Semantischer Vergleich von Fahrplandaten

Bachelorarbeit von Phong Tran

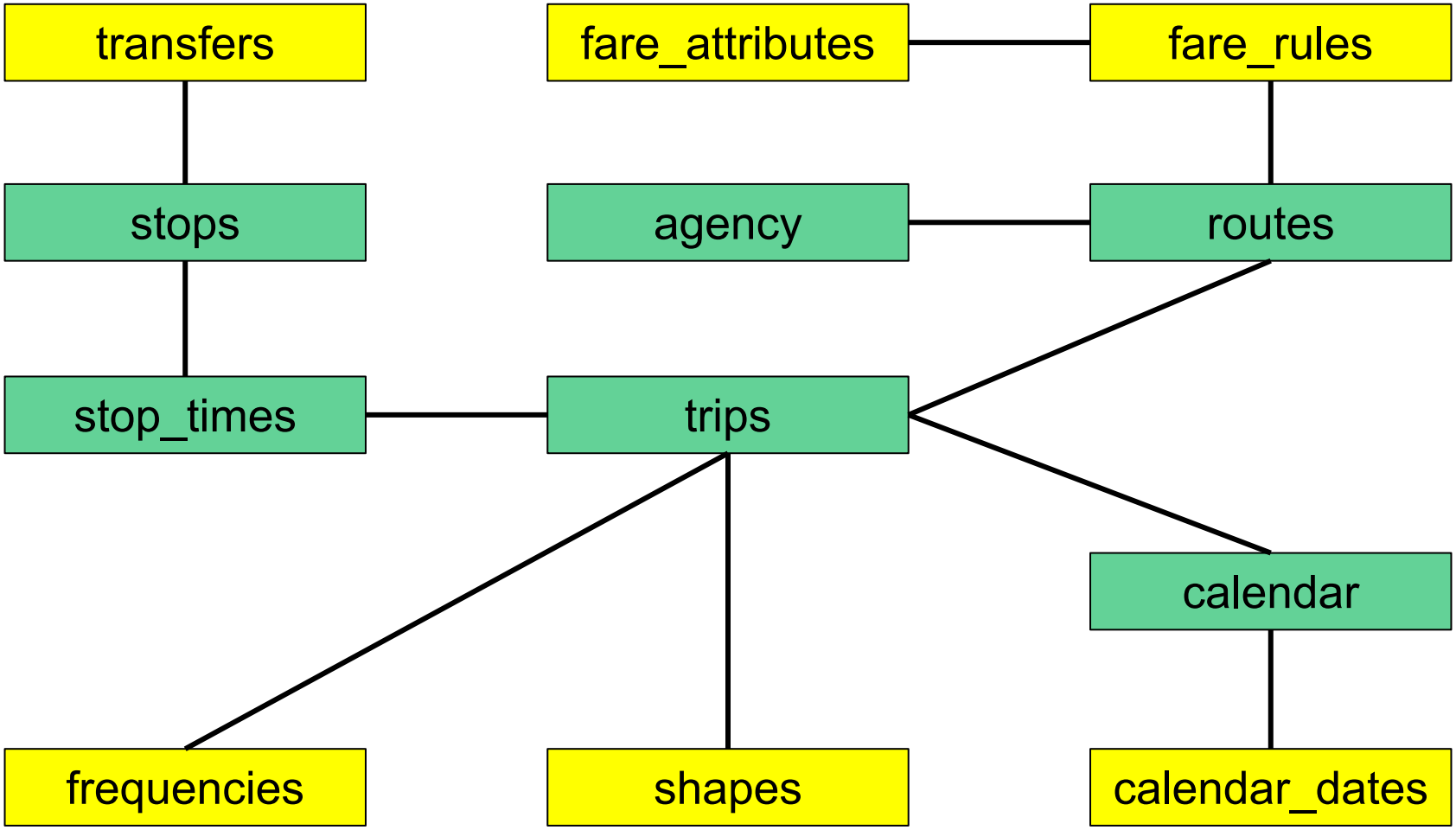
Übersicht

- Einführung GTFS
- Aufbau der Datenstruktur & Implementierung
- Evaluation & Erweiterungen

Was ist GTFS?

- General Transit Feed Specification
- Format für Darstellung von Fahrplandaten des öffentlichen Nahverkehrs
- Format:
  - Eine zip-Datei bestehend aus mehreren CSV-Dateien
  - Im Standard: 13 CSV-Dateien
  - Beliebig erweiterbar

Übersicht der CSV-Dateien



Beispiel von Haltestellen

stops.txt

stop_id,stop_name,stop_lat,stop_lon
de:8311:30400:0:1,Dorfstraße,47.964962,7.857308
de:8311:30401:0:1,Klosterplatz,47.96583,7.8558292
de:8311:30402:0:1,Wiesenweg,47.96695,7.852604
de:8311:30403:0:1,Wonnhalde,47.975033,7.847777

stop_times.txt

trip_id,arrival_time,departure_time,stop_id,stop_sequence
741.T2.11-2-I-j16-1.45.H,5:15:00,5:15:00,de:8311:30400:0:1,1
741.T2.11-2-I-j16-1.45.H,5:16:00,5:16:00,de:8311:30401:0:1,2
741.T2.11-2-I-j16-1.45.H,5:17:00,5:17:00,de:8311:30402:0:1,3
741.T2.11-2-I-j16-1.45.H,5:19:00,5:19:00,de:8311:30403:0:1,4

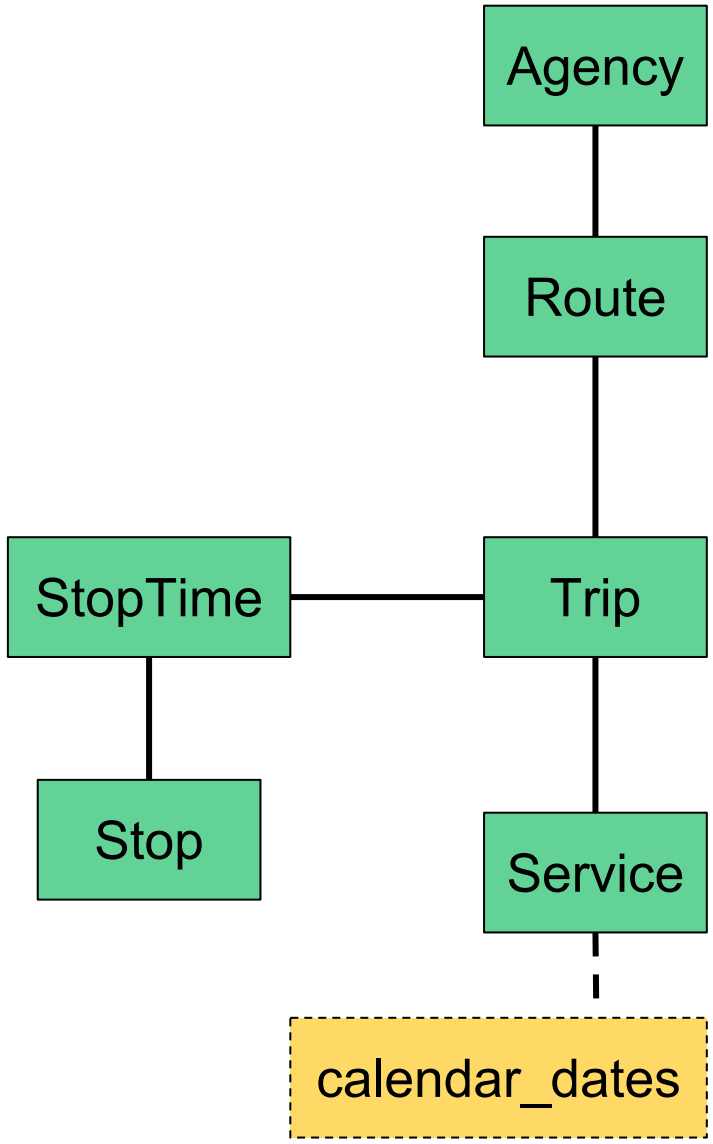
Beispiel von Servicezeiten

calendar.txt
service_id,monday,tuesday,wednesday,thursday,friday,saturday,sunday,start_date,end_date
T2,0,0,0,0,0,1,0,20151211,20161210

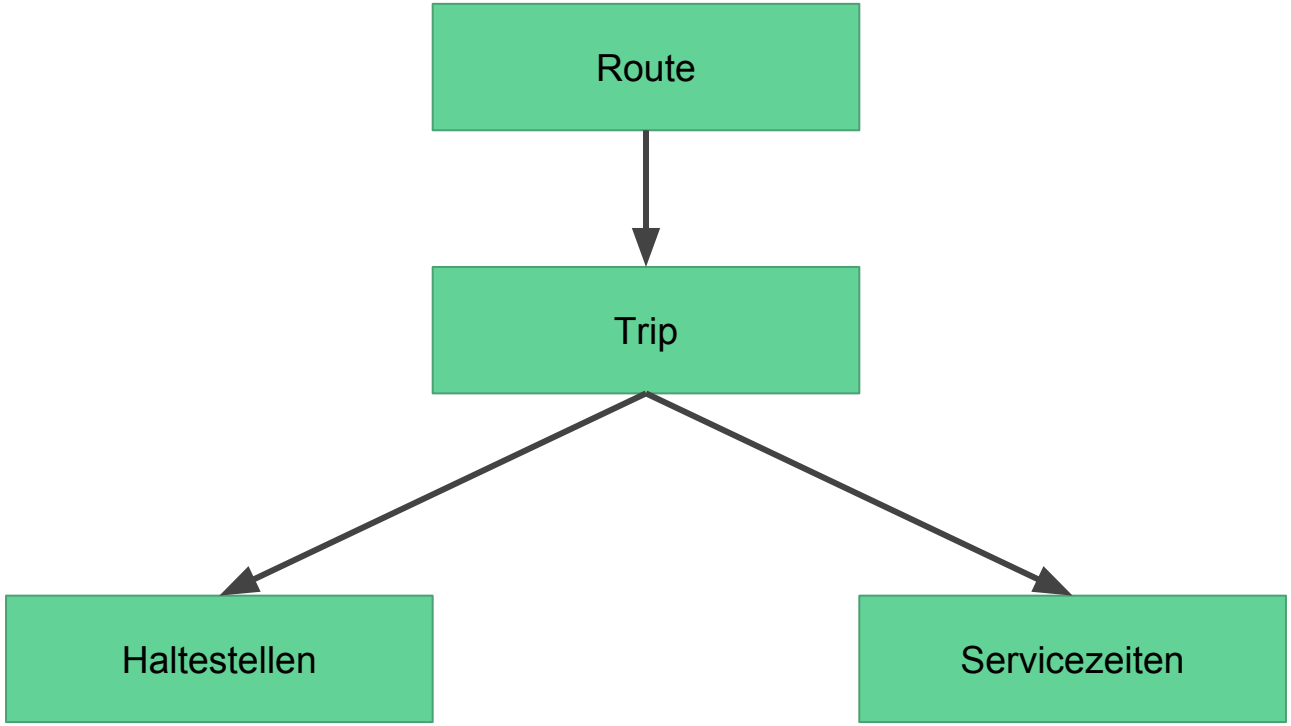
calendar_dates.txt
service_id,date,exception_type
T2,20151226,2

Aufbau der Datenstruktur

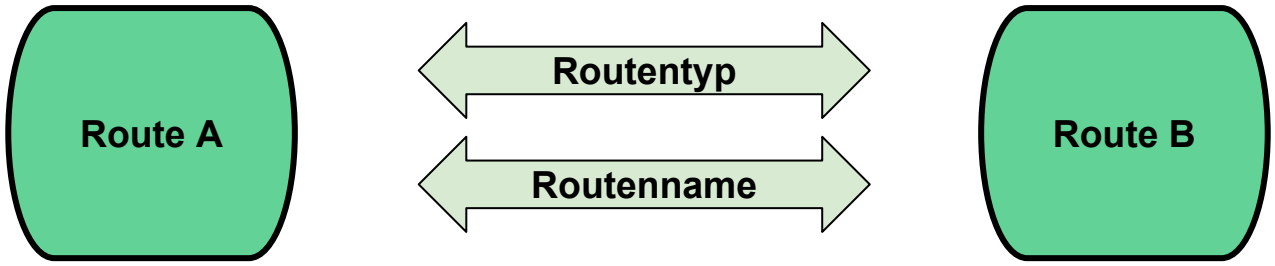
- Hash-Tabellen für Objekte mit ID’s als Schlüssel (Stop, Agency, Route, Trip, Service)
- “calendar_dates”-Einträge in einer Hash-Tabelle im Service-Objekt
- StopTime-Objekte als sortierte Liste im Trip-Objekt



Implementierung der Vergleiche

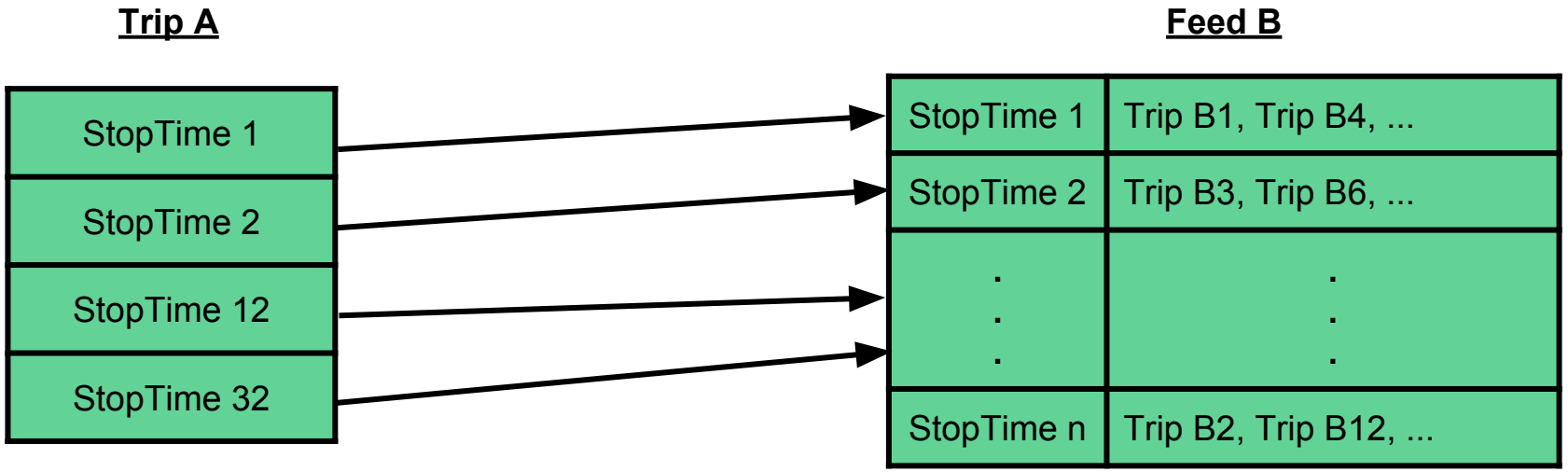


Vergleich von Routen



- Namen müssen mind. 75% übereinstimmen
- Berechnung mittels Levenshtein-Distanz
  - Anzahl von Einfüge,-Lösch und Tausch-Operationen um eine Zeichenkette in eine andere umzuwandeln

Vergleich von Trips

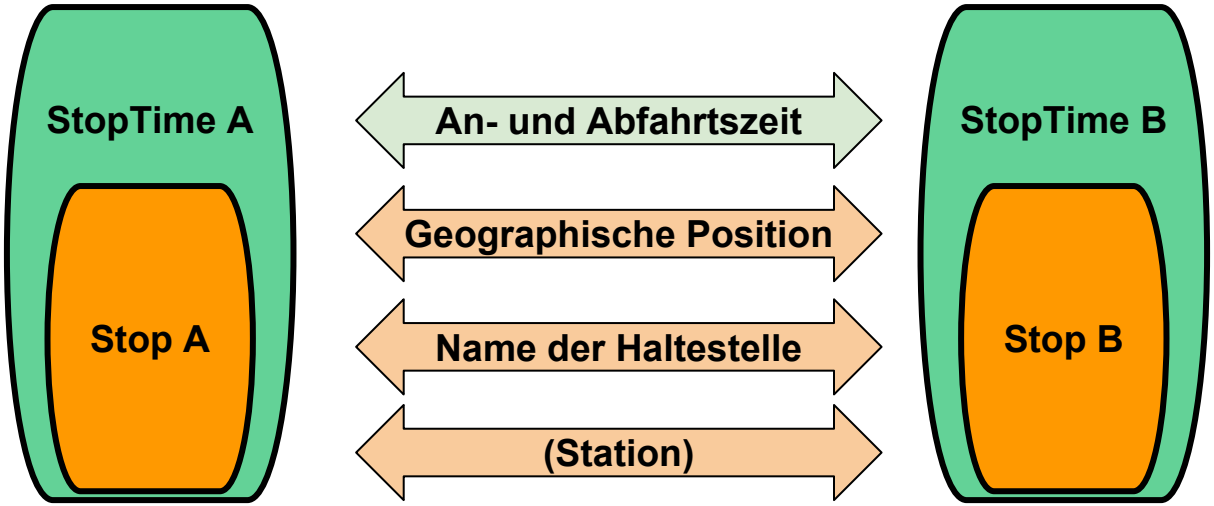


- Trips vom zweiten Feed (Feed B) werden umsortiert um Anzahl der zu vergleichenden Trips zu verringern
- Neue HashMap mit StopTime-Objekt als Schlüssel und Liste von Trips, welche die Haltestelle anfahren

Übereinstimmungstypen

- **FULL_MATCH:**
  - Haltestellen und Servicezeiten sind identisch.
- **FULL_STOP_MATCH:**
  - Haltestellen stimmen überein, Servicezeiten sind unterschiedlich.
- **MATCH:**
  - Teilweise Übereinstimmung von Haltestellen und Servicezeiten.
- **NO_MATCH:**
  - Keine Übereinstimmung von Haltestellen.

Vergleich von Haltestellen



- An- und Abfahrtszeiten müssen exakt übereinstimmen
- Namen müssen mind. 75% übereinstimmen
- Geographische Position muss bis auf 5 Meter übereinstimmen
- Anzahl der gemeinsamen Haltestellen durch LCS

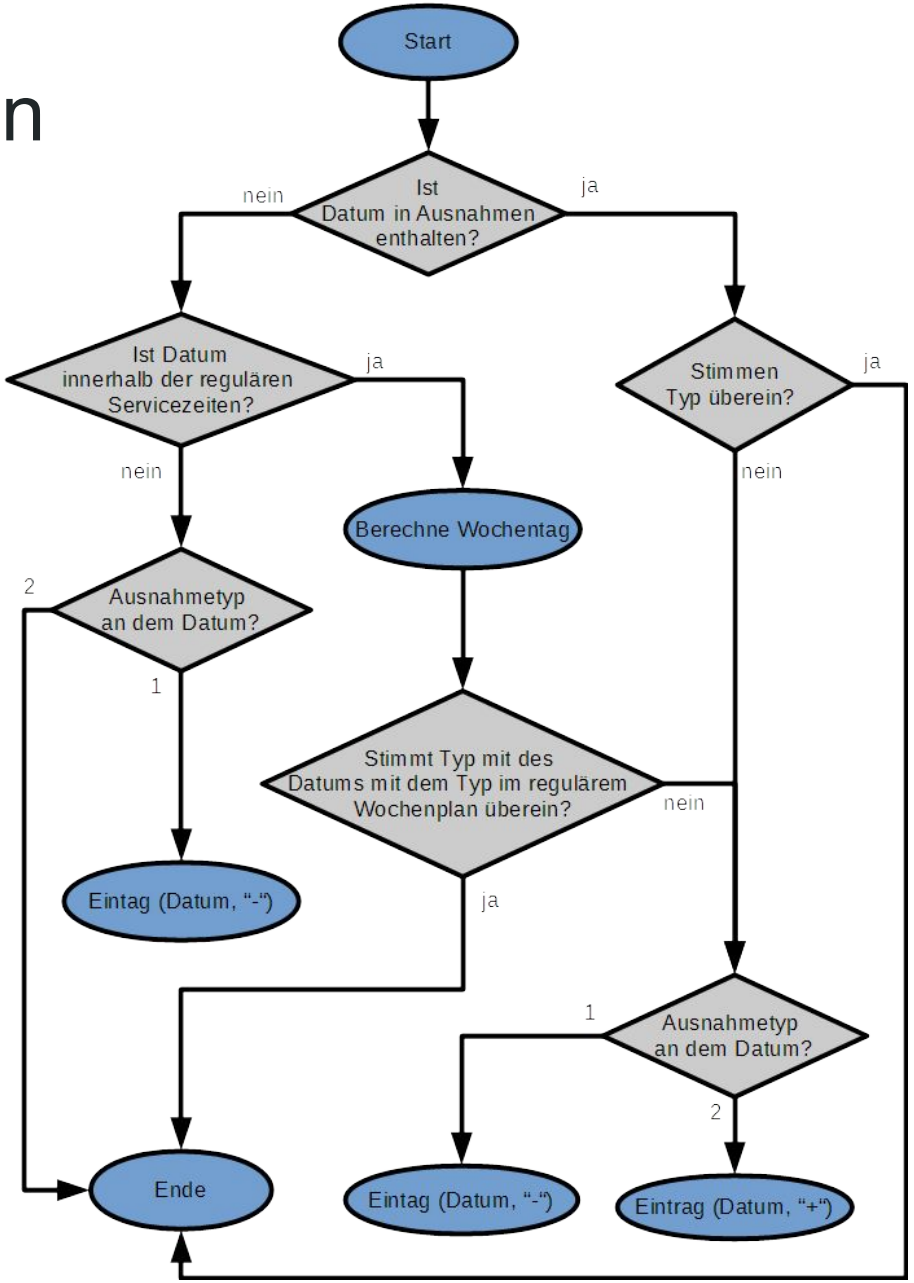
Vergleich von Servicezeiten

- Zwei Vergleiche:
  - Vergleich der Ausnahmen (calendar_dates.txt)
  - Vergleich der regulären Servicezeiten (calendar.txt)
- Ergebnisse in Hash-Tabelle gespeichert:
  - Schlüssel: Datum
  - Wert: “-” falls Service A aktiv ist, “+” falls Service B aktiv ist
- Ist die Hash-Tabelle leer nach dem Vergleich und Übereinstimmungstyp war **FULL_STOP_MATCH**

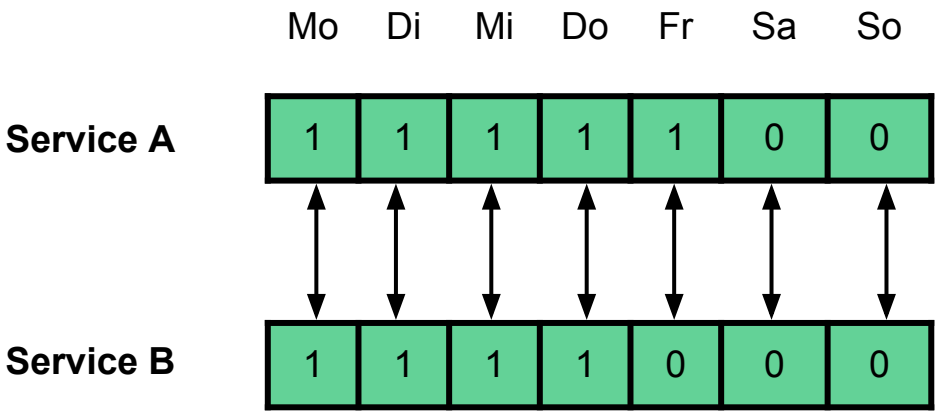
  ➔ Änderung des Übereinstimmungstyp auf **FULL_MATCH**

Vergleich der Ausnahmen

- Vergleich der Ausnahmen von Service A mit Service B
- “-” und “+” vertauschen für Vergleich von Service B mit Service A



Vergleich der regulären Servicezeiten



- Wochentage werden miteinander verglichen
- Bei Unterschied: Eintrag in die Ergebnistabelle für den aktiven Service

Ergebnisse der Vergleiche

- Ergebnisse in einem **TripData-**Objekt gespeichert
- TripData:
  - Übereinstimmungstyp
  - Prozent der gleichen Haltestellen
  - Anzahl der sich unterscheidenden Servicetage
  - ID’s der Trips
  - Liste mit exklusiven StopTime-Objekten beider Feeds
- TripData-Objekte werden in eine Prioritätswarteschlange gespeichert

Ausgabe der Ergebnisse

- Aggregierte Ausgabe:

```
1 9 trip(s) stayed the same
2 1 trip(s) with same stop(s), 2 different service date(s)
3 2 partial trip match(es), 4 stop(s) removed, 3 stop(s) added, 2 different date(s)
4 4 trip(s) removed, 1 trip(s) added, 26 stop(s) removed, 4 stop(s) added
5 6 route(s) stayed the same, 0 route(s) removed, 0 route(s) added
```

Ausgabe der Ergebnisse

- Vollständige Ausgabe

```
1 ~ trip; id_a: 39744A361B1981A; id_b: 39744A361B1981B
2 D:\Dokumente\Eclipse\DiffToolForGtfs\TestFeedOne.zip\stop_times.txt:142: -stop; on trip: 39744A361B1981A
      ; id: 19016; name: Sutter Hill Transit Center, arr_time: 09:05:00, dep_time: 09:05:00
3 D:\Dokumente\Eclipse\DiffToolForGtfs\TestFeedOne.zip\stop_times.txt:143: -stop; on trip: 39744A361B1981A
      ; id: 20474; name: Sutter Creek Auditorium, arr_time: 09:08:00, dep_time: 09:08:00
4 D:\Dokumente\Eclipse\DiffToolForGtfs\TestFeedTwo.zip\stop_times.txt:143: +stop; on trip: 39744A361B1981B
      ; id: 19014; name: Safeway, arr_time: 09:20:00, dep_time: 09:20:00
5 D:\Dokumente\Eclipse\DiffToolForGtfs\TestFeedTwo.zip\stop_times.txt:151: +stop; on trip: 39744A361B1981B
      ; id: 18993; name: Petkovich Park, arr_time: 09:47:00, dep_time: 09:47:00
6 D:\Dokumente\Eclipse\DiffToolForGtfs\TestFeedTwo.zip\stop_times.txt:152: +stop; on trip: 39744A361B1981B
      ; id: 18994; name: 150 Main St., arr_time: 09:48:00, dep_time: 09:48:00
7 D:\Dokumente\Eclipse\DiffToolForGtfs\TestFeedOne.zip\trips.txt:13: -trip; id: AAA-FREQ5; short_name: ;
      route_name: Sacramento Express, first_station: J St. & 3rd St., last_station: Murieta Dr. & Lone
      Pine Dr., start: 05:50:00, end: 06:20:00
```

Evaluation (1)

VAG Freiburg

Inhalte der Feeds		original	gtfstidy
Stops		713	698
StopTimes		234 533	234 533
Services		263	261
Routes		24	24
Trips durch frequencies.txt		0	16 652
Trips		18 847	18 847

Unterschiede der Feeds

Gleiche Routen		24 (100%)	24 (100%)
Gleiche Trips		18 847 (100%)	18 847 (100%)
Gleiche Trips mit unterschiedlichen Servicezeiten	Trips	0 (0%)	0 (0%)
Trips mit teilweiser Übereinstimmung	Trips	0 (0%)	0 (0%)
	StopTimes	0 (0%)	0 (0%)
Exklusive Trips		0 (0%)	0 (0%)
Exklusive StopTimes		0 (0%)	0 (0%)
Exclusive Routen		0 (0%)	0 (0%)

Evaluation (2)

Metra Chicago

Inhalte der Feeds		April 2016	April 2017
Stops		239	239
StopTimes		150 716	130 962
Services		29	25
Routes		11	11
Trips durch frequencies.txt		0	0
Trips		8817	7700

Unterschiede der Feeds

Gleiche Routen		11 (100%)	11 (100%)
Gleiche Trips		0 (0%)	0 (0%)
Gleiche Trips mit unterschiedlichen Servicezeiten	Trips	6119 (69,4%)	6119 (79,5%)
Trips mit teilweiser Übereinstimmung	Trips	1460 (16,6%)	1460 (16,6%)
	StopTimes	7018 (4,7%)	6617 (5,1%)
Exklusive Trips		1238 (14,0%)	121 (1,6%)
Exklusive StopTimes		21 273 (14,1%)	1920 (24,9%)
Exclusive Routen		0 (0%)	0 (0%)

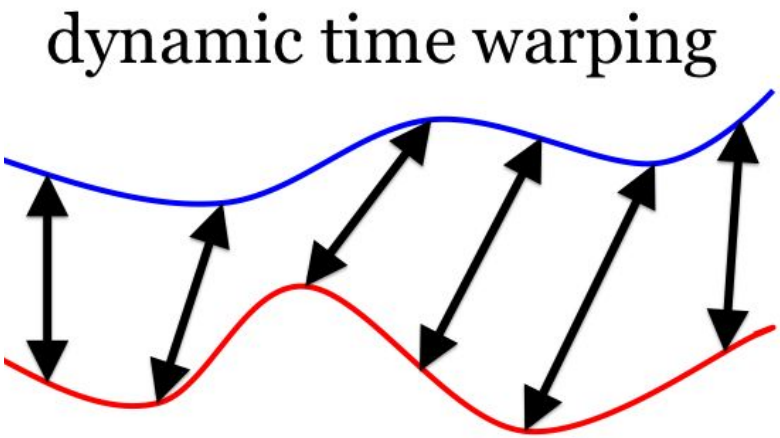
Evaluation (3)

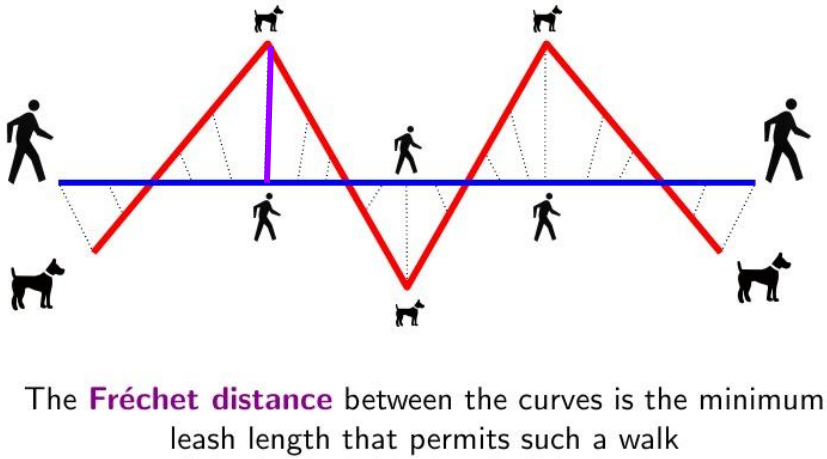
Nahverkehr der DB in Südwestdeutschland

Inhalte der Feeds		offiziell	Zhang
Stops		2202	1418
StopTimes		36 668	59 136
Services		2580	1
Routes		2579	132
Trips durch frequencies.txt		0	0
Trips		2579	7896
Unterschiede der Feeds			
Gleiche Routen		2 (0,1%)	2 (1,5%)
Gleiche Trips		0 (0%)	0 (0%)
Gleiche Trips mit unterschiedlichen Servicezeiten	Trips	0 (0%)	0 (0%)
Trips mit teilweiser Übereinstimmung	Trips	0 (0%)	0 (0%)
	StopTimes	0 (0%)	0 (0%)
Exklusive Trips		2579 (100%)	7896 (100%)
Exklusive StopTimes		36668 (100%)	59136 (100%)
Exclusive Routen		2577 (99,9%)	130 (98,5%)

Erweiterungen (1)

- Vergleich von Pfaden (shapes.txt)





Erweiterungen (2)

- Generierung eines neuen Feeds
  - Sammlung der Trips in neuer Hash-Tabelle
  - ID’s müssen evtl. angepasst werden (keine Duplikate)
  - **FULL_MATCH** und exklusive Trips: Trips übernehmen
  - **FULL_STOP_MATCH**: Einen Trip übernehmen und Servicezeiten anpassen
  - **MATCH**: Haltestellen und Servicezeiten eines Trips anpassen
- Vergleich auf Abdeckungsähnlichkeit
  - Anzahl der Fahrten zwischen zwei Haltestellen

Bilderquellen

- Dynamic Time Warping: https://de.mathworks.com/matlabcentral/fileexchange/43156-dynamic-time-warping--dtw-
- Fréchet-Distanz: https://www.slideshare.net/shripadthite/frechettalk