

Analyse und Auswertung von Echtzeit-Fahrplänen der Deutschen Bahn (Project D-Railing)

STUDIENARBEIT

für die Prüfung zum
Bachelor of Engineering
des Studienganges Informationstechnik
an der
Dualen Hochschule Baden-Württemberg Karlsruhe
von
Alexander Bierenstiel, André Schmitt, Dominik Schmitt

Abgabedatum 14. Mai 2018

Bearbeitungszeitraum	900 Stunden
Matrikelnummer	2496963, 3272367, 7191584
Kurs	TINF15B3
Ausbildungsfirma	Sick AG, E.G.O. Gerätebau, netcup GmbH Waldkirch, Oberderdingen, Karlsruhe
Gutachter der Studienakademie	Prof. Dr. Jürgen Vollmer

Erklärung

Ich versichere hiermit, dass ich meine Studienarbeit mit dem Thema: „Analyse und Auswertung von Echtzeit-Fahrplänen der Deutschen Bahn“ selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort Datum

Unterschrift

Sofern von der Ausbildungsstätte ein Sperrvermerk gewünscht wird, ist folgende Formulierung zu verwenden:

Sperrvermerk ja oder nein

Sperrvermerk

Der Inhalt dieser Arbeit darf weder als Ganzes noch in Auszügen Personen außerhalb des Prüfungsprozesses und des Evaluationsverfahrens zugänglich gemacht werden, sofern keine anders lautende Genehmigung der Ausbildungsstätte vorliegt.

Zusammenfassung

Dieses Abstract besser schreiben und eventuell eine englische Übersetzung anfertigen

Die vorliegende Studienarbeit befasst sich mit dem Thema der deutschen Bahn und ihrer Verspätungen. Es soll die von der Bahn zu Verfügung gestellten API genutzt werden, um Daten zu sammeln. Anhand dieser Daten soll ein neuronales Netz modelliert werden, welches genutzt werden kann, um Verspätungen und Abhängigkeiten im Schienenverkehr zu erkennen und vorherzusagen.

Inhaltsverzeichnis

1	Einleitung	7
1.1	Motivation	7
1.2	Stand der Technik	7
1.3	Ziel der Studienarbeit	7
1.4	Begriffsdefinitionen	8
2	Grundlagen	9
2.1	Die DB Timetable API	9
2.1.1	Station	9
2.1.2	Plan	10
2.1.3	fchg	10
2.1.4	rchg	10
2.2	Planung	11
2.3	Data Mining	11
2.4	Datenmodell	13
2.5	Modellierung realer Größen	13
2.6	Aufbereitung von Daten	14
2.7	Neuronalen Netzen an simplen Beispielen erklärt	15
2.8	Eingabe- und Ausgabe-Parameter für das Neuronale Netz	15
2.9	Literaturhinweise und Empfehlungen	15
3	Datenbeschaffung	16
3.1	Programmierung des Data Miners	16
3.2	Datenbank und Schema	17
4	Datenverarbeitung mit Data Mining	19
4.1	Stochastische Analyse	19
4.2	Visualisierung	19
5	Datenverarbeitung mit neuronalem Netz	20
5.1	Programmierung der Automatischen Datenverarbeitung	20
5.2	Vorverarbeitung der Datensätze	20
5.3	Programmieren der Modellstruktur des neuronalen Netzes	23
5.4	Vermeidung von Overfitting und Anpassungen um die Genauigkeit zu erhöhen	23

<i>INHALTSVERZEICHNIS</i>	2
5.5 Anlernen des Netzes	23
5.6 Verifizieren des Netzes	23
5.7 Vorhersagen anhand des Netzes	23
5.8 Auswertung und Fehlerbehandlung	24
6 Schlussfolgerung	25
6.1 Rückblick	25
6.2 Fazit	25
6.3 Ausblick	25
Anhang	26
Literaturverzeichnis	26
Liste der ToDo's	26

Abbildungsverzeichnis

3.1 Grundablauf des Miners	17
--------------------------------------	----

Tabellenverzeichnis

5.1	Vorverarbeitung der Datenbank-Daten	22
-----	---	----

Liste der Quellcodeausschnitte

Abkürzungsverzeichnis

BRV	Bahnhof-respektive Verzögerung	14
BRVD	Bahnhof-respektiver Verzögerungsdurchschnitt	14
SARV	Streckenabschnitt-respektive Verzögerung	14

Kapitel 1

Einleitung

1.1 Motivation

Wieso wollen wir das machen und warum ist das für uns wichtig.

Verspätung im öffentlichen Nah- und Fernverkehr treten täglich auf. Da häufig die Ursachen der Verspätung nicht direkt erkennbar sind, soll mit dieser Studienarbeit die Vorhersage von Verspätung ermöglicht werden. Dies kann für Pendler von Vorteil sein, um nicht zu spät zu Meetings oder zur Arbeit zu kommen. Ein weiterer gewünschter Nebeneffekt ist die Einsparung unnötiger Wartezeiten auf den gegebenenfalls nächsten Pünktlichen Zug.

1.2 Stand der Technik

Hier etwas zum Stand der Technik schreiben, neuronale Netze, Tensorflow, KI, Data-mining, OpenData, etc

1.3 Ziel der Studienarbeit

Hier das Ziel aus der Anmeldung schön definieren und klar Abgrenzen was Ziel und was optional nice to have ist.

Feststellungen von Verspätungen und Analyse nach

- Ort
- Zeit
- Strecke
- kritische Punkte

Visualisierung der Analyseergebnisse Optional: Vorhersage von weiteren Verspätung durch

- Ort
- Zeit
- Strecke
- kritische Punkte
- Wetterdaten
 - Wind
 - Regen
 - Temperatur
- Höhenlage eines Bahnhofs (z. B. Schneefall)

1.4 Begriffsdefinitionen

Im Rahmen dieser Arbeit werden bestimmte Begriffe verwendet, den eine spezielle Bedeutung beigemessen wird. Damit der Leser diese Begriffe nicht mit der alltäglichen Bedeutung verwechselt, werden sie im Folgenden definiert.

Streckenabschnitt Ein Streckenabschnitt besteht aus einem Gleis oder mehreren Gleise und verbindet zwei Bahnhöfe. Ein Streckenabschnitt wird eindeutig durch die von ihm verbundenen Bahnhöfe identifiziert.

Linie Im Sinne eines Verkehrsnetzes beschreibt die Linie eine Folge von anzufahrenden Bahnhöfen. Um eine Linie eindeutig zu beschreiben, bedarf es einer Menge von Bahnhöfe, die in ihrer anzufahrenden Reihenfolge angeordnet sind.

Kapitel 2

Grundlagen

2.1 Die DB Timetable API

Was bekommen wir eigentlich alles über die Api geliefert

API-URL: <http://api.deutschebahn.com/timetables/v1>
API-Swagger: https://editor.swagger.io/?_ga=2.234759646.1724072740.1516449724-126494731510747057#/

2.1.1 Station

Dieser Endpunkt gibt Informationen über ein Bahnhof zurück. Dafür kann sowohl der Name der Station, die eindeutige EVA Nummer oder die ds100 bzw. rl100 Nummer zur Identifikation angegeben werden. Der Klin'sche Stern kann verwendet werden, um alle Stationen abzurufen. Wurde der Server nicht gefunden, wird der Http-Code **404** zurückgegeben. War der Aufruf erfolgreich, so gibt die API den Status **200** zurück.

Außerdem wird ein Container mit den angefragten Stationen zurückgegeben. Innerhalb eines Stations-Objekt, werden die verschiedenen Identifikationsmöglichkeiten angegeben. Darunter auch die von der Timetable oft genutzte EVA-Nummer. Mit ihr kann jede Bahnstation in Deutschland eindeutig identifiziert werden.

Des Weiteren werden die Plattformen der Bahnstation mit Pipe („|“) angegeben. Der Meta-Eintrag gibt weitere EVA-Nummern an, die mit diesem Bahnhof zusammenhängen (Subbahnhof). Konnte der Bahnhof nicht identifiziert werden, so wird ein leeres Objekt zurückgegeben. Beispiel:

Request:

```
https://api.deutschebahn.com/timetables/v1/station/Heidelberg%20HBF
```

Response:

```
<stations>
  <station p="4|5" meta="518168|8070043"
    name='Heidelberg Hbf' eva="8000156" ds100="RH"/>
```

</stations>

2.1.2 Plan

Durch Angabe der EVA nummer (String), eines Datums und einer Stunde, können planmäßige Abfahrten an dem gewählten Bahnhof innerhalb der angegebenen Stunde abgefragt werden. Dabei ist das Datum als String im „YYMMDD“ Format anzugeben. Die Stunde ist ebenfalls als String anzugeben, diese soll im „HH“ Format angegeben werden.

```
/timetable/plan/{evaNo}/{date}{hour}:  
    evaNo: Angabe des Bahnhofs  
    date: angabe des gesuchten datums (YYMMDD)  
    hour: gesuchte stunde (HH)
```

Gibt ein Timetable-Objekt zurück, in dem alle geplanten Abfahrten in der angegebenen Stunde enthält. Dabei werden keine Änderungen durch Verspätungen berücksichtigt.

Responses:

200 Successfull operation

Gibt ein Timetable-Objekt zurück. In ihm ist der Stationsname, und die EVA-Nummer der Station gekapselt. Außerdem enthält es Listen von Timetable-Stop und Message-Objekten. In einer Plan-Response werden keine Messages übertragen. Es werden nur die "planend" Attribute genutzt.

2.1.3 fchg

Der "fchg" Endpunkt nimmt eine EVA-Nummer (String) entgegen und gibt ein Timetable-Objekt zurück. Darin werden alle Änderungen vom Zeitpunkt der Anfrage an gespeichert.

```
/timetable/fchg/{evaNo}:  
    evaNo: Angabe des Bahnhofs
```

Innerhalb des Timetabele wird der Name der Station, die EVA Nummer eine Liste von Timetable-Stops und Messages.

2.1.4 rchg

Durch Angabe einer EVA-Nummer können alle Änderungen der letzten zwei Minuten zurückgegeben. Alle 30 Sekunden werden diese aktualisiert.

```
/timetable/rchg/{evaNo}:  
    evaNo: Angabe des Bahnhofs
```

Der rchg Endpunkt ist sowohl von den Eingabeparametern als auch von den Ausgabeparametern gleich. der einzige Unterschied ist, dass die Änderungen die Übertragen werden in der Vergangenheit liegen.

Timetablestop In einem Timetablestop werden eine ID aus einer Daily-Trip-ID, Abfahrtsdatum des Zuges am Beginn der Linie und der Nummer des Stops gespeichert. Außerdem die aktuelle EVA-Nummer, die Bezeichnung der Stecke, eine Referenz zum eigentlichen Zug, wenn es ein Ersatzzug ist, die Events Ankunft und Abfahrt, in denen vor allem die An- bzw. Abfahrtszeiten und das Gleis untergebracht sind. Wobei jeweils die geplante als auch die prognostizierte Information enthalten sein kann, eine Message, warum eine Änderung gemacht worden ist, sowie Informationen, die angeben wie viel Verspätung die Bahn hat und ob sie auf ein anderes Gleis geleitet wurde.

Message Eine Message besteht aus einer Message-Id, einem Message-Typ und einen Timestamp. Des Weiteren können noch folgende Informationen angehängt werden: Eine Information auf welche Uhrzeit der Zug verlegt wurde, aber auch wann der Zug eigentlich geplant war. Ein Code um die Message zu identifizieren, den Text der Nachricht, die Kategorie der Nachricht, die Priorität, der Eigentümer, ein externer Link, der Indikator ob die Nachricht gelöscht ist, eine Nachricht des Verteilers, sowie der Name des Zuges.

2.2 Planung

Zeitliche Einteilung, beachten 5. Semester ist weniger Zeit, Hauptteil wird im 6. Semester passieren

Kurzplan: 5.Semester Grundlagen Maschine Learning und Data Mining. 6.Semester sollen die Daten aus dem 5.Semester verarbeitet werden, diese Datensätze sollen genutzt werden, um später das neuronale Netz zu erstellen. Hauptaugenmerk soll daher im 6. Semester der Aufbau und die Erstellung eines funktionierenden neuronalen Netzes sein, welches zum Vorhersagen der Zugverspätung genutzt werden kann.

2.3 Data Mining

Data Mining Einführung und dessen Bedeutung für das Projekt

Data Mining ist ein wichtiger Bestandteil des Projektes, ohne die Daten kann dieses Projekt nicht funktionieren. Denn um ein neuronales Netz zu trainieren, sind Unmengen an Daten nötig. Als Faustregel gilt, je mehr Daten, desto genauer das neuronale Netz. Zum Speichern der Datensätze sollte ein offenes weiterverwendbares Format genutzt werden. Dies soll zudem der weiteren Automatisierbarkeit des Datenflusses dienen.

Datenformat und Aufbau erklären. Wieso sollte im ersten Schritt beim Mining nicht direkt alles angepasst werden? Wieso müssen die Daten aufbereitet werden? Stichwort: FehlerAPI, Fehlende Datensätze, Bucketlist, Konvertierung

Dinge die wir brauchen:

- Bahnhaltsnummer

- Linie als Folge von angefahrenen Bahnhöfen (z.B. ICE 690, EC 378, R856)
- Zugreferenz (gleicher Zug auf Linie?)
- Ankunftszeit geplant
- Ankunftszeit real
- Abfahrtszeit geplant
- Abfahrtszeit real
- Historic Delay Element?
Angeblich kann man damit die vorherigen Verspätungen auf der Linie auslesen
- Wetter je PLZ[Postleitregionen] (Wind, Niederschlag, Temperatur)
- Die Bahnhof Tabelle mit PLZ ergänzen, um Wetterdaten zuordnen zu können (Postleitregionen)

Mögliche Auswertungen:

- Relative Verspätung pro Streckenabschnitt
Pro Streckenabschnitt kann ein Zug Verspätung aufbauen oder abbauen. Jedem Streckenabschnitt wird die Summe aller Verspätungen, die die Züge auf diesem Streckenabschnitt aufbauen oder abbauen zugeordnet. Diese Summe aller relativen Verspätungen pro Streckenabschnitt wird anschließend visualisiert.
- Verzögerung im Bahnhof
Pro Bahnhof kann die Verspätung eines Zuges zunehmen oder abnehmen. Pro Bahnhof werden von allen Zügen die Verspätungen, die sie in dem jeweiligen Bahnhof aufbauen oder abbauen, aufsummiert. Anschließend wird für jeden Bahnhof die gebildete Summe visualisiert.
- Welche Wetterlagen bringen Verspätungen

Mögliche Arten der Visualisierung

- Welche Strecken bringen die meiste Verspätung? Heatmap? Top10?
- Welche Bahnhöfe haben die größte Verzögerung? Heatmap? Top 10? Diagramm?

Auswahl der Wetterstationen: Die Wetterstationen werden pro Postleitregion so gewählt, sodass diese möglichst im Zentrum der jeweiligen Region liegen.

2.4 Datenmodell

Datenmodell erläutern, welche Rohdaten aus der DB-API

Ein Datenmodell ist sowohl erforderlich, um Datenobjekte bezüglich ihrer Bedeutung zu interpretieren, als auch, um Beziehungen zwischen Datenobjekten festzustellen oder zu beschreiben. Im Rahmen dieser Arbeit gilt es, ein Datenmodell zu definieren, das unterschiedliche Aufgaben erfüllen soll:

Modellierung realer Größen Zu aller erst definiert das Datenmodell die Modellierung von Größen der realen Welt, die später für die folgende Datenverarbeitung benötigt werden. Hierbei werden mathematische Definitionen entwickelt, die die Bedeutung der jeweiligen Größe, wie zum Beispiel Verspätung, beschreibt.

Modellierung der Rohdaten Anschließend definiert das Datenmodell, wie die beschriebenen Größen der realen Welt in den Rohdaten abstrahiert und abgebildet werden. Dies ist wichtig, um die Rohdaten, wie sie beispielsweise von der Timetable-API der Deutschen Bahn geliefert werden, interpretieren und weiterverarbeiten zu können. Insbesondere muss die Modellierung die Beziehungen unter den Datenobjekten der Rohdaten definieren, um aus diesen wieder die realen Größen ableiten zu können.

Modellierung der Auswertung Nachdem die Bedeutung von realen Größen und deren Abbildung in den Rohdaten definiert ist, muss die Auswertung der Daten konzipiert und modelliert werden. Hierzu zählen sowohl die Beschreibung der internen Darstellung der Daten zum Zwecke der weiteren Auswertung als auch die Beschreibung des auswertenden Algorithmus. Zu den internen Darstellungen können Datenstrukturen in Programmen oder Datenbank-Schemata gezählt werden.

Um die Gliederung der Arbeit übersichtlich zu halten, sind die Modellierungen der oben genannten Punkte in separaten Kapiteln dargestellt.

2.5 Modellierung realer Größen

In diesem Abschnitt werden die realen Größen, die zur Datenauswertung benötigt werden, modelliert. Eine der wichtigsten realen Größen in dieser Arbeit ist die Verspätung oder Verzögerung von Zügen. Im folgenden werden die verschiedenen Arten von Verzögerungen dargestellt.

Ankunftsverzögerung Die Verzögerung der Ankunft Δan eines Zuges zug_n im Bahnhof bhf_m ist definiert als

$$\Delta an(bhf_m, zug_n) := an_{real}(bhf_m, zug_n) - an_{plan}(bhf_m, zug_n) \quad (2.1)$$

Abfahrtsverzögerung Die Verzögerung der Abfahrt Δab eines Zuges zug_n im Bahnhof bhf_m ist definiert als

$$\Delta ab(bhf_m, zug_n) := ab_{real}(bhf_m, zug_n) - ab_{plan}(bhf_m, zug_n) \quad (2.2)$$

Bahnhof-respektive Verzögerung (BRV)

$$brv(bhf_m, zug_n) := \Delta ab(bhf_m, zug_n) - \Delta an(bhf_m, zug_n) \quad (2.3)$$

Anhand der BRV lässt sich erkennen, ob der Zug Verspätung während dem Verweilen in dem Bahnhof aufbaut. Ist die BRV positiv, so nimmt die Verspätung des Zuges durch die außerplanmäßige verlängerte Haltedauer zu. Ist die BRV hingegen negativ, so verringert sich die Verspätung des Zuges durch eine verkürzte Haltedauer im Bahnhof. Ist $brv = 0$, so entspricht die Haltedauer des Zuges der geplanten Haltedauer.

Bahnhof-respektiver Verzögerungsdurchschnitt (BRVD)

$$brvd(bhf_m, Z) := \sum_{i=0}^n \frac{brv(bhf_m, z_i)}{n} \quad (2.4)$$

Der BRVD berechnet den Durchschnitt des BRV bezüglich einer Zugmenge Z . Mithilfe des BRVD lässt sich interpretieren, wie stark die Züge im Durchschnitt durch den Halt in dem jeweiligen Bahnhof verzögert werden.

Streckenabschnitt-respektive Verzögerung (SARV)

$$sarv(bhf_{ab}, bhf_{an}, zug_n) := \frac{[an_{real}(bhf_{an}, zug_n) - ab_{real}(bhf_{ab}, zug_n)] - [an_{plan}(bhf_{an}, zug_n) - ab_{plan}(bhf_{ab}, zug_n)]}{n} \quad (2.5)$$

Die SARV erlaubt es festzustellen, ob der Zug auf dem Streckenabschnitt von den letzten Bahnhof zum nächsten Bahnhof Verzögerung aufbaut oder abbaut.

Mit den oben definierten Verzögerungen ist es bereits möglich, erste statistische Auswertungen auszuführen über die Verspätung von Zügen, die sich entweder in Bahnhöfen oder auf den Strecken zwischen Bahnhöfen ereignen.

2.6 Aufbereitung von Daten

Wie werden Daten aufbereitet, vorbereitet für das neuronale Netz, welche Dinge gibt es zu beachten (DATENTYPEN!)

Bei der Aufbereitung der Datensätze geht es die Vorhandenen Daten aufzuteilen, zu kategorisieren, zu formatieren und vorzubereiten. Da im nächsten Schritt ein neuronales Netz trainiert und geprüft werden soll, ist eine Aufteilung der Datensätze in diese drei Kategorien sinnvoll. Später kann dann der Echtzeit Datensatz vorhergesagt werden.

2.7 Neuronale Netze an simplen Beispielen erklärt

Kleine Einleitung an einem Simplen Beispiel, Linear Regression oder so. Wieso wir sowas brauchen und weshalb es von Relevanz ist.

2.8 Eingabe- und Ausgabe-Parameter für das Neuronale Netz

Erläuterung welche Informationen in das Neuronale Netz eingegeben werden und welche Daten von dem Netz ausgegeben werden.

Endnutzereingaben: Startbahnhof Zielbahnhof Einsteige-Zeit Zugeingabe (welcher Zug genau?)

Eingabe: Zug-ID Ziel-Bahnhof Um Vorraussagen treffen zu können, braucht das neuronale Netz noch zusätzliche Informationen: Strecke des Zuges? Vergangene Fahrten des Zuges und dessen Verspätung?

Zug-ID

Soll-Ankunftszeit des Zuges

Ausgabe: Voraussichtliche Verspätung in Minuten

2.9 Literaturhinweise und Empfehlungen

Weiterführende Literatur sollte bis zum Abschluss erwähnt werden, verwendete Quellen zum Einlesen in neuronale Netze und gute Erklärungen, event. Zitate auch benutzen. Diese Autoren sind sehr wichtig für dieses Projekt und sollte auch genannt werden.

Kapitel 3

Datenbeschaffung

3.1 Programmierung des Data Miners

Der Data Miner wird im Laufe der Studienarbeit immer weiter entwickelt und stetig verbessert. Die erste Version zeigte nach nur wenigen Wochen erhebliche Schwachstellen im Quellcode auf. Die erste Version des Data Miners besitzt folgende Funktionen:

- Bahn API aufrufen
- Daten ungeprüft in Datenbank schreiben

Durch die geringere Datenmenge (anstatt der 6600 Stationen wurden nur 1200 abgerufen) konnte die Umsetzung schnell realisiert werden. Da es sehr viele Optionen und Probleme gab, wurde die erste Version nach etwa

Anzahl Wochen

Wochen durch die zweite Version des Miners ersetzt. Diese besitzt neben neuen Funktionen auch die Erweiterung der vollständigen Abfrage der API. Die zweite Version konnte die Probleme auf der Seite des Miners minimieren. Die zweite Version kann zudem alle Daten abfragen und nutzt deutlich mehr Informationen, welche in der API der Bahn bereitgestellt werden. Die wichtigste Änderung ist die Fehlererkennung in der Abfrage von Datensätzen. Dadurch soll ein übermäßiges Fehlen von Datensätzen vermieden werden. Die zweite Version des Data Miners ist in der Lage über 600.000 Datensätze am Tag zu verarbeiten. Für die dritte Version des Data Miners wird ein Webinterface zur besseren Überwachung und Monitoring des Miners geplant. Des Weiteren soll eine grundlegende Refaktorisierung vorgenommen werden, um den Quellcode in Zukunft besser wartbar und strukturierter zu machen.

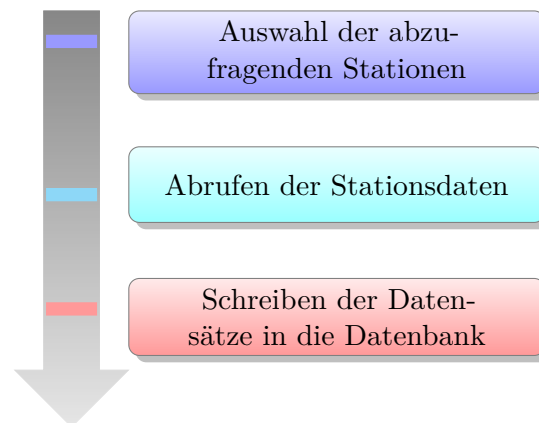
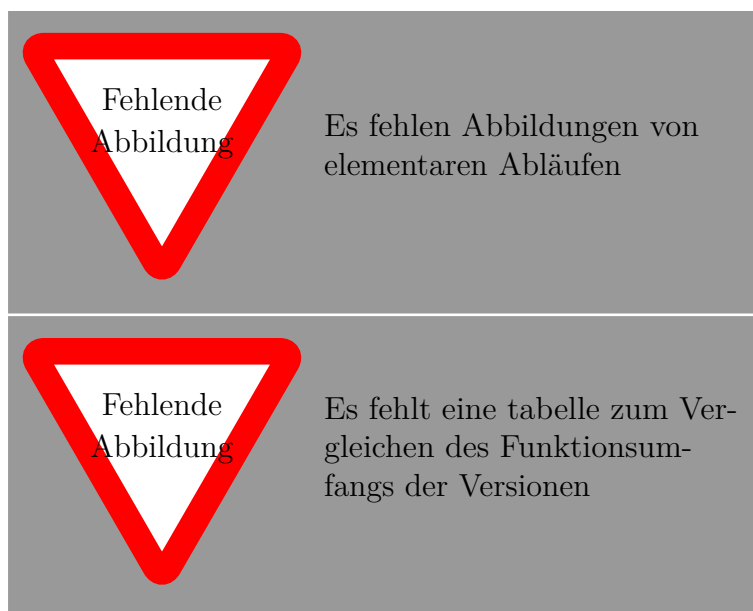


Abbildung 3.1: Grundablauf des Miners



3.2 Datenbank und Schema

Wie werden die Datensätze abgespeichert und verwaltet? Das Schema der Datenbank befindet sich im Ressourcen Ordner

Ein wichtiger Bestandteil des Projektes ist neben dem Abrufen der API das dauerhafte Abspeichern von Datensätzen. Die Struktur dieser Datensätze hat sich mit der Entwicklung des Data Miners ebenfalls verändert. Es werden mit der zweiten Version deutlich mehr Informationen aus der API abgespeichert. Ein Datensatz benötigt in der ersten Version 140 Bytes und in der zweiten Version 320 Bytes. Viele der neuen Informationen sind für die spätere Arbeit sehr wahrscheinlich wichtig, daher wurden diese in der zweiten Version des Miners ausgewählt. So kann nun der Verlauf eines Zuges besser verfolgt werden und es werden Informationen zum Zugstatus und der Pünktlichkeit strikt getrennt. In Abbildung

x.y

ist das Schema von der ersten Version abgebildet.

Hier etwas darüber erläutern

In Abbildung

x.y

ist dagegen das Schema der zweiten Version zu sehen. Dieses Schema besitzt deutlich mehr Spalten pro Datensatz und benötigt daher auch etwas mehr Speicherplatz. Trotzdem beträgt die Größe der Datenbank nach mehr als 20 Millionen Datensätzen unter 6 Gigabyte. Ein wichtiger Punkt hierbei ist die Menge an Datensätzen. In der Literatur gilt häufig die Faustregel, je mehr Datensätze, desto besser kann das neuronale Netz trainiert werden.

Literatur verweise einfügen

In wie weit diese Aussagen auf dieses Projekt zutreffen wird in Kapitel

x.y

geprüft.



Es fehlt die Abbildung für
Schema Version 1



Es fehlt die Abbildung für
Schema Version 2

Kapitel 4

Datenverarbeitung mit Data Mining

4.1 Stochastische Analyse

4.2 Visualisierung

Kapitel 5

Datenverarbeitung mit neuronalem Netz

5.1 Programmierung der Automatischen Datenverarbeitung

Wie kommen die Datensätze aus der DB zum neuronalen Netzwerk, wie wird die Formatierung vorgenommen

5.2 Vorverarbeitung der Datensätze

Kurze Einführung schreiben

id Id als Primärschlüssel zur Speicherung in der Datenbank.

zugid Beispiel: **-7714364757423921343-1712081222-8**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

zugverkehrstyp Beispiel: **F**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

zugtyp Beispiel: **p**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

zugowner Beispiel: **80**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

zugklasse Beispiel: **ICE**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

zugnummer Beispiel: **788**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

zugnummerfull Beispiel: **ICE788**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

linie Beispiel: **–leerer String–**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

evanr Beispiel: **8000152**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

arzeitsoll Beispiel: **16:32:00**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

arzeitist Beispiel: **16:33:00**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

dpzeitsoll Beispiel: **16:36:00**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

dpzeitist Beispiel: **16:38:00**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

gleissoll Beispiel: **7**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

gleisist Beispiel: **7**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

datum Beispiel: **2017-12-08**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

streckengeplanthash Beispiel: **4d0bc383**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

streckenchangedhash Beispiel: **bd84c25a**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

zugstatus Beispiel: **n**

Zug-Id die von der Timetable-API zur Zugidentifizierung genutzt wird.

Hier die `generate_csv.py` beschreiben

Die Datensätze aus der Datenbank müssen vor dem Import in das neuronale Netz preprocessed werden. Da Tensoren nur aus numerischen Typen bestehen sollten. Hierfür werden kombinierte Datentypen getrennt und in einen passenden Zieltypen konvertiert. Bei manchen Typen ist es jedoch besser, die Möglichen Werte in ein Vocabfile zu schreiben, um diese mit einer $n \times n$ Identitätsmatrix im Tensor darzustellen. Alle Datensätze sollen bevor sie in die einzelnen CSV Dateien geschrieben werden in ein einheitliches Format gebracht werden. Ziel ist es Training, Test und Prediction in einem Schritt umzusetzen.

Datum	Beispiel	Datenbank Datentyp	Konvertierter Datentyp (Python)
id	4092195	VARCHAR	
zugid	-7714364757423921343-1712081222-8	VARCHAR	
zugverkehrstyp	F	VARCHAR	
zugtyp	p	VARCHAR	
zugowner	80	VARCHAR	
String zugklasse	ICE	VARCHAR	
String zugnummer	788	VARCHAR	
zugnummerfull	ICE788	VARCHAR	
linie	#leerer String#	VARCHAR	
String evanr	8000152	INT	
arzeitsoll	16:32:00	TIME	
IntType arzeitist	16:33:00	TIME	
IntType dpzeitsoll	16:36:00	TIME	
IntType dpzeitist	16:38:00	TIME	
IntType gleissoll	7	VARCHAR	
String gleisist	7	VARCHAR	
String datum	2017-12-08	DATE	
String streckengeplanthash	4d0bc383	VARCHAR	
streckenchangedhash	bd84c25a	VARCHAR	
zugstatus	n	VARCHAR	

Tabelle 5.1: Vorverarbeitung der Datenbank-Daten

5.3 Programmieren der Modellstruktur des neuronalen Netzes

Wie ist das neuronale Netz strukturiert, welche Neuronen-Architekturen verwenden wir und weshalb, wie finden wir das beste Netz

5.4 Vermeidung von Overfitting und Anpassungen um die Genauigkeit zu erhöhen

Overfitting ist häufig ein Problem wie erkennt man es und wie kann man overfitting vermeiden.

Overfitting ist die zu hohe Genauigkeit, welche es nicht mehr ermöglicht, eventuelle Fehler sinnvoll zu erkennen. Diese Fehler werden als korrekte Daten angesehen. Häufig ist Overfitting an einer starken Schwankung in der Genauigkeit beim Trainieren des Netzes zu erkennen.

5.5 Anlernen des Netzes

Welche Datensätze werden zum Anlernen verwendet, weshalb ist es wichtig nie alle zu nehmen im Bezug auf Test, Predict und welche Verhältnisse sind bei uns sinnvoll anzusetzen

Aufzeigen wie sich die Menge an Daten auf die Genauigkeit auswirkt

Welche Optionen und Parameter können optimiert werden, wie ändert sich dadurch das Ergebnis.

Hier Tabellen mit Vergleich der Methoden und Genauigkeit, Geschwindigkeit, Erläuterungen weshalb das Ergebnis so ist.

5.6 Verifizieren des Netzes

Testen des neuronalen Netzes, Verifikation der Genauigkeit und deren Steigerung durch Training oder Anpassungen des Netzes

5.7 Vorhersagen anhand des Netzes

Vorhersagen aus Daten treffen und anschauen wie gut sie sind, wo gibt es Probleme, welche Probleme treten auf.

5.8 Auswertung und Fehlerbehandlung

Was passiert im Fehlerfall, wie erkennt man Fehler, müssen wir Fehler erkennen oder sind Fehler egal", wie stellen wir eine GUI bereit, um anderen Menschen die Ergebnisse zu testen, genauere Statistiken zu Zügen je nach Strecke, Uhrzeit etc., vlt. Visuelle Darstellung wie bei Travic oder mit eigenen Heatmaps bzw. Openstreetmap.

Kapitel 6

Schlussfolgerung

6.1 Rückblick

Was ist geschehen, was würden wir anders machen, was waren wichtige Schritte





















6.2 Fazit

Ergebnis der Studienarbeit, was war gut, was war schlecht, hat alles so geklappt, wo gab es Probleme, wie wurden diese gelöst (kurz und knapp zusammengefasst).

6.3 Ausblick

Wie geht es weiter, könnte es weiter gehen, was sollte verbessert werden, wo befinden sich Schwachstellen, event. ungelöste Probleme

Liste der ToDo's

	Sperrvermerk ja oder nein	1
	Dieses Abstract besser schreiben und eventuell eine englische Übersetzung anfertigen	1
	Wieso wollen wir das machen und warum ist das für uns wichtig.	7
	Hier etwas zum Stand der Technik schreiben, neuronale Netze, Tensorflow, KI, Datamining, OpenData, etc	7
	Hier das Ziel aus der Anmeldung schön definieren und klar Abgrenzen was Ziel und was optional nice to have ist.	7
	Was bekommen wir eigentlich alles über die Api geliefert	9
	Zeitliche Einteilung, beachten 5. Semester ist weniger Zeit, Hauptteil wird im 6. Semester passieren	11
	Data Mining Einführung und dessen Bedeutung für das Projekt	11
	Datenformat und Aufbau erklären. Wieso sollte im ersten Schritt beim Mining nicht direkt alles angepasst werden? Wieso müssen die Daten aufbereitet werden? Stichwort: FehlerAPI, Fehlende Datensätze, Bucketlist, Konvertierung	11
	Datenmodell erläutern, welche Rohdaten aus der DB-API	13
	Wie werden Daten aufbereitet, vorbereitet für das neuronale Netz, welche Dinge gibt es zu beachten (DATENTYPEN!)	14
	Kleine Einleitung an einem Simplen Beispiel, Linear Regression oder so. Wieso wir sowas brauchen und weshalb es von Relevanz ist.	15
	Erläuterung welche Informationen in das Neuronale Netz eingegeben werden und welche Daten von dem Netz ausgegeben werden.	15
	Weiterführende Literatur sollte bis zum Abschluss erwähnt werden, verwendete Quellen zum Einlesen in neuronale Netze und gute Erklärungen, event. Zitate auch benutzen. Diese Autoren sind sehr wichtig für dieses Projekt und sollte auch genannt werden.	15
	Anzahl Wochen	16
	Abbildung: Es fehlen Abbildungen von elementaren Abläufen	17
	Abbildung: Es fehlt eine tabelle zum Vergleichen des Funktionsumfangs der Versionen	17
	Wie werden die Datensätze abgespeichert und verwaltet? Das Schema der Datenbank befindet sich im Ressourcen Ordner	17
	x.y	17
	Hier etwas darüber erläutern	18
	x.y	18
	Literatur verweise einfügen	18

■ x.y	18
Abbildung: Es fehlt die Abbildung für Schema Version 1	18
Abbildung: Es fehlt die Abbildung für Schema Version 2	18
■ Wie kommen die Datensätze aus der DB zum neuronalen Netzwerk, wie wird die Formatierung vorgenommen	20
■ Kurze Einführung schreiben	20
■ Hier die generate_csv.py beschreiben	21
■ Wie ist das neuronale Netz strukturiert, welche Neuronen-Architekturen verwenden wir und weshalb, wie finden wir das beste Netz	23
■ Overfitting ist häufig ein Problem wie erkennt man es und wie kann man overfitting vermeiden.	23
■ Welche Datensätze werden zum Anlernen verwendet, weshalb ist es wichtig nie alle zu nehmen im Bezug auf Test, Predict und welche Verhältnisse sind bei uns sinnvoll anzusetzen	23
■ Aufzeigen wie sich die Menge an Daten auf die Genauigkeit auswirkt	23
■ Welche Optionen und Parameter können optimiert werden, wie ändert sich dadurch das Ergebnis.	23
■ Hier Tabellen mit Vergleich der Methoden und Genauigkeit, Geschwindigkeit, Erläuterungen weshalb das Ergebnis so ist.	23
■ Testen des neuronalen Netzes, Verifikation der Genauigkeit und deren Steigerung durch Training oder Anpassungen des Netzes	23
■ Vorhersagen aus Daten treffen und anschauen wie gut sie sind, wo gibt es Probleme, welche Probleme treten auf.	23
■ Was passiert im Fehlerfall, wie erkennt man Fehler, müssen wir Fehler erkennen oder sind Fehler egal", wie stellen wir eine GUI bereit, um anderen Menschen die Ergebnisse zu testen, genauere Statistiken zu Zügen je nach Strecke, Uhrzeit etc., vlt. Visuelle Darstellung wie bei Travic oder mit eigenen Heatmaps bzw. Openstreetmap.	24
■ Was ist geschehen, was würden wir anders machen, was waren wichtige Schritte	25
■ Ergebnis der Studienarbeit, was war gut, was war schlecht, hat alles so geklappt, wo gab es Probleme, wie wurden diese gelöst (kurz und knapp zusammengefasst.	25
■ Wie geht es weiter, könnte es weiter gehen, was sollte verbessert werden, wo befinden sich Schwachstellen, event. ungelöste Probleme	25