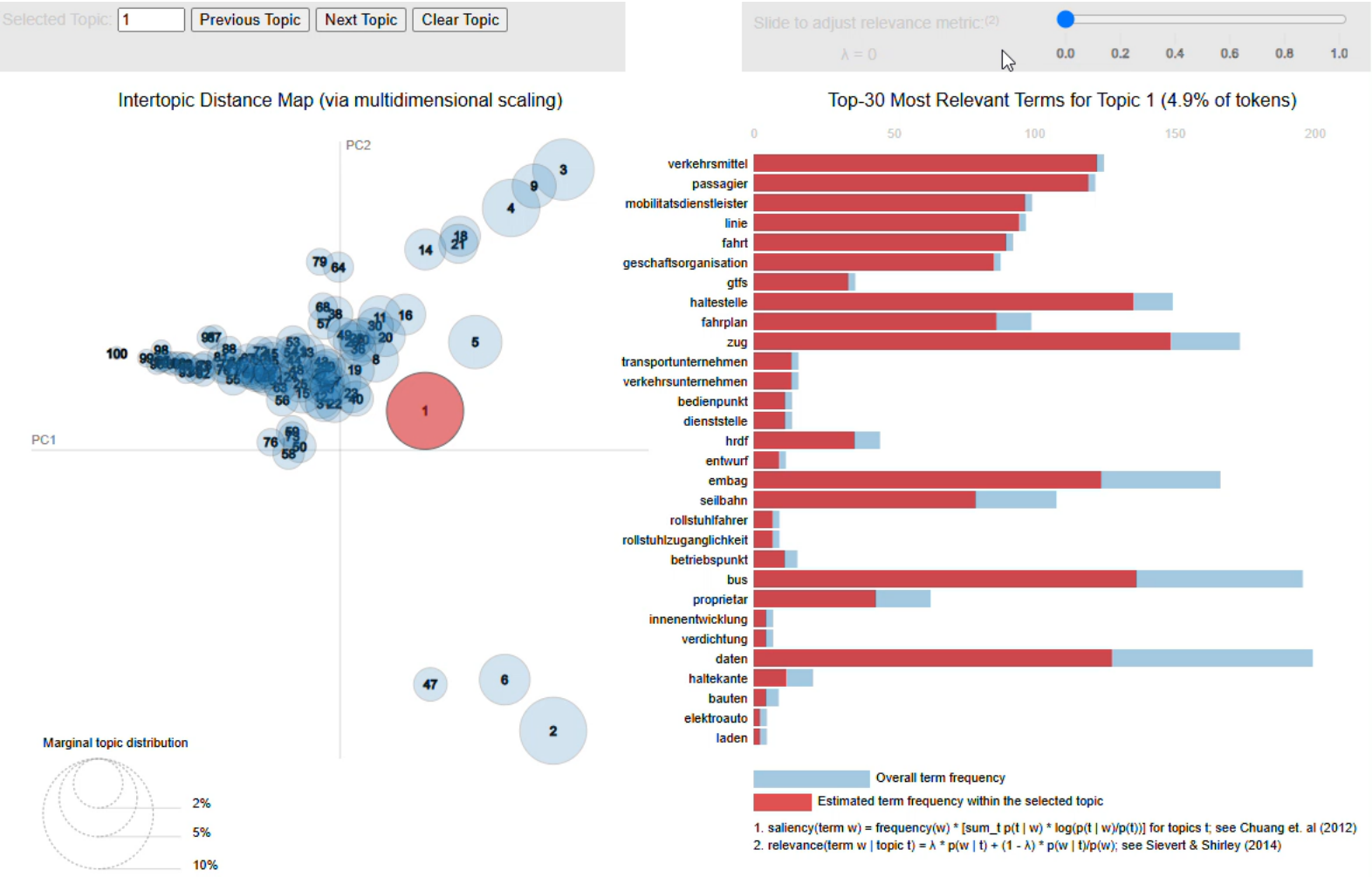
© IGEO FHNW

Themencluster – 2 Varianten

Ziel
- Finden von zusätzlichen Mobilitätsdatensätze in der Datenbank, die thematisch mit Suchbegriff übereinstimmen.
- Zuweisung des Suchbegriffs zu einem besten Cluster.

Weiches Clustering (LDA)
- Über semantische Ähnlichkeiten der Begriffe
- Zuweisung der Zugehörigkeit [%] zu **jedem** Thema

Hartes Clustering (DBSCAN)
- Über Wortvektoren der Keywords
- Reduktion der Vektoren auf 50 Dimensionen
- Zuweisung zu **einem** Thema



Video: Interaktive Erkundung des LDA-Clustering. Der Lambda-Slider (λ) oben rechts steuert die Gewichtung: Bei λ = 1 erscheinen die häufigsten Begriffe, bei kleineren Werten die themenspezifischeren.

© IGEO FHNW

Detektion geografischer Abdeckung von Metadaten

Kanton Bern
Bern BE Kt. Bern KT Bern Kantons Bern canton de Berne Chantun Berna Canton of Bern

Wortliste mit 2’290 Toponymen

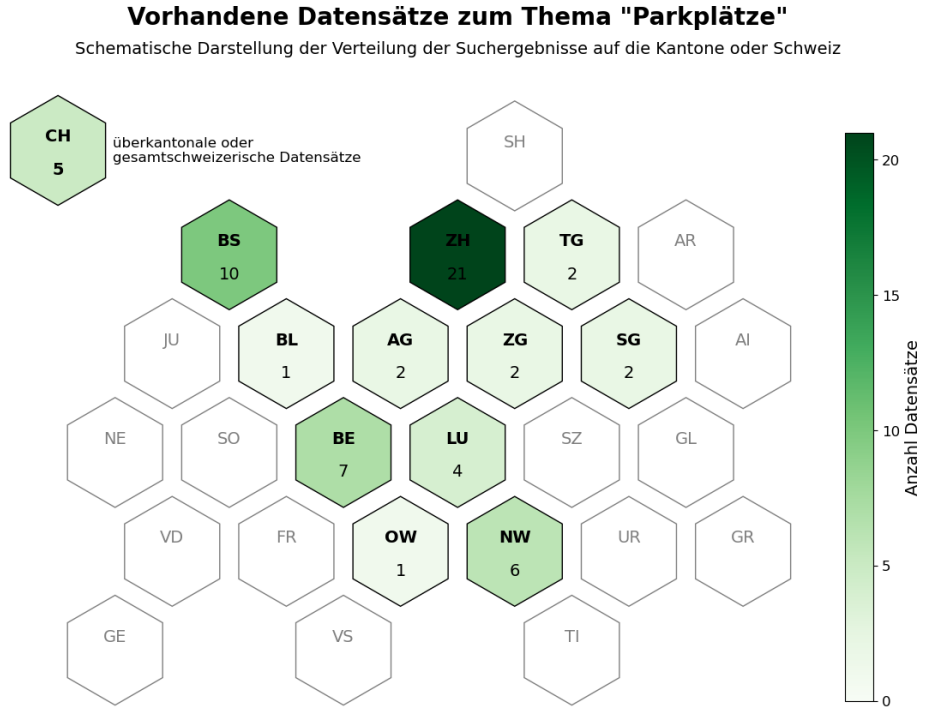
Kombination von 27’976 Begriffen mit Suffixen und Präfixen in fünf verschieden Sprachen. Gesamtheit dieser Wortliste nennt sich Gazetteer

Lösung Nr. 1: Gazetteer

Annotierte Trainingsdaten: 44’000 Sätze über die Gemeinden der Schweiz von Wikipedia Artikel

Lösung Nr. 2: NER-Modell

Lösung Nr. 1 für Auswahl Toponyme nach hierarchischem Entscheidungsablauf



Zusätzlich 42 Suchergebnisse ohne erkannte räumliche Ausdehnung gefunden.

Abb. 7: Beispiel der geografischen Abdeckung vom Themengebiet Parkplätze. Toponym Extraktion mit Gazetteer und Ermitteln der Themenzugehörigkeit mit DBSCAN.

Autoren: Fabian Gross, Silas Haab, Examinator:in: Prof. Dr. Pia Bereuter, Friedrich Striewski, Betreuer: Stefan Eberlein, Experte: Lukas Zurbuchen (BAV)

© IGEO FHNW

Qualitätsbeurteilung von Metadaten

Neue Erkenntnisse:

- Informationstiefe vergleichbar
- Anreiz schaffen für mehr Keywords
- In Kombination mit Clustering sind Trends von Datenformaten erkennbar

Fokus von data.europa.eu:
Beurteilung der Metadatenqualität aller Datensätze einer Plattform

Fokus von MODI (Vorschlag)
Beurteilung der Qualität der jedes einzelnen Metadatensatzes

- **43,7 %** Datensätze mit korrekter Sprachzuordnung.
- **29,3 %** verwenden identische Beschreibungen in mehreren Sprachen
- **241 Zeichen** lang ist der Median der Beschreibungslängen

Qualitätsmerkmale	**Beispiel Datensatz 1**	**Beispiel Datensatz 2**
Beschreibung vorhanden	Ja	Ja
Keywords vorhanden	Ja	Ja
Download Erreichbarkeit	Nein	Ja
Beschreibungen [Anzahl Zeichen]	35	1920
Anzahl Keywords	2	13
Anzahl Formate	0	2
HTTP-Statuscode	503	200
Sprachkonsistenz	Korrekt	Inkorrekt Sprache ist DE

Tabelle 1: Beispiel unterschiedlicher Qualitätsmerkmale von Metadatensätzen. Oben, Qualitätsmerkmale die von opendata.swiss verwendet werden. Unter, erarbeitet Qualitätsmerkmale

Autoren: Fabian Gross, Silas Haab, Examinator:in: Prof. Dr. Pia Bereuter, Friedrich Striewski, Betreuer: Stefan Eberlein, Experte: Lukas Zurbuchen (BAV)
© IGEO FHNW

Resultat, Diskussion und Erkenntnisse

Resultat: Metadatenbank

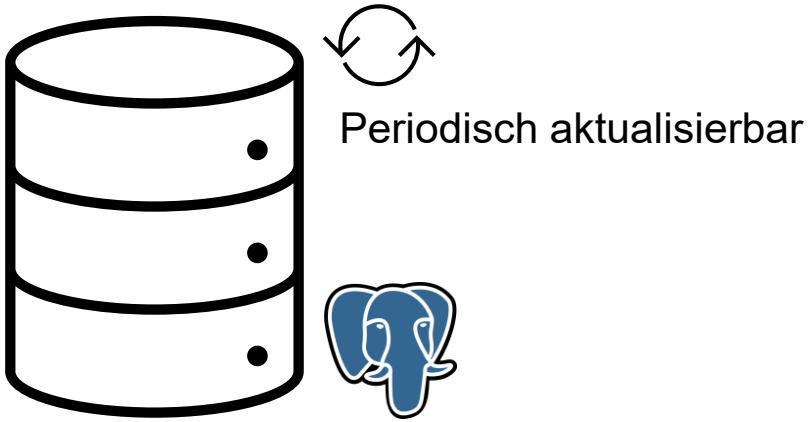


Abb. 10: (PostgreSQL, 2025)

- **25’055 Datensätze** davon sind **3’639 Mobilitätsdaten**
- **85’967** Beschreibungen in 5 Sprachen
- 219’724 Keywords
- PostgreSQL Datenbank

Erkenntnisse

- NLP ist der Way to go für dieses Aufgabe
- LLM sind gute Universal-Werkzeuge, die sich für unterschiedlichste Aufgaben eignen
- Nutzen eines gemeinsamen Metadatenstandards erleichtert Datenintegration

Diskussion

- Korrektur der Daten an Datenpublizierende durch Feedbackloop mitteilen
- Auffindbar sind nur Daten, die Open Government Data sind.

© IGEO FHNW

Literatur

- BAV, Bundesamt für Verkehr (2025): «Daten für ein effizientes Mobilitätssystem». URL: https://www.bav.admin.ch/bav/de/home/allgemeine-themen/mmm.html [Stand: 17.6.2025].
- geocat.ch (2025): «geocat.ch: THE SWISS GEOGRAPHIC CATALOGUE». URL: https://www.geocat.ch/datahub/news [Stand: 17.6.2025].
- geoharvester.ch (2025): «GeoHarvester». URL: https://geoharvester.ch/ [Stand 17.6.2025].
- Google Gemini (2025): «Gemini». URL: https://gemini.google.com/ [Stand 17.6.2025].
- opendata.swiss (2025): «Finden Sie Schweizer Open Government Data». URL: https://opendata.swiss/de [Stand: 17.6.2025].
- opentransportdata.swiss (2025): «opentransportdata.swiss». URL: https://opentransportdata.swiss/de/ [Stand: 17.6.2025].
- PostgresSQL (2025): «PostgreSQL: The World's Most Advanced Open Source Relational Database». URL: https://www.postgresql.org/ [Stand: 17.6.2025].

© IGEO FHNW