

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	Einleitung	7
1.2	Motivation	7
1.3	Stand der Technik	8
1.4	Ziel der Studienarbeit	8
1.5	Begriffsdefinitionen	9
2	Grundlagen	10
2.1	Die DB Timetable API	10
2.1.1	Station	10
2.1.2	Plan	11
2.1.3	fchg	11
2.1.4	rchg	11
2.2	Planung	12
2.2.1	Zeitliche Einteilung der Studienarbeit	12
2.2.2	Versionsverwaltung	12
2.3	Data Mining	13
2.4	Datenmodell	14
2.5	Modellierung realer Größen	15
2.6	Aufbereitung von Daten	16
2.7	Neuronalen Netzen an simplen Beispielen erklärt	16
2.8	Eingabe- und Ausgabe-Parameter für das Neuronale Netz	16
2.9	Einrichten der Tensorflow Umgebung	16
2.10	Literaturhinweise und Empfehlungen	17
3	Datenbeschaffung	18
3.1	Programmierung des Data Miners	18
3.2	Datenbank und Schema	21
3.3	Backup der Datenbank	23
4	Datenverarbeitung mit Data Mining	24
4.1	Vorverarbeitung der Daten	24
4.2	Stochastische Analyse	25
4.3	Visualisierung	25

5	Datenverarbeitung mit neuronalem Netz	26
5.1	Programmierung der Automatischen Datenverarbeitung	26
5.2	Vorverarbeitung der Datensätze	26
5.3	Begriffsdefinitionen für ein neuronales Netz	29
5.4	Eingabe der Datensätze in Tensorflow	30
5.5	Anlernen des Netzes	30
5.6	Verifizieren des Netzes	30
5.7	Vorhersagen anhand des Netzes	31
5.8	Auswertung und Fehlerbehandlung	31
6	Visualisierung und Bereitstellung der Daten im Internet	32
6.1	Aufbau der Website	32
6.2	Visualisierung der Datensätze	33
7	Schlussfolgerung	34
7.1	Rückblick	34
7.2	Fazit	34
7.3	Ausblick	34
	Anhang	35
	Literaturverzeichnis	35
	Liste der ToDo's	35

Abbildungsverzeichnis

3.1 Grundablauf des Miners	19
--------------------------------------	----

Tabellenverzeichnis

5.1	Vorverarbeitung der Datenbank-Daten	28
-----	---	----

Liste der Quellcodeausschnitte

3.1	Drei Ausschnitte aus einer Datei	20
4.1	Some Python File	25

Abkürzungsverzeichnis

HTTP	Hypertext Transfer Protocol	21
BRV	Bahnhof-respektive Verzögerung	15
BRVD	Bahnhof-respektiver Verzögerungsdurchschnitt	15
SARV	Streckenabschnitt-respektive Verzögerung	15

Kapitel 1

Einleitung

1.1 Einleitung

Wie kam es dazu, eventuell mit Motivation kombinieren.

Mit der Bereitstellung der Livefahrplandaten durch die Open Data Bewegung der deutschen bahn¹ kam der Gedanke mit den nun vorhandenen Daten etwas anzufangen. Zuallererst wird eine Diskussion geführt, welche das Ziel der Studienarbeit in der Zukunft bestimmen soll. Aufgrund der Diskussion wurde die Anmeldung und somit die Basis der Studienarbeit geschaffen. Ein Bestandteil der Anmeldung sieht eine Aufbereitung der Rohdaten vor. Hierzu muss eine passende Software entwickelt werden. Die Analyse der Daten soll durch verschiedene Parameter vorgenommen werden, so sollen etwa Verspätungen nach Zeit, Ort oder Strecke betrachtet und ausgewertet werden.

1.2 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. Dafür soll zuerst eine statistische Auswertung der im Laufe der Studienarbeit gesammelten Daten durchgeführt werden. Die Auswertung soll die Daten mit Zusammenhang auf ihre Relevanz visualisieren und entsprechend aufbereiten. 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. Durch die gewünschte Erkennung von Regelmäßigkeiten und deren Einflüsse soll die Reisedauer verringert werden.

¹Siehe URL der DB Seite einfügen

1.3 Stand der Technik

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

Derzeit ist der Begriff: Maschinelles Lernen ein wichtiger Punkt im Fortschritt von Software. In dieser Studienarbeit sollen verschiedene Disziplinen von maschinellem Lernen über Data Mining und Visualisierungstechniken bis hin zur Bereitstellung der Ergebnisse behandelt werden. Es ist wichtig vor Beginn der Arbeit die Gebiete voneinander abzugrenzen, um die Bearbeitung in kleineren Schritten durchzuführen. Eine gewisse Reihenfolge muss dabei beachtet werden, weshalb im ersten Abschnitt der Studienarbeit auf die Grundlagen eingegangen wird. Zuerst muss der Begriff der Datenbeschaffung und des Data Minings geklärt werden.

Hier Zitat aus Buch Definition zu datamining und datenbeschaffung einfügen

Erst nach der Beschaffung können die Daten in Zusammenhang gebracht werden. Die sinnvolle Visualisierung der Datensätze ist sehr wichtig, um eventuelle Zusammenhänge besser erkennen zu können.

1.4 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)

1.5 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

2.2.1 Zeitliche Einteilung der Studienarbeit

Im ersten Teil der Studienarbeit steht die Erfassung der Daten der deutschen bahn im Vordergrund des programmieraspektes. Neben der Programmierung des Data Miners für die Bahn API wird Literatur, welche für die anschließende Aufbereitung und Visualisierung der Datensätze benötigt wird, gelesen. Da die Modulwahl des Teams im fünften Semester eine deutlich höhere zeitliche Belastung durch die Vorlesungseinheiten ausweist, wurde ein Großteil der Hauptarbeit in das sechste Semester verlegt.

Hier Gantt Diagramm oder Tabelle einfügen mit was wurde in welchem Semester gemacht.

2.2.2 Versionsverwaltung

Zur Planung gehört neben der zeitlichen Planung auch die Planung, wie der entstandene Quellcode und die Studienarbeit als Dokument einer Versionsverwaltung unterzogen wird. Die Entscheidung der Gruppe fiel auf Github

eventuell Verlinken

, da damit bereits gute Erfahrungen gemacht wurde. Dort wird eine öffentliche Organisation angelegt, welcher alle Gruppenteilnehmer beitreten. In der Organisation werden die Repositories zur Verwaltung von Website, Data Miner, Visualisierungstoolkit und

Dokumentation angelegt. Alle Teilnehmer bekommen Zugriff auf den Gesamten Quellcode. Damit ist gleichzeitig Backup und ein aktueller Wissensaustausch zwischen den Teilnehmern sichergestellt. Die gemeinsame Arbeit an Quellcode wird durch die Versionsverwaltung erleichtert, da parallel in verschiedenen Branches gearbeitet 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:

- Bahnhofsnummer
- 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.

Achtung siehe Begriffsdefinitionen von neuronalen Netzen, dieses Kapitel vllt hier her

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.

Achtung eventuell doppelter Eintrag siehe spätere Kapitel.

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

Eingabe: Zug-ID Ziel-Bahnhof Um Voraussagen 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 Einrichten der Tensorflow Umgebung

Was wird alles für Tensorflow benötigt

Eventuell how to install tf verlinken

Bevor ein neuronales Netz mit Tensorflow realisiert wird, muss die Umgebung auf den jeweiligen Computern eingerichtet werden. Hier unterscheiden sich die Schritte der Einrichtung je nach Betriebssystem. Unter Windows wird die Einrichtung von Python 3.5.2+ via Installer fertiggestellt. Daraufhin wird mit PIP Installs Packages

Verweis einfügen

das Packet von Tensorflow heruntergeladen und installiert. Daraufhin steht die Grundversion von Tensorflow dem Nutzer bereit. Da Tensorflow vor allem durch eine GPU beschleunigt wird, sollte bei der Verwendung als langfristige Umgebung, die GPU Unterstützung installiert werden. Dies spart vor allem Zeit und somit auch unnötigen Leerlauf beim ausprobieren eines neuen Modells. Unter Linux müssen die Schritte ebenfalls vorgenommen werden, da jedoch die Unterstützung für Linux Server bereits vorhanden ist, wird die Installation vereinfacht und benötigt mehrere Stunden weniger, im Falle einer Fehlersuche. Unter Windows gab es beim einrichten des GPU Support unerwartete Probleme mit den Systemumgebungsvariablen, wodurch die Treiber für die Grafikkarte nicht geladen werden konnten. Da jedoch keine hilfreiche Fehlermeldung erschien, musste die Installation manuell verifiziert werden. Nach diesen Schritten kann Tensorflow mit und ohne GPU Unterstützung auf den Rechnern ausgeführt werden. Ein kurzer Vergleich zeigte, dass die Geschwindigkeit bei Berechnungen mit der Grafikkarte verzehnfacht.

2.10 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 eine 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. Zu Beginn gab es jedoch noch Probleme mit den aus der API Dokumentation erhalten Datenstrukturen, so sollte ein Gleis angeblich ein Integer sein. Dies trifft jedoch im Falle von "3 A-G", also Gleis 3 Abschnitt A bis G nicht zu. Daher musste die Datenbankspalte für das Gleis angepasst werden. Ebenfalls von Fehlern betroffen war die Zugnummer, diese sollte eine gewisse Länge nicht überschreiten, es gab jedoch Zugnummern mit einer Ziffer zu viel, dadurch konnten Anfangs nicht alle Züge gespeichert werden.

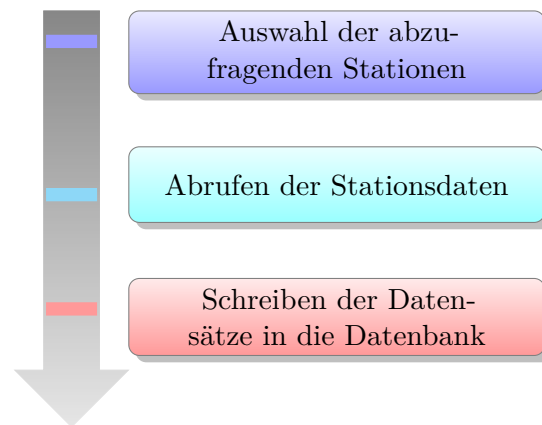


Abbildung 3.1: Grundablauf des Miners

```

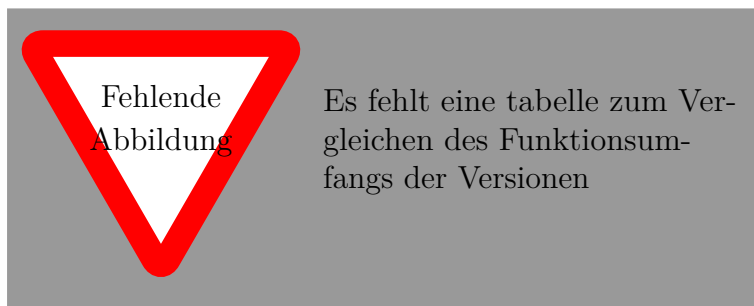
1 <?php
2
3 include_once './settings.php';
4 require_once 'classes/MysqliDb.php';
5 require_once 'classes/appgati.php';
6 // currently not used maybe later
7 } else {
8     $bahnapi = new bahnapi($apikey2);
9 }
10
11
12 // Using old querie here because limit dosnt seem to work in
    rawquery
13 $params = date("Y-m-d_H:i:s", time() - 3600);
14 $mysqlslave = new mysqli(SETTING_DB_IP, SETTING_DB_USER,
    SETTING_DB_PASSWORD, SETTING_DB_NAME);
15
16 if($minute == 0 || $minute == "00" || $minute == "0") {
17     $stationsquery = $mysqlslave->query("UPDATE haltestellen2
        set fetchtime='2017-12-01 00:00:00'"); // all stations
        should be fetched
18
19     // Insert twitter fetch here last 200 tweets lasted over 3
        0 days...
20
21
22 }
23
24 $stationsquery = $mysqlslave->query("SELECT EVA_NR as nr,
    NAME FROM haltestellen2 WHERE fetchactive2=1 AND fetchtime <
    '$params' ORDER BY fetchtime ASC LIMIT 0,135");
25
26 $station = array();
27 while ($row = $stationsquery->fetch_assoc()) {

```

Quellcode 3.1: Drei Ausschnitte aus einer Datei



Es fehlen Abbildungen von
elementaren Abläufen



Die zweite Version des Miner kann zudem mit den HTTP Status Codes automatisch erkennen, ob es auf der Seite der API gerade ein Problem gibt. So wird auch erkannt, dass es Abends öfter zu kurzen Ausfällen der API mit dem Hypertext Transfer Protocol (HTTP) Statuscode 503

Hier noch was bedeutet 503 und eventuell zitat aus RFC
<https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html>

kommt. Dies hilft herauszufinden, ob ein Fehler auf der Seite der BahnAPI oder des Miners vorliegt. Auch ein häufiger Fehler der fehlerhaften initialisierung von Variablen wurde behoben.

Nach der Migration des Miners auf eine größere und schnellere Seite wird die Performance der Datenbank erheblich verbessert. Die Datenbank profitiert hier vor allem von deutlich mehr Arbeitsspeicher (anstatt 16 GigaByte nun 64 GigaByte), um Abfragen zwischenspeichern. Des weiteren ist der Miner nun IPv6 fähig, da der alte Hostserver noch keine eigene IPv6 Adresse hatte. Dies sichert die Funktionalität im Falle einer IPv6 Umstellung der API Schnittstellen.

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

Bei dem Umzug des Data Miners samt Datenbank auf einen neuen Server mussten zehn GigaByte an Datenbank migriert werden. Dies erwies sich als komplizierter als angenommen, denn zum einen Dauert der Export und Import mehrere Stunden und zum anderen müssen die nicht exportierten Einträge des Miners in der Zeit des Umzuges mit dem neuen Server synchronisiert werden. Dies ist bei einer Datenbanktabelle, welche dauerhaft mehrere Transaktionen des Miners bekommt sehr mühsam umzusetzen. Um den Prozess so schonend wie möglich zu machen, wurde ein Skript geschrieben, welches nach der fertigen Migration der Datenbank die Tabellen miteinander Synchronisiert, da ein MySQL Sharding mit Master- und Slave-Modus aufgrund inkompatibler Versionen nicht möglich war. Nachdem das Skript die Tabellen synchronisiert hatte, wurde der alte Miner gestoppt und der Miner auf dem neuen Server gestartet. Die Downtime des Miners betrug nur etwa 60 Sekunden, danach wurde noch einen Fehler in der Installation entdeckt, die durch die Anpassung .

3.3 Backup der Datenbank

Datenbanken sind toll, aber es muss bei einer kritischen Stelle ein Backup vorhanden sein.

Jeder Datensatz des Data Miners ist wichtig. Daher soll für diese kritische Stelle, der persistenten Speicherung der Daten ein automatisiertes und verifizierbares Backup entstehen. Hierbei gibt es zwei Hauptprobleme zu lösen. Zum einen muss während des Backup eine große Transaktion im Cache oder auf der Festplatte zwischengespeichert werden. Zum anderen ist durch die Menge an Datensätzen eine manuelle Verifikation, ob die Datensätze auch wieder einspielbar sind sehr aufwändig. Daher wird ein kleines Skript geschrieben, welches mit linearem Aufwand (Größe der Datei) die Datensätze an bestimmten Stellen aufsplittet. So entstehen viele kleinere Dateien. Diese können in unter einer Minute mit einem Datenbank Import auf funktionierende Constraints und geprüft werden. Dies ermöglicht nach einem vollständigen Backup die einzelnen Dateien automatisiert nach und nach in einer kleineren Datenbank zu prüfen. Sollte ein Fehler auftreten, wird dieser in dem MySQL eigenen Fehler Log geschrieben. Hier gilt der Grundsatz, nutzen was schon vorhanden ist. In Listing x.y wird der Quellcode des Skriptes zum Aufteilen der Datensätze gezeigt. Die Laufzeit wird grundsätzlich durch die I/O-Geschwindigkeit der Festplatte bestimmt. Die Begrenzung der erstellten Dateien erwies sich bei der Implementierung als Hilfe, um ein fehlerhaftes Anlegen von tausenden kleinen Dateien zu vermeiden.

Listing mit `splitfile.php` und `vllt` von Andre das Import script.

Kapitel 4

Datenverarbeitung mit Data Mining

4.1 Vorverarbeitung der Daten

Bevor die gesammelten Daten analysiert werden können, müssen Teile der Datensätze vorverarbeitet werden, um sie in ein brauchbares Datenformat zu bringen.

Strecken eines Zuges werden in langen Zeichenketten statt EVA-Nummern abgelegt

Die Datenbank enthält mehrere tausend Datensätze von Zügen die an verschiedenen Haltestellen und Bahnhöfen halten. Die Strecke, die ein Zug fährt ist eine der wichtigsten Informationen, die aus den Datensätzen herausgelesen werden muss. Jedoch ist das in dem ursprünglichen Format der Datensätze sehr ineffizient auszulesen. Die einzelnen

```
1 import sys
2 from glob import glob
3 import os
4 import pymysql
5 import json
6 import re
7 import io
8     hhmss = input
9     (h, m, s) = hhmss.split(':')
10    result = int(h) * 60 + int(m)
11    return result
12
13
14 #used instead of hash buckets to get a better idea of the
15 meaing of the values
16 #warning this function is slow
17 def coloumntovocalfileold(name, input):
18     # ii#
19     filename = "./vocabfiles/" + str(name) + ".txt"
20     with io.open(filename, mode="r+", encoding="utf-8") as
21         file:
22             for line in file:
23                 if input in line:
24                     break
25                 else: # not found, we are at the eof
26                     file.write(input) # append missing data
27
28 def openvocalfile(name):
29     # ii#
```

Quellcode 4.1: Some Python File

4.2 Stochastische Analyse

4.3 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 vorverarbeitet 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 Begriffsdefinitionen für ein neuronales Netz

Welche Begriffe werden häufig verwendet, sollte man gehört haben und zuordnen können

Beim Einstieg in das Themengebiet neuronale Netze fallen viele fremde Begriffe. Diese sollten vorab geklärt sein, um Missverständnisse zu vermeiden. In folgender Auflistung werden die allerwichtigsten Begriffe erklärt, weitere Begriffe können im Internet (siehe Quellenangaben) nachgelesen werden.

Die Liste vervollständigen und eventuell Quelle mit weiterführenden Definitionen angeben

Feature wird ein Attribut einer Zeile genannt, in diesem Fall zählt zum Beispiel die evanr als Feature in dem Datensatz.

Label wird als die Spalte des Datensatzes definiert, welche am Ende vom neuronalen Netz vorhergesagt werden soll. In unserem Fall wäre die Ankunftszeit (IST) eine solche Spalte.

Layer beschreibt eine Schicht von Neuronen, die Anzahl der Neuronen eines Layers wird anhand der sogenannten Hidden Units festgelegt. Diese gibt gleichzeitig die Anzahl der Layer vor. Ein Beispiel: [20,5,10] bedeutet 20 Neuronen im ersten Layer, fünf Neuronen im zweiten Layer und zehn Neuronen im Output Layer

Loss

Accuracy

Optimizer

Estimator

Input Function nennt man die Funktion, welche für die Eingabe von Datensätzen im Training, Testen und Vorhersagen verwendet wird. Die Funktion liest die Datensätze auf der Festplatte ein (zum Beispiel eine .csv-Datei) und gibt zwei Tensoren zurück. Der erste Tensor beinhaltet alle Feature Spalten der Datensätze und der zweite Tensor die Label der Datensätze.

Model Function

Activation Function

Dropout ist ein Float Wert zwischen 0.0 und 1.0, wobei 0.0 für keine fehlenden Verbindungen zwischen den Layern der Neuronen steht und 1.0 bedeuten würde, dass es keine Verbindungen gäbe. Ein guter Wert liegt zwischen 0.0 (ein sogenanntes "fully connected neuronal network" oder 0.3). Der Dropout verhindert, dass alle Datenwerte direkt von Relevanz sind und vermeidet somit ein sogenanntes Overfitting des Modells auf die Trainingsdatensätze.

Tensor

Epochs ist die Anzahl an Epochen, welche das Modell durchlaufen soll.

Steps ist die Anzahl der Schritte, die pro Epoche von dem Modell trainiert werden soll. Bei einer Vorhersage wird die Schrittzahl auf die Anzahl der eingegebenen Datensätze gesetzt beziehungsweise automatisch von Tensorflow erkannt.

5.4 Eingabe der Datensätze in Tensorflow

Input Funktion beschreiben

Die Eingabe von Datensätzen und die Vorverarbeitung sind bei der Erstellung eines neuronalen Netzes von hoher Bedeutung. Die Zeit, eine gut funktionierende und schnelle Eingabefunktion zu schreiben macht sich beim Trainieren des neuronalen Netzes bemerkbar. Da beim Training viele Datensätze in kurzer Zeit benötigt werden, muss ein Engpass an dieser Stelle wenn möglich vermieden werden. Bevor die Eingabefunktion geschrieben wird, müssen die Spalten der Datensätze im Modell angelegt werden. Es wird also ein Modell mit den Spalten als Variablen angelegt, in welches zu späterer Zeit von der Eingabefunktion echte Werte eingesetzt werden. Deshalb befinden sich in einem Modell des neuronalen Netzes auch niemals echte Datensätze sondern nur die Parameter, welche durch das Trainieren erstellt wurden.

Hier ein Code Snippet der inputfn anzeigen und beschreiben

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.

Die Vorhersage mit neuronalen Netzen unterliegen einer Grundlegenden Struktur. Durch die Input Funktion werden die bekannten Größen des Modells an Tensorflow gegeben. Dort wird die Vorhersage durchgeführt und liefert einen Tensor als Antwort zurück. In diesem Falle besitzt der Tensor 1441 Klassen, welche jeweils eine Uhrzeit darstellen. Jeder Uhrzeit wird über eine Softmax Funktion eine relative Wahrscheinlichkeit zugeordnet. Dies bedeutet im Klartext, dass die Summe aller Klassen gleich 100% entsprechen. Ein erwartetes Ergebnis einer Vorhersage wäre also eine Normalverteilung über einen bestimmten Wert. Dies hätte die Bedeutung, dass ein Zug zum Beispiel mit der Wahrscheinlichkeit 75% genau zu dieser Zeit kommt oder mit 95% Wahrscheinlichkeit in einer Zeitspanne von fünf Minuten um diesen Wert. Je nachdem, wie genau das Modell die Realität vorhersagen kann, kann diese Kurve schmaler werden, wodurch die Wahrscheinlichkeit einer genaueren Vorhersage größer ist.

Hier ein vergleichsgraphen erwartete Verteilung, echte verteilung zeigen.

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

Visualisierung und Bereitstellung der Daten im Internet

6.1 Aufbau der Website

Wie wird die Website bereitgestellt, was kann sie und welche Views existieren für die Nutzer

Heutzutage ist die Bereitstellung einer Website eine einfache Methode Daten mit anderen Menschen zu teilen. Da unsere Datenbank und Website einen gewissen Sicherheitsstandard erfüllen soll, wird sich für ein Framework entschieden, welches bereits integrierte Sicherheitsfunktionen bietet. Der name des Frameworks lautet Laravel

Hier reflink zu laravel seite einfügen

. Dies spart vor allem Zeit bei der Entwicklung der neuen Funktionen für die Bereitstellung der einzelnen Webviews. Ein View ist eine Seite oder der Teil einer Website, welcher in eine weitere Seite eingebettet sein kann. In Abbildung x.y kann die Struktur der einzelnen Ansichten erkannt werden.

Hier Abbildung View mit mehreren Subviews

Auf der Website gibt es folgende Hauptpunkte, welche jedem Nutzer zur Verfügung stehen.⁴

Home ist die Startseite der Nutzer. Hier sollen Grundinformationen an die Nutzer gegeben werden, wie zum Beispiel die Anzahl an Datensätzen (gesamt). Diese Seite soll optimiert sein schnell zu laden, weshalb sie relativ wenig Daten an den Nutzer senden soll. Dies ist vor allem in betracht auf die Mobile Nutzung der Website wichtig.

Toplist

Hier die Punkte der Website updaten wenn sich etwas ändert.

Map ist eine Karte, welche als Basisoberfläche Kartenmaterial von OpenStreetmaps.org verwendet. Darauf werden mithilfe von leaflet einzelnen Schichten gezeichnet, wie zum beispiel die Bahnhöfe und die Streckenverläufe der deutschen bahn.

Hier links zu OSM und leaflet einbinden, sowie zu den rohdaten von db bzw. dem github repo

Stationen ist die Hauptansicht für Statisten der einzelnen Stationen. Auf der Hauptseite befindet sich eine Suchfunktion mit grundlegenden Einstellungen. Nach erfolgreicher Suche nach einem Bahnhof kann sich der Nutzer eine der vielen erzeugten Ansichten anschauen.

Impressum ist eine Verlinkung auf das nach deutschem Recht benötigte Impressum einer Website gemäß § 5 Telemediengesetz (TMG)

6.2 Visualisierung der Datensätze

Dimensionen der Daten, ORT; ZEIT; NAME; STRECKE; etc.

Zur Visualisierung der Datensätze werden vorberechnete und LiveDaten verwendet. Je nachdem, ob dem nutzer eine interaktionasmöglichkeit gegeben werden soll, wird dies entschieden. Die dafür verwendeten tools stammen aus dem Python Modul matplotlib oder aus der javascript Library d3.js.

Verlinken auf matplotlib und d3js

Die Anzahl der Dimensionen der Datensätzen macht es schwierig zu entscheiden, ob eine Ansicht oder Grafik für diese überhaupt Sinn ergibt. Daher werden zu Beginn der Visualisierungsprozesses verschiedene Techniken ausprobiert. Alle Skripts werden mit dem Bedacht auf die Wiederverwendbarkeit geschrieben. Ein wichtiger Schritte von der Datenbank zur fertigen Grafik ist die SQL-Abfrage. Diese soll optimiert sein, um den Datenbankserver nicht unnötig zu belasten. Dafür kann das MySQL Schlüsselwort EXPLAIN verwendet werden.

Verlinkung auf Explain bzw. erklären mit Screenshot

Nachdem die Abfrage ausgeführt ist, wird die Antwort als Objekt abgespeichert, hier gibt es grundlegende Unterschiede, ob das Objekt weitere methoden enthält, oder ob das Objekt als simple Datensammlung darstellt.

Kapitel 7

Schlussfolgerung

7.1 Rückblick

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

7.2 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).

7.3 Ausblick

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

Liste der ToDo's

<input type="checkbox"/>	Sperrvermerk ja oder nein	1
<input type="checkbox"/>	Dieses Abstract besser schreiben und eventuell eine englische Übersetzung anfertigen	1
<input type="checkbox"/>	Wie kam es dazu, eventuell mit Motivation kombinieren.	7
<input type="checkbox"/>	Wieso wollen wir das machen und warum ist das für uns wichtig.	7
<input type="checkbox"/>	Hier etwas zum Stand der Technik schreiben, neuronale Netze, Tensorflow, KI, Datamining, OpenData, etc	8
<input type="checkbox"/>	Hier Zitat aus Buch Definition zu datamining und datenbeschaffung einfügen	8
<input type="checkbox"/>	Hier das Ziel aus der Anmeldung schön definieren und klar Abgrenzen was Ziel und was optional nice to have ist.	8
<input type="checkbox"/>	Was bekommen wir eigentlich alles über die Api geliefert	10
<input type="checkbox"/>	Zeitliche Einteilung, beachten 5. Semester ist weniger Zeit, Hauptteil wird im 6. Semester passieren	12
<input type="checkbox"/>	Hier Gantt Diagramm oder Tabelle einfügen mit was wurde in welchem Semester gemacht.	12
<input type="checkbox"/>	eventuell Verlinken	12
<input type="checkbox"/>	Data Mining Einführung und dessen Bedeutung für das Projekt	13
<input type="checkbox"/>	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	13
<input type="checkbox"/>	Datenmodell erläutern, welche Rohdaten aus der DB-API	14
<input type="checkbox"/>	Wie werden Daten aufbereitet, vorbereitet für das neuronale Netz, welche Dinge gibt es zu beachten (DATENTYPEN!)	16
<input type="checkbox"/>	Kleine Einleitung an einem Simplen Beispiel, Linear Regression oder so. Wieso wir sowas brauchen und weshalb es von Relevanz ist.	16
<input type="checkbox"/>	Achtung siehe Befriffsdefintionen von enuronalen Netzen, dieses kapitel vllt hier her	16
<input type="checkbox"/>	Erläuterung welche Informationen in das Neuronale Netz eingegeben werden und welche Daten von dem Netz ausgegeben werden.	16
<input type="checkbox"/>	Achtung eventuell doppelter Eintrag siehe spätere Kapitel.	16
<input type="checkbox"/>	Was wird alles für Tensorflow benötigt	16
<input type="checkbox"/>	Eventuell how to install tf verlinken	16
<input type="checkbox"/>	Verweis einfügen	17

■ 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.	17
■ Anzahl Wochen	18
Abbildung: Es fehlen Abbildungen von elementaren Abläufen	20
Abbildung: Es fehlt eine tabelle zum Vergleichen des Funktionsumfangs der Versionen	20
■ Hier noch was bedeutet 503 und eventuell zitat aus RFC https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html	21
■ Wie werden die Datensätze abgespeichert und verwaltet? Das Schema der Datenbank befindet sich im Ressourcen Ordner	21
■ x.y	21
■ Hier etwas darüber erläutern	21
■ x.y	21
■ Literatur verweise einfügen	22
■ x.y	22
Abbildung: Es fehlt die Abbildung für Schema Version 1	22
Abbildung: Es fehlt die Abbildung für Schema Version 2	22
■ Datenkanken sind toll, aber es muss bei eine kritischen Stelle ein backup vorhanden sein.	23
■ Listing mit splitfile.php und vllt von Andre das Import script.	23
■ Strecken eines Zuges werden in langen Zeichenketten statt EVA-Nummern abgelegt	24
■ Wie kommen die Datensätze aus der DB zum neuronalen Netzwerk, wie wird die Formatierung vorgenommen	26
■ Kurze Einführung schreiben	26
■ Hier die generate_csv.py beschreiben	27
■ Welche Begriffe werden häufig verwendet, sollte man gehört haben und zuordnen können	29
■ Die Liste vervollständigen und eventuell Quelle mit weiterführenden Definitionen angeben	29
■ Input Funktion beschreiben	30
■ Hier ein Code Snippet der inputfn anzeigen und beschreiben	30
■ 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	30
■ Aufzeigen wie sich die Menge an Daten auf die Genauigkeit auswirkt	30
■ Welche Optionen und Parameter können optimiert werden, wie ändert sich dadurch das Ergebnis.	30
■ Hier Tabellen mit Vergleich der Methoden und Genauigkeit, Geschwindigkeit, Erläuterungen weshalb das Ergebnis so ist.	30
■ Testen des neuronalen Netzes, Verifikation der Genauigkeit und deren Steigerung durch Training oder Anpassungen des Netzes	30

■ Vorhersagen aus Daten treffen und anschauen wie gut sie sind, wo gibt es Probleme, welche Probleme treten auf.	31
■ Hier ein vergleichsgraphen erwartete Verteilung, echte verteilung zeigen.	31
■ 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.	31
■ Wie wird die Website bereitgestellt, was kann sie und welche Views existieren für die Nutzer	32
■ Hier reflink zu laravel seite einfügen	32
■ Hier Abbildung View mit mehreren Subviews	32
■ Hier die Punkte der Website updaten wenn sich etwas ändert.	32
■ Hier links zu OSM und leaflet einbinden, sowie zu den rohdaten von db bzw. dem github repo	32
■ Dimensionen der Daten, ORT; ZEIT; NAME; STRECKE; etc.	33
■ Verlinken auf mathplot und d3js	33
■ Verlinkung auf Explain bzw. erklären mit Screenshot	33
■ Was ist geschehen, was würden wir anders machen, was waren wichtige Schritte	34
■ 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.	34
■ Wie geht es weiter, könnte es weiter gehen, was sollte verbessert werden, wo befinden sich Schwachstellen, event. ungelöste Probleme	34