Automatische Korrektur von ÖV-Stationen in OpenStreetMap

FOSSGIS 2020
Freiburg im Breisgau

Patrick Brosi
12. März 2020

Lehrstuhl für Algorithmen und Datenstrukturen
Universität Freiburg

Vorstellung Lehrstuhl

Lehrstuhl für Algorithmen und Datenstrukturen

- Natural Language Processing aqqu.cs.uni-freiburg.de

Vorstellung Lehrstuhl

Lehrstuhl für Algorithmen und Datenstrukturen

- Natural Language Processing aqqu.cs.uni-freiburg.de
- Texterkennung in PDFs icecite.cs.uni-freiburg.de

Vorstellung Lehrstuhl

Lehrstuhl für Algorithmen und Datenstrukturen

- Natural Language Processing — aqqu.cs.uni-freiburg.de
- Texterkennung in PDFs — icecite.cs.uni-freiburg.de
- Routenplanung — maps.google.de

Vorstellung Lehrstuhl

Lehrstuhl für Algorithmen und Datenstrukturen

- Natural Language Processing — aqqu.cs.uni-freiburg.de
- Texterkennung in PDFs — icecite.cs.uni-freiburg.de
- Routenplanung — maps.google.de
- Graphdatenbanken — qlever.cs.uni-freiburg.de

Vorstellung Lehrstuhl

Lehrstuhl für Algorithmen und Datenstrukturen

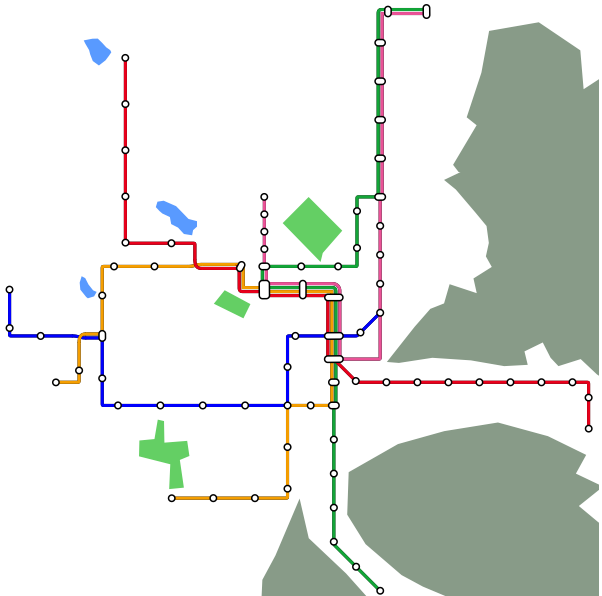
- Natural Language Processing — aqqu.cs.uni-freiburg.de
- Texterkennung in PDFs — icecite.cs.uni-freiburg.de
- Routenplanung — maps.google.de
- Graphdatenbanken — qlever.cs.uni-freiburg.de
- ÖV Map Matching (pfaedle) — travic.cs.uni-freiburg.de

Vorstellung Lehrstuhl

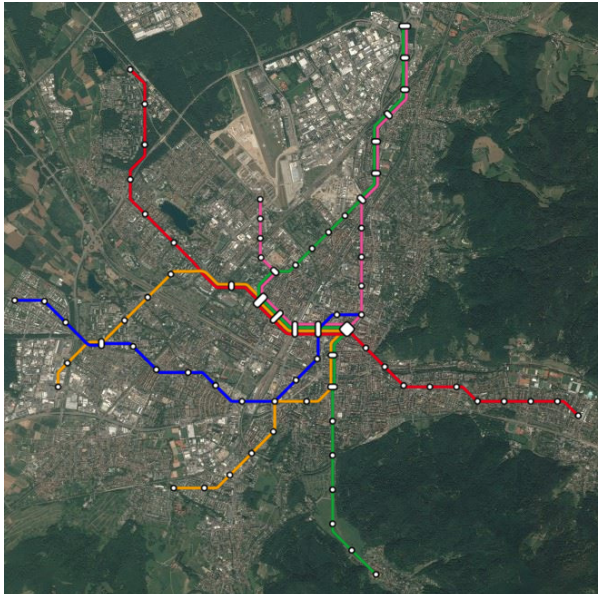
Lehrstuhl für Algorithmen und Datenstrukturen

- Natural Language Processing — aqqu.cs.uni-freiburg.de
- Texterkennung in PDFs — icecite.cs.uni-freiburg.de
- Routenplanung — maps.google.de
- Graphdatenbanken — qlever.cs.uni-freiburg.de
- ÖV Map Matching (pfaedle) — travic.cs.uni-freiburg.de
- Graph Drawing — loom.cs.uni-freiburg.de
  octi.cs.uni-freiburg.de

Vorstellung Lehrstuhl - octi + Obstacles



Vorstellung Lehrstuhl - octi als Overlay



Ähnlichkeiten von ÖV-Stationen

Häufiges Problem:

Ähnlichkeiten von ÖV-Stationen

Häufiges Problem:

Gegeben zwei Identifier von ÖV-Stationen s_1, s_2, jeweils bestehend aus einem Label (z.B. "Freiburg Hauptbahnhof") und einer Position (z.B. 47.997533, 7.840999)

Ähnlichkeiten von ÖV-Stationen

Häufiges Problem:

Gegeben zwei Identifier von ÖV-Stationen s_1, s_2, jeweils bestehend aus einem Label (z.B. "Freiburg Hauptbahnhof") und einer Position (z.B. 47.997533, 7.840999)

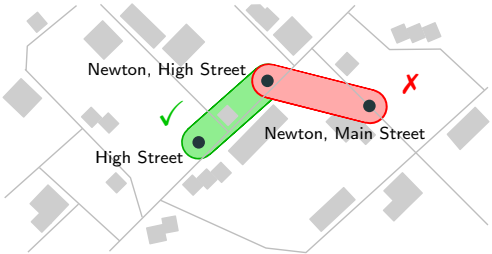
Beschreiben s_1 und s_2 dieselbe Station?

Ähnlichkeiten von ÖV-Stationen

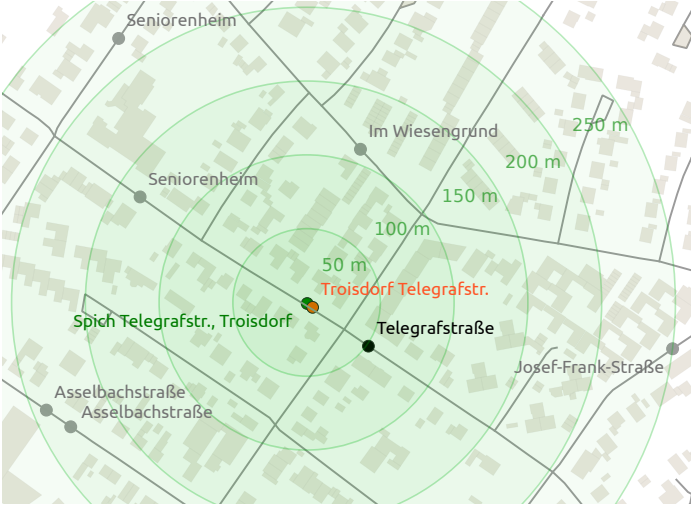
Häufiges Problem:

Gegeben zwei Identifier von ÖV-Stationen s_1, s_2, jeweils bestehend aus einem Label (z.B. "Freiburg Hauptbahnhof") und einer Position (z.B. 47.997533, 7.840999)

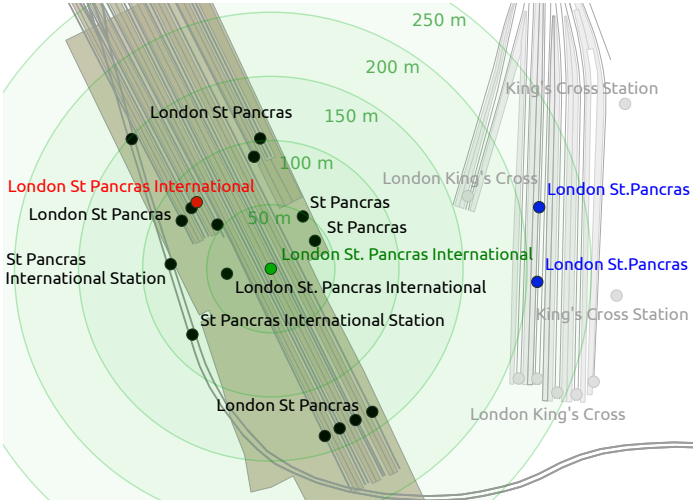
Beschreiben s_1 und s_2 dieselbe Station?



Einfache Heuristiken haben es schwer...



Einfache Heuristiken haben es schwer...



Einfache Heuristiken - Setup OSM D-A-CH

... aber wie schlecht sind sie?

- **Ground truth:** Paarweise Stationen (innerhalb 1 km) aus OpenStreetMap-Daten für D-A-CH die Mitglied einer public_transport=stop_area relation sind.

Einfache Heuristiken - Setup OSM D-A-CH

... aber wie schlecht sind sie?

- **Ground truth:** Paarweise Stationen (innerhalb 1 km) aus OpenStreetMap-Daten für D-A-CH die Mitglied einer public_transport=stop_area relation sind.
- 1,2 Millionen Stationen

Einfache Heuristiken - Setup OSM D-A-CH

... aber wie schlecht sind sie?

- **Ground truth:** Paarweise Stationen (innerhalb 1 km) aus OpenStreetMap-Daten für D-A-CH die Mitglied einer public_transport=stop_area relation sind.
- 1,2 Millionen Stationen
- Erweiterte Version: relevante Namesattribute (name, uic_name, ref_name, gtfs_name, ...) gelten als eigenständige Station, die im jeweiligen Node gruppiert ist.

Einfache Heuristiken - Setup OSM D-A-CH

... aber wie schlecht sind sie?

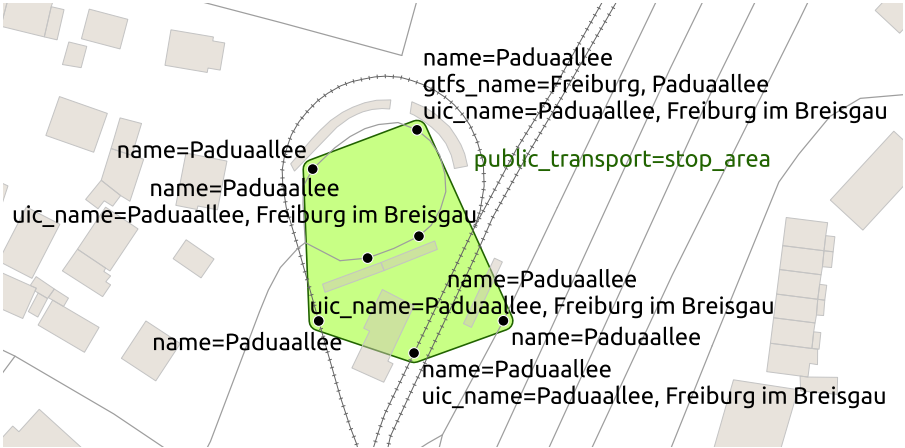
- **Ground truth:** Paarweise Stationen (innerhalb 1 km) aus OpenStreetMap-Daten für D-A-CH die Mitglied einer public_transport=stop_area relation sind.
- 1,2 Millionen Stationen
- Erweiterte Version: relevante Namesattribute (name, uic_name, ref_name, gtfs_name, ...) gelten als eigenständige Station, die im jeweiligen Node gruppiert ist.
- Stationspaare sind **ähnlich**, wenn sie in derselben Relation (Gruppe) sind.

Einfache Heuristiken - Setup OSM D-A-CH

... aber wie schlecht sind sie?

- **Ground truth:** Paarweise Stationen (innerhalb 1 km) aus OpenStreetMap-Daten für D-A-CH die Mitglied einer public_transport=stop_area relation sind.
- 1,2 Millionen Stationen
- Erweiterte Version: relevante Namesattribute (name, uic_name, ref_name, gtfs_name, ...) gelten als eigenständige Station, die im jeweiligen Node gruppiert ist.
- Stationspaare sind **ähnlich**, wenn sie in derselben Relation (Gruppe) sind.
- Stationspaare sind **nicht ähnlich**, wenn sie in verschiedenen Relationen (Gruppen) sind.

Einfache Heuristiken - Setup OSM D-A-CH



Einfache Heuristiken - Ergebnisse OSM D-A-CH

Ergebnisse für D-A-CH, Schwellwertparameter optimiert für besten F1-Score.

Methode	Bester Schwellwert	Precision	Recall	F1
Geo-Distanz	20 m	0.84	0.68	0.75
Editierdistanz	0.85	0.97	0.68	0.8
Präfix-Editierdistanz	0.9	0.93	0.74	0.82
Jaro	0.85	0.93	0.72	0.81
Jaro-Winkler	0.9	0.9	0.73	0.8
Jaccard Index	0.45	0.85	0.88	0.87
TF-IDF	0.65	0.87	0.88	**0.88**

Einfache Heuristiken - Ergebnisse OSM D-A-CH

Ergebnisse für D-A-CH, Schwellwertparameter optimiert für besten F1-Score.

Methode	Bester Schwellwert	Precision	Recall	F1
Geo-Distanz	20 m	0.84	0.68	0.75
Editierdistanz	0.85	0.97	0.68	0.8
Präfix-Editierdistanz	0.9	0.93	0.74	0.82
Jaro	0.85	0.93	0.72	0.81
Jaro-Winkler	0.9	0.9	0.73	0.8
Jaccard Index	0.45	0.85	0.88	0.87
TF-IDF	0.65	0.87	0.88	**0.88**

⇒ In der Praxis ungeeignet ☹

Einfache Heuristiken - 2. Versuch

Idee: verschiedene Methoden mittels Soft Voting kombinieren

Einfache Heuristiken - 2. Versuch

Idee: verschiedene Methoden mittels Soft Voting kombinieren

Ergebnisse für D-A-CH, Schwellwertparameter optimiert für besten F1-Score.

Methode	Bester Schwellwert	Precision	Recall	F1
Geo-Distanz + Editierdistanz	20 m + 0.99	0.91	0.82	0.86
Geo-Distanz + TF-IDF	40 m + 0.55	0.95	0.88	**0.92**

Einfache Heuristiken - 2. Versuch

Idee: verschiedene Methoden mittels Soft Voting kombinieren

Ergebnisse für D-A-CH, Schwellwertparameter optimiert für besten F1-Score.

Methode	Bester Schwellwert	Precision	Recall	F1
Geo-Distanz + Editierdistanz	20 m + 0.99	0.91	0.82	0.86
Geo-Distanz + TF-IDF	40 m + 0.55	0.95	0.88	**0.92**

⇒ Immer noch werden mehr als 10% der ähnlichen Paare nicht entdeckt ☹

Schreibweise von Stationen

Hauptbahnhof

Freiburg

Freiburg Hauptbahnhof

Freiburg Hbf

Freiburg im Breisgau

Freiburg im Breisgau Hauptbahnhof

Freiburg (Breisgau) Hauptbahnhof

Hauptbahnhof, Freiburg im Breisgau

Hauptbahnhof Freiburg

Freiburg (Breisgau), Hauptbahnhof

Freiburg(Brsg) Hauptbahnhof

Typische Fehler von TF-IDF

FN = Falsches Negatives, FP = Falsches Positives

FN

Auerbach (Karlsbad), Rosenweg @ (48.9161, 8.5341)
Rosenweg @ (48.9160, 8.5343)

FP

Cottbus, Kiekebusch Alte Schule @ (51.7215, 14.3646)
Kiekebusch Friedhof, Cottbus @ (51.7179, 14.3672)

Schreibweise von Stationen - Beobachtungen

- Bestimmte Tokens haben regional wenig bis keine Bedeutung (“Freiburg im Breisgau”)

Schreibweise von Stationen - Beobachtungen

- Bestimmte Tokens haben regional wenig bis keine Bedeutung (“Freiburg im Breisgau”)
- Die Reihenfolge der Tokens hat oft keine Bedeutung (“Freiburg Bertoldsbrunnen” vs. “Bertoldsbrunnen, Freiburg”)

Schreibweise von Stationen - Beobachtungen

- Bestimmte Tokens haben regional wenig bis keine Bedeutung (“Freiburg im Breisgau”)
- Die Reihenfolge der Tokens hat oft keine Bedeutung (“Freiburg Bertoldsbrunnen” vs. “Bertoldsbrunnen, Freiburg”)
- Bestimmte Tokens haben gängige Ab- und Verkürzungen (“Hbf”, “ZOB”, “Straße”, “Str.”)

Schreibweise von Stationen - Beobachtungen

- Bestimmte Tokens haben regional wenig bis keine Bedeutung (“Freiburg im Breisgau”)
- Die Reihenfolge der Tokens hat oft keine Bedeutung (“Freiburg Bertoldsbrunnen” vs. “Bertoldsbrunnen, Freiburg”)
- Bestimmte Tokens haben gängige Ab- und Verkürzungen (“Hbf”, “ZOB”, “Straße”, “Str.”)
- Bestimmte Tokens sind ein Indikator für eine große geografische Ausdehnung der Station (“Hauptbahnhof”)

Lernbasierter Ansatz

Idee: Nutze die folgenden Features für einen lernbasierten Klassifikator (jedes Feature pro Stationspaar (s_1, s_2)!)

- Distanz in Metern zwischen s_1 und s_2

Lernbasierter Ansatz

Idee: Nutze die folgenden Features für einen lernbasierten Klassifikator (jedes Feature pro Stationspaar (s_1, s_2)!)

- Distanz in Metern zwischen s_1 und s_2
- Position des Centroids von s_1 und s_2 auf einem verschachtelten Grid

Lernbasierter Ansatz

Idee: Nutze die folgenden Features für einen lernbasierten Klassifikator (jedes Feature pro Stationspaar (s_1, s_2)!)

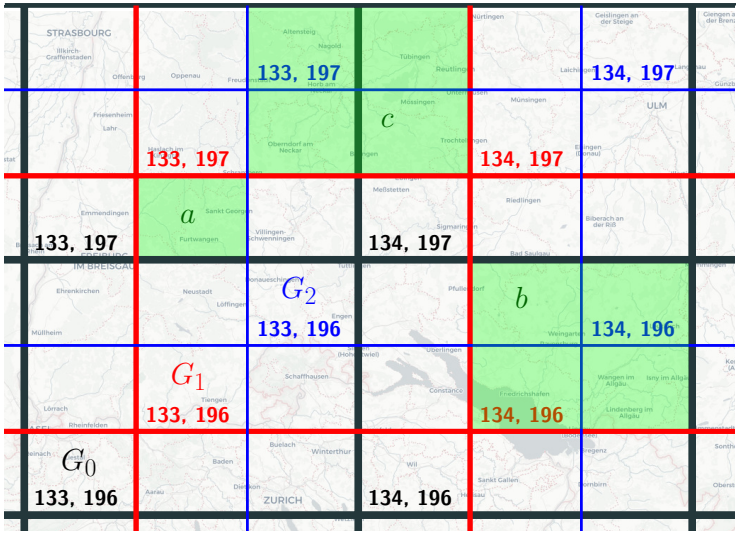
- Distanz in Metern zwischen s_1 und s_2
- Position des Centroids von s_1 und s_2 auf einem verschachtelten Grid
- Die Anzahl der 3-Gramme die nur in einem der beiden Labels vorkommt.

Lernbasierter Ansatz

Idee: Nutze die folgenden Features für einen lernbasierten Klassifikator (jedes Feature pro Stationspaar (s_1, s_2)!)

- Distanz in Metern zwischen s_1 und s_2
- Position des Centroids von s_1 und s_2 auf einem verschachtelten Grid
- Die Anzahl der 3-Gramme die nur in einem der beiden Labels vorkommt.
- Für die top-k 3-Gramme, die absolute Differenz der Vorkommen des jeweiligen 3-Gramme zwischen s_1 und s_2

Lernbasierter Ansatz - Verschachteltes Grid



Lernbasierter Ansatz - Beispiel

Drei Stationspaare in Freiburg:

1. "Freiburg im Breisgau Hauptbahnhof" (47.9966, 7.8404) vs. "Hauptbahnhof" (47.9965, 7.8407)

Lernbasierter Ansatz - Beispiel

Drei Stationspaare in Freiburg:

1. "Freiburg im Breisgau Hauptbahnhof" (47.9966, 7.8404) vs. "Hauptbahnhof" (47.9965, 7.8407)
2. "Okenstraße" (48.0105, 7.8545) vs. "Nordstraße" (48.0111, 7.8541)

Lernbasierter Ansatz - Beispiel

Drei Stationspaare in Freiburg:

1. "Freiburg im Breisgau Hauptbahnhof" (47.9966, 7.8404) vs. "Hauptbahnhof" (47.9965, 7.8407)
2. "Okenstraße" (48.0105, 7.8545) vs. "Nordstraße" (48.0111, 7.8541)
3. "Zentraler Omnibusbahnhof, Freiburg im Breisgau" (47.9960, 7.8407) vs. "ZOB" (47.9959, 7.8405)

Lernbasierter Ansatz - Beispiel

Drei Stationspaare in Freiburg:

1. "Freiburg im Breisgau Hauptbahnhof" (47.9966, 7.8404) vs. "Hauptbahnhof" (47.9965, 7.8407)
2. "Okenstraße" (48.0105, 7.8545) vs. "Nordstraße" (48.0111, 7.8541)
3. "Zentraler Omnibusbahnhof, Freiburg im Breisgau" (47.9960, 7.8407) vs. "ZOB" (47.9959, 7.8405)

Lernbasierter Ansatz - Beispiel

Drei Stationspaare in Freiburg:

1. "Freiburg im Breisgau Hauptbahnhof" (47.9966, 7.8404) vs. "Hauptbahnhof" (47.9965, 7.8407)
2. "Okenstraße" (48.0105, 7.8545) vs. "Nordstraße" (48.0111, 7.8541)
3. "Zentraler Omnibusbahnhof, Freiburg im Breisgau" (47.9960, 7.8407) vs. "ZOB" (47.9959, 7.8405)

d_m	d_{3g}	x_0	y_0	x_1	y_1	rei	tra	raß	aße	urg	bur	ibu	␣Fr	Fre	eib	rg␣	eis	"ähnlich"
24	20	133	196	133	195	2	0	0	0	1	1	1	1	1	1	1	1	**Ja**
72	10	133	196	133	195	0	0	0	0	0	0	0	0	0	0	0	0	**Nein**
12	47	133	196	133	195	2	1	0	0	1	1	2	1	1	1	1	1	**Ja**

Lernbasierter Ansatz - Ergebnisse

Ergebnisse für D-A-CH mit lernbasiertem Ansatz (Random Forest Classifier)

Methode	Bester Schwellwert	Precision	Recall	F1
TF-IDF	0.65	0.87	0.88	0.88
Geo-Distanz + TF-IDF	40 m + 0.55	0.95	0.88	0.92
ML (Random Forest)	—	**0.99**	**0.99**	**0.99**

Lernbasierter Ansatz - Ergebnisse

Ergebnisse für D-A-CH mit lernbasiertem Ansatz (Random Forest Classifier)

Methode	Bester Schwellwert	Precision	Recall	F1
TF-IDF	0.65	0.87	0.88	0.88
Geo-Distanz + TF-IDF	40 m + 0.55	0.95	0.88	0.92
ML (Random Forest)	—	**0.99**	**0.99**	**0.99**

⇒ Annähernd perfekt ☺

Lernbasierter Ansatz - Beobachtung

Bei der Analyse der verbleibenden falschen Positiven (FP) und falschen Negativen (FN) stellt man fest, dass der Klassifikator häufig Recht hatte, aber die **Ground Truth nicht korrekt war**.

Lernbasierter Ansatz - Beobachtung

Bei der Analyse der verbleibenden falschen Positiven (FP) und falschen Negativen (FN) stellt man fest, dass der Klassifikator häufig Recht hatte, aber die **Ground Truth nicht korrekt war**.

Idee: Nutze das Modell zur Fehlerkorrektur von Stationsdaten in OpenStreetMap.

staty

staty macht genau das:

- Für jedes Paar (s_1, s_2) von Stationen klassifiziert das Modell, ob s_1 und s_2 ähnlich sind.

staty

staty macht genau das:

- Für jedes Paar (s_1, s_2) von Stationen klassifiziert das Modell, ob s_1 und s_2 ähnlich sind.
- Sind sie es, aber in OSM sind sie nicht gruppiert, wird vorgeschlagen eine Gruppierung durchzuführen.

staty

staty macht genau das:

- Für jedes Paar (s_1, s_2) von Stationen klassifiziert das Modell, ob s_1 und s_2 ähnlich sind.
- Sind sie es, aber in OSM sind sie nicht gruppiert, wird vorgeschlagen eine Gruppierung durchzuführen.
- Sind sie es nicht, aber in OSM sind sie gruppiert, wird empfohlen, die Station aus der Gruppe zu lösen.

staty

staty macht genau das:

- Für jedes Paar (s_1, s_2) von Stationen klassifiziert das Modell, ob s_1 und s_2 ähnlich sind.
- Sind sie es, aber in OSM sind sie nicht gruppiert, wird vorgeschlagen eine Gruppierung durchzuführen.
- Sind sie es nicht, aber in OSM sind sie gruppiert, wird empfohlen, die Station aus der Gruppe zu lösen.

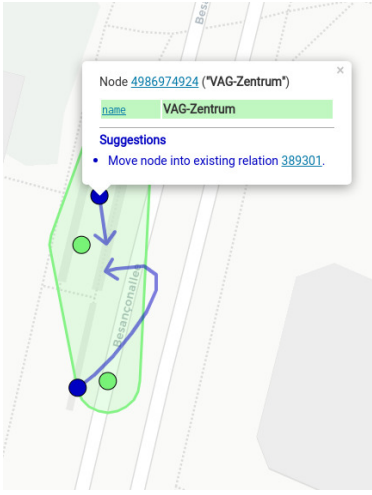
staty

staty macht genau das:

- Für jedes Paar (s_1, s_2) von Stationen klassifiziert das Modell, ob s_1 und s_2 ähnlich sind.
- Sind sie es, aber in OSM sind sie nicht gruppiert, wird vorgeschlagen eine Gruppierung durchzuführen.
- Sind sie es nicht, aber in OSM sind sie gruppiert, wird empfohlen, die Station aus der Gruppe zu lösen.

Aktuell noch in Entwicklung, eine aktuelle Version für D-A-CH ist aber online:

staty.cs.uni-freiburg.de



Vielen Dank!