

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

Wir versichern hiermit, dass wir unsere 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 haben. Wir versichern zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort

Datum

Unterschrift

Zusammenfassung

Dieses Abstract besser schreiben und eventuell eine englische Übersetzung anfertigen

Jeder muss ab und an mit der Bahn fahren. Die einen mehr die anderen weniger. Egal wie oft jemand mit der Deutschen Bahn fährt, es ist immer ärgerlich, wenn die Bahn zu spät kommt. Festzustellen, an was jede einzelne Verspätung liegt, ist nicht in Rahmen unserer Möglichkeiten. Aber wir können herausfinden, ob sich Verspätungen der Deutschen Bahn an bestimmten Stellen des Netzes Häufen. Durch die Timetables API der Deutschen Bahn sind seit dem 09.04.2018 alle Texte der Anzeigetafeln frei verfügbar. Dadurch auch, wann Züge an Bahnhöfen wirklich ankommen und wann sie abfahren. Die vorliegende Studienarbeit befasst sich mit von der Bahn zu Verfügung gestellten API. Mit ihrer Hilfe sollen bundesweit Daten über Zug Halte gespeichert werden. Mit diesen Daten sollen verschiedene statistische Analysen durchgeführt werden, die im Nachgang auf einer Website veröffentlicht werden sollen. Außerdem soll mit diesen Daten ein neuronales Netz modelliert werden, welches genutzt werden kann, um Verspätungen und Abhängigkeiten im Schienen Verkehr zu erkennen und vorherzusagen.

Inhaltsverzeichnis

1	Einleitung	7
1.1	Motivation	7
1.2	Ziele und Ablauf der Studienarbeit	8
1.3	Stand der Technik	10
1.4	Planung	11
1.4.1	Zeitliche Einteilung der Studienarbeit	11
1.4.2	Versionsverwaltung	11
1.5	Begriffsdefinitionen	12
2	Grundlagen	13
2.1	Data Mining	13
2.2	Datenmodell	14
2.3	Modellierung realer Größen	15
3	Datenbeschaffung	16
3.1	Die DB Timetable API	16
3.1.1	Station	16
3.1.2	Plan	17
3.1.3	fchg	17
3.1.4	rchg	17
3.2	Programmierung des Data Miners	18
3.3	Weatherminer	22
3.3.1	OpenWeatherMap	24
3.4	Datenbank und Schema	26
3.4.1	Datenbank Schema des Wetterminers	28
3.5	Backup der Datenbank	28
4	Datenverarbeitung mit Data Mining	30
4.1	Grundlagen von Data Science und KDD	30
4.1.1	Data Mining	30
4.1.2	Knowledge Discovery in Databases (KDD)	30
4.2	Grundlagen	30
4.3	Vorverarbeitung der Daten	31
4.4	Software-Architektur der Datenauswertung	33

4.4.1	Query Suite	34
4.4.2	Processing Utils	36
4.5	Statistische Auswertungen	36
4.5.1	Definition der Verspätungen	36
4.5.2	Analyse der Verspätungen eines Zuges	40
4.6	Stochastische Analyse	43
4.6.1	Durchschnitt der Zeiten einer Haltestelle	43
4.6.2	Durchschnitt der Zeiten einer Strecke	44
4.7	Visualisierung	46
5	Datenverarbeitung mit neuronalem Netz	47
5.1	Programmierung der Automatischen Datenverarbeitung	47
5.2	Vorverarbeitung der Datensätze	48
5.3	Einrichten der Tensorflow Umgebung	52
5.4	Begriffsdefinitionen für ein neuronales Netz	53
5.5	Eingabe der Datensätze in Tensorflow	54
5.6	Eingabe- und Ausgabe-Parameter für das Neuronale Netz	56
5.7	Anlernen des Netzes	56
5.8	Verifizieren des Netzes	57
5.9	Vorhersagen anhand des Netzes	57
5.10	Bewertung der Ergebnisse	60
6	Visualisierung und Bereitstellung der Daten im Internet	61
6.1	Aufbau der Website	61
6.2	Erstellung der Webrouten	63
6.3	Erstellung der Seiten	64
6.3.1	Idee des dynamischen Nachladen	64
6.3.2	Die Stationsübersicht	65
6.3.3	Die Zugübersicht	65
6.4	Testen der Seiten mit Unit Test	66
6.5	Visualisierung der Datensätze	66
6.6	Darstellung der Datensätze	69
7	Schlussfolgerung	70
7.1	Rückblick	70
7.2	Fazit	70
7.3	Ausblick	71
	Anhang	72
	Literaturverzeichnis	72
	Liste der ToDo's	73

Abbildungsverzeichnis

1.1	Ablauf der Implementierung	9
3.1	Grundablauf des Miners	20
3.2	Übersicht der Postleitregionen von Deutschland. Erstellt von Stefan Kühn, vom Wikimedia Commons	23
3.3	Datenbank Schema Version 1	27
3.4	Datenbank Schema Version 1	27
4.1	Übersicht des KDD-Prozesses [USAMA FAYYAD 1996, S. 41].	31
4.2	Grundablauf des Data Minings	34
4.3	Verspätungsanalyse von RE4725 von Karlsruhe nach Konstanz	41
4.4	Verspätungsanalyse von ICE74 von Basel Bad. Bf. nach Kiel mit Verspätung bei Abfahrt	42
4.5	Verspätungsanalyse von ICE74 von Basel Bad. Bf. nach Kiel mit Verspätung bei Abfahrt	42
4.6	Prozentuale Anteile der Durchschnittlichen Verspätung bei ankommenden Zügen	45
5.1	Der Verzeichnisbaum mit der Aufteilung der Datensätze	50
5.2	Ausschnitt aus einem Vocabfile, hier zu sehen Gleis (Soll)	51
5.3	Verteilung der Wahrscheinlichkeiten bei 288 Klassen und allen Spalten als Eingabeparameter	58
5.4	Verteilung der Wahrscheinlichkeiten bei 288 Klassen mit nur der Ankunfts- uhrzeit als Eingabeparameter	59
6.1	Struktur der einzelnen Views der Website	62
6.2	Dreidimensionales Histogramm aus dem TensorBoard	69

Tabellenverzeichnis

1.1	Aufgaben über die Zwei Semester	11
3.1	Tabelle mit allen Wetterverhältnisse	25
4.1	Struktur der Durchschnitts Tabelle	45
5.1	Vorverarbeitung der Datenbank-Daten	52

Liste der Quellcodeausschnitte

3.1	Drei Ausschnitte aus einer Datei	21
3.2	Befehl zum einfügen der SQL Dateien	29
4.1	Some Python File	32
4.2	Zerlegen der Zug ID in seine Komponenten	33
4.3	Beispiel einer Query-Methode	36
4.4	Berechnung der SARV	39
4.5	SQL Query für neue Halte einer Haltestelle	44
5.1	Ausschnitt aus der Datei generate_csv.py	49
5.2	Ausschnitt von der Input Funktion aus der Datei train_test_predict.py .	55
6.1	Beispiel einer Webrountendefinition	63

Abkürzungsverzeichnis

HTTP	Hypertext Transfer Protocol.....	22
PLZ	Postleitzahl.....	22
PLR	Postleitregion.....	22
SARV	Streckenabschnitt-respektive Verzögerung	
KDD	Knowledge Discovery in Databases.....	1

Kapitel 1

Einleitung

TODO

Ratschlage von meinem Ausbilder: Roter Faden ausbauen

Einleitung überarbeiten: Den Scope einengen: Wir haben Grundsteine gelegt auf dem Weg zur Prognose Implementierungsablauf vorstellen, wie sind wir vorgegangen und weshalb? (Diagram) Datenaquirierung: Wo bekommen wir sie her? Dateneingrenzung: Welche Daten verwenden wir? Datenspeicherung: Wie speichern wir sie? etc. Motivation starker ausformulieren: Praxisbezug zum Leser herstellen

Website mit Visualisierungen besser darstellen + Screenshots! Ansätze des Neuronalen Netzes beschreiben + Probleme schildern, die sich dabei ergeben haben

Ausblick: Prognose im Ausblick beschreiben und wie die bisherige Arbeit für die Prognosenbildung hilft

TODO

1.1 Motivation

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

Planbarkeit soll verbessert werden, dadurch verringerte Reisedauer

Im Rahmen der Open Data Bewegung veröffentlicht auch die Deutschen Bahn¹ die Echtzeitdaten der Fahrpläne ihrer Personenzüge. Aus diesem Umstand heraus, hat sich der Wunsch aufgetan, die Daten der Deutschen Bahn für eine Studienarbeit zu verwenden. Um die Daten der Deutschen Bahn in einer Studienarbeit zu verwenden, benötigt es noch eine Idee, was aus den Daten gewonnen werden kann. Hierbei hat sich folgende Idee entwickelt:

Verspätungen im öffentlichen Nah- und Fernverkehr treten täglich auf. Oftmals wartet der Fahrgast mehrere Minuten, bis sein Zug endlich einfährt. Besonders früh morgens oder im Winter wünscht sich der Fahrgast, dass er die Wartezeit Zuhause hätte verbringen

¹Siehe <https://developer.deutschebahn.com/store/apis/info?name=Timetables&version=v1&provider=DBOpenData>

können, anstatt im Bahnhof auf die Ankunft des Zuges zu warten. Die Idee ist es hierbei, mit den Echtzeitdaten der Deutschen Bahn Vorhersagen abzuleiten, die dem Fahrgast mitteilen, wie groß die Verspätung des Zuges ist, auf den er wartet. Mit diesem Wissen kann es sich der Fahrgast mit einem Kaffee daheim nochmals gemütlich machen, bevor er zum Bahnhof aufbricht. Genauso gut kann der Fahrgast die Prognosen nutzen, um abschätzen zu können, ob er Verbindungszüge erreicht oder gegebenenfalls mehr Zeit einplanen muss.

Da zur Prognosenbildung die Ursachen der Verspätung häufig nicht direkt erkennbar sind, soll mit dieser Studienarbeit Grundsteine für die Vorhersage von Verspätungen gelegt werden. Da die Prognose von Verspätungen als komplexe Aufgabe eingeschätzt wird, beschränkt sich diese Studienarbeit auf die Analyse der Daten. Die Erkenntnisse aus der Analyse der Daten sollen die Grundsteine für die Entwicklung eines Prognose-Modells bilden. Die Prognosenbildung für Verspätungen wird somit als optionales Ziel definiert.

1.2 Ziele und Ablauf der Studienarbeit

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

In diesem Abschnitt werden die vorgesehenen und die optionalen Ziele der Studienarbeit formuliert und dem Leser aufgezeigt in welche Schritte sich die Studienarbeit aufteilt.

Vorgesehene Ziele Die Ziele der Studienarbeit sind es, Verspätungen von Zügen festzustellen und anschließend zu analysieren. Hierbei sollen unterschiedliche Aspekte bei der Analyse betrachtet werden: Es sollen die Verspätungen nach Abhängigkeit der Zeit, Ort und Strecke, sowie kritische Punkte analysiert werden. Dabei sollen die Analysen auch visualisiert werden, um Schlüsse aus ihnen ziehen zu können.

Optionale Ziele Da die Prognosenbildung für Verspätungen als komplexe Aufgabe eingeschätzt wird, ist die Vorhersage der Verspätung eines Zuges als optionales Ziel eingestuft. Für die Prognosenbildung sollen neben den Erkenntnissen aus der Analyse der Daten zusätzlich auch Wetterdaten einbezogen werden. Es wird vermutet, dass das Wetter besonders bei Stürmen, Kälte oder Hitze einen Einfluss auf die Pünktlichkeit von Zügen hat. Aus diesem Grund werden bei dem Wetter die Windgeschwindigkeit, die Art des Niederschlags, die Menge des Niederschlags und die Höhenlage der Bahnhöfe als zu nutzende Wetterdaten in Betracht gezogen.

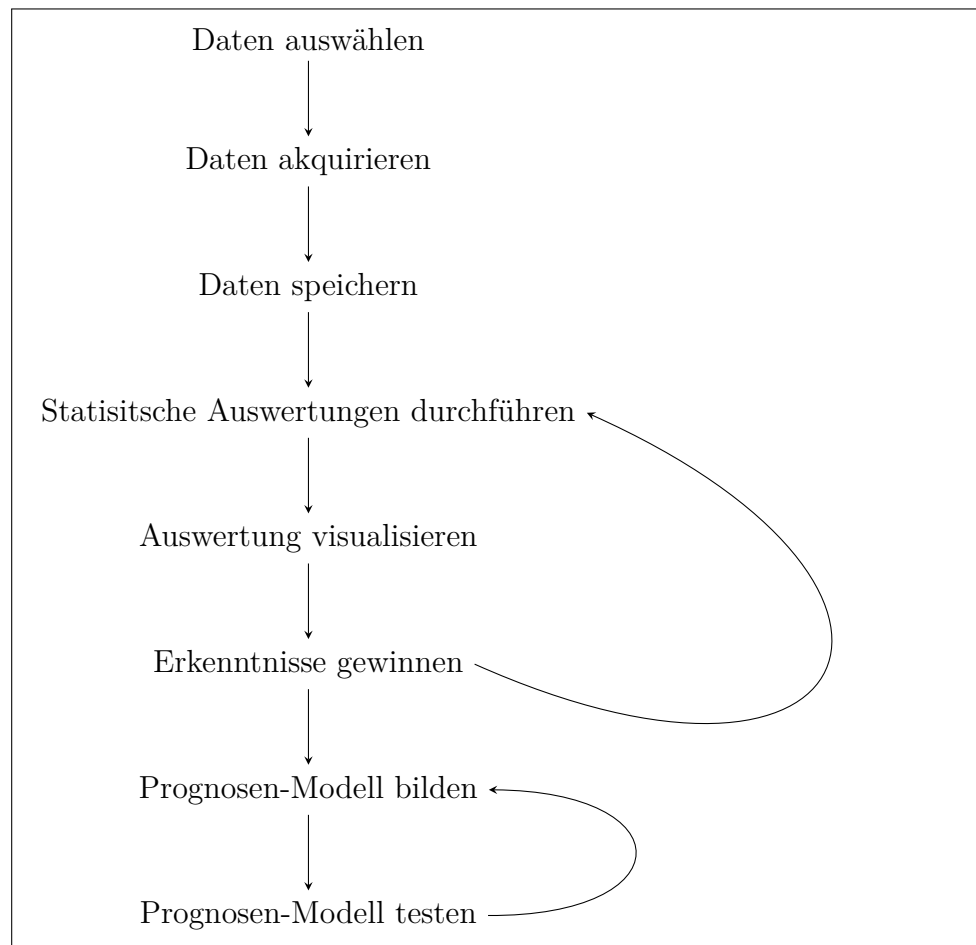


Abbildung 1.1: Ablauf der Implementierung

Struktur der Studienarbeit Die Studienarbeit teilt sich in einzelne Schritte auf. In Abbildung 1.1 werden die einzelnen Schritte visualisiert. Hierbei führt jeder Schritt auf das Ziel zu, Prognosen für die Verspätung von Zügen vorhersagen zu können. Der erste Schritt, der in der Studienarbeit durchzuführen ist, ist die Beschaffung der Daten über die API der Deutschen Bahn. Dieser Schritt unterteilt sich in drei Teilschritte. Der erste Teilschritt ist, diejenigen Daten auszuwählen, die nützlich für die Auswertung und Prognose sein können und diejenigen Daten zu ignorieren, die für die Analyse und Prognose unnütz erschienen. Dieser Schritt wird in der Abbildung “Daten auswählen” genannt. Danach folgt der Schritt “Daten akquirieren” bei dem die Daten von der API abgerufen und im nächsten Schritt “Daten speichern” in der Datenbank gespeichert werden. Diese genannten drei Teilschritte werden in dem Kapitel 3 *Datenbeschaffung* behandelt.

Der nächste Schritt ist das statistische Auswerten der Daten, um anschließend das Ergebnis der Auswertung zu visualisieren. Mithilfe der Visualisierung soll die Interpretation der Auswertung leichter fallen, um anschließend Erkenntnisse über die analysierten Daten zu gewinnen. Diese Schritte werden in Kapitel 4 *Datenverarbeitung mit Data Mining* durchgeführt. Hierbei können die drei Schritte als Schleife durchlaufen werden, um mit

den bisherigen Kenntnissen weitere Hypothesen aufstellen zu können und durch weitere Analysen zu bestätigen oder zu widerlegen.

Mit den erlangten Kenntnissen aus den Auswertungen kann anschließend ein Prognosen-Modell entworfen werden. Dieses Modell muss anschließend auf seine Qualität getestet werden. Diese zwei Schritte können ebenfalls als Schleife durchlaufen werden, um das Prognosen-Modell solange anzupassen, bis es die gewünschte Qualität erreicht.

Neuronales Netz für Prognosenbildung In dieser Studienarbeit wird parallel zu der Analyse der Daten versucht, ein neuronales Netz für die Vorhersage der Verspätung eines Zuges zu trainieren. Das neuronale Netz wird unter der Annahme entwickelt, dass es Zusammenhänge selbst aus den Trainingsdaten erlernt. Unter dieser Annahme ist die Entwicklung des Netzes unabhängig von den zu erlangenden Erkenntnissen aus den Datenanalysen, sodass das Netz parallel zu den Datenanalysen entwickelt werden kann. Die Entwicklung des neuronalen Netzes wird in Kapitel 5 *Datenverarbeitung mit neuronalem Netz* beschrieben.

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 maschinellen Lernens, 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.

Data mining refers to extracting or „mining“ knowledge from large amounts of data [JIAWEI HAN 2000]

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. Die Visualisierung wird in späteren Kapiteln genauer betrachtet.

Eine wichtige Änderung in den letzten Jahren war die Entscheidung einiger Großunternehmen, Daten über API Schnittstellen für Entwickler verfügbar/nutzbar zu machen.

Eine wichtige Änderung in den letzten Jahren ist der Wille von Großunternehmen einen Daten über eine API Schnittstelle Entwicklern bereitzustellen.

1.4 Planung

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

Die Planung der Studienarbeit sieht vor, wie in Abbildung 4.2 gezeigt, dass wir zuerst die Daten die wir aus der DB API bekommen können sichten. Die bei der Sichtung als Nützlich

1.4.1 Zeitliche Einteilung der Studienarbeit

Im ersten Teil der Studienarbeit steht die Erfassung der Daten der Deutschen Bahn im Vordergrund. Neben der Programmierung des Data Miners für die Bahn API wird Literatur, welche für die anschließende Aufbereitung und Visualisierung der Datensätze benötigt wird. Des weiteren wird Literatur gelesen und auf nutzbare Themen durchsucht. Da die Erfassung der Daten einige Zeit in Anspruch nimmt, und die Zeitlichen Freiräume fehlen, wird die Hauptarbeit in das sechste Semester verschoben. In diesem soll die Analyse der Daten sowie deren Auswertung erfolgen. Auch die Prognose soll im sechsten Semester implementiert werden.

schwerer Fehler

5. Semester	6. Semester
Data Miner Literatur Recherche	Analyse der Daten Statistische Analyse Prognose Visualisierung

Tabelle 1.1: Aufgaben über die Zwei Semester

1.4.2 Versionsverwaltung

Wichtig für die Planung ist neben der Zeiteinteilung auch die Versionsverwaltung des Quellcodes und der Ausarbeitung selbst. Da ständig eine aktuelle Version beider Arbeiten zu Verfügung stehen muss und alle Autoren parallel schreiben und editieren können müssen, wird die Wahl der Gruppe auf den Onlinedienst Github für den Quellcode und ShareLaTeX für die Ausarbeitung. In Github wird eine öffentliche Organisation angelegt, welcher alle Autoren beitreten. Innerhalb 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 ein Backup und der aktuelle Wissensaustausch zwischen den Teilnehmern sichergestellt. Die gemeinsame Arbeit an Quellcode wird durch die Versionsverwaltung erleichtert, da parallel in verschiedenen Branches gearbeitet werden kann.

1.5 Begriffsdefinitionen

Im Rahmen dieser Arbeit werden Begriffe verwendet, die 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 Gleisen 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 Bahnhöfen, die in ihrer anzufahrenden Reihenfolge angeordnet sind.

Data Mining Data Mining ist die Anwendung statistischer Methoden auf große und komplexe Datenmengen mit dem Ziel neue Rückschlüsse und Zusammenhänge zu erkennen.

Kapitel 2

Grundlagen

Dieses gesamte Kapitel aufsplitten und in die Grundlagen der einzelnen Chapter einordnen

2.1 Data Mining

Data Mining Einführung und dessen Bedeutung für das Projekt

Das Data Mining ist ein essentieller Bestandteil des Projektes. Ohne genügend Daten als Grundlage kann dieses Projekt nicht funktionieren, da ein Neuronales Netz nur mithilfe von Datenmaterial trainiert werden kann. Hierbei gilt: Je mehr Datenmaterial zur Verfügung steht, desto genauer das Neuronale Netz. Um die weitere Automatisierung des Datenflusses zu ermöglichen, werden die Datensätze in einem offenen und weiterverwendbaren Format gespeichert.

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:

- Bahnstationsnummer
- Linie als Folge von angefahrenen Stationen (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.

2.2 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.

2.3 Modellierung realer Größen

Schauen, ob Kapitel noch Sinn macht

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.

Kapitel 3

Datenbeschaffung

3.1 Die DB Timetable API

Was bekommen wir eigentlich alles über die Api geliefert

Folgende Endpunkte können in der Deutschen Bahn API¹ abgefragt werden.

3.1.1 Station

Hier noch besser formatieren und nochmal lesen Kapitel x.1.1 -x.1.4

Dieser Endpunkt gibt Informationen über ein Bahnhof zurück. Dafür kann sowohl der Name der Station, als auch die eindeutige EVA Nummer oder die ds100 beziehungsweise 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>
```

¹API-URL: <http://api.deutschebahn.com/timetables/v1>

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

In diesem gesamten Kapitel fehlen noch Quellenangaben und weitere Literaturhinweise

3.1.2 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 „planned“ Attribute genutzt.

3.1.3 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 Timetable wird der Name der Station, die EVA Nummer, eine Liste von Timetable-Stops und Messages zurückgegeben.

3.1.4 rchg

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

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

Der rchg Endpunkt unterscheidet sich ausschließlich durch die übertragenen Änderungen von dem fchg Endpunkt, die im Fall des rchg 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 Strecke, eine Referenz zum eigentlichen Zug - falls es sich um einen Ersatzzug handelt und die Events *Ankunft* und *Abfahrt*, in denen vor allem die An- bzw. Abfahrtszeiten und das Gleis untergebracht sind. Es kann jeweils die geplante, als auch die prognostizierte Information enthalten sein, eine Message, weshalb eine Änderung stattfand, sowie die Informationen wie viel Verspätung die Bahn hat und ob sie auf ein anderes Gleis umgeleitet wurde.

Message Eine Message besteht aus einer Message-Id, einem Message-Typ und einem Timestamp. Folgende Informationen können zusätzlich angehängt werden:

- Information auf welche Uhrzeit der Zug verlegt wurde, aber auch wann der Zug eigentlich geplant war
- Code um die Message zu identifizieren
- Text der Nachricht
- Kategorie der Nachricht
- Priorität der Nachricht
- Eigentümer der Nachricht
- externer Link
- Indikator ob Nachricht gelöscht
- Nachricht des Verteilers
- Name des Zuges

3.2 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 in die Datenbank schreiben
- Fehler in einer Tabelle protokollieren

Durch die geringere Datenmenge (Reduzierung der Anzahl abgerufener Stationen von 6600 auf 1200), konnte die Umsetzung schnell realisiert werden. Da es sehr viele Optionen und Probleme gab, wurde die erste Version nach etwa x Wochen durch die zweite Version des Miners ersetzt. Diese besitzt neben neuen Funktionen auch die Erweiterung zur vollständigen Abfrage der API. Außerdem konnten die Probleme des Miners minimiert werden. Die zweite Version kann zudem alle Daten abfragen und nutzt deutlich mehr Informationen, die in der API der Bahn bereitgestellt werden. Die wichtigste Änderung ist die Fehlererkennung in der Abfrage der Datensätze. Hierdurch wird vermieden, dass zu viele Datensätze fehlen. 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 denen aus der API Dokumentation erhalten Datenstrukturen. So sollte zum Beispiel 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.

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.

Hier noch verfeinern und grafiken anpassen

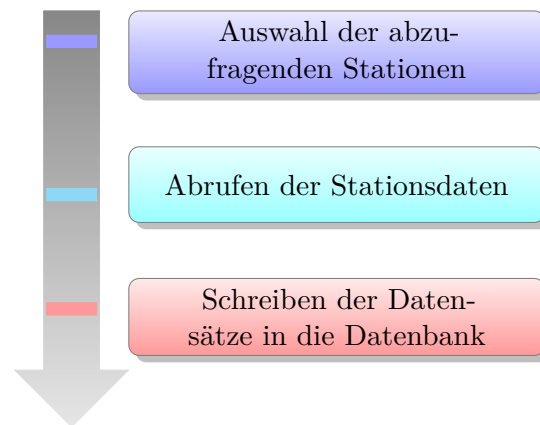


Abbildung 3.1: Grundablauf des Miners

Hier Quellcode updaten und anzeigen, beschreiben

```

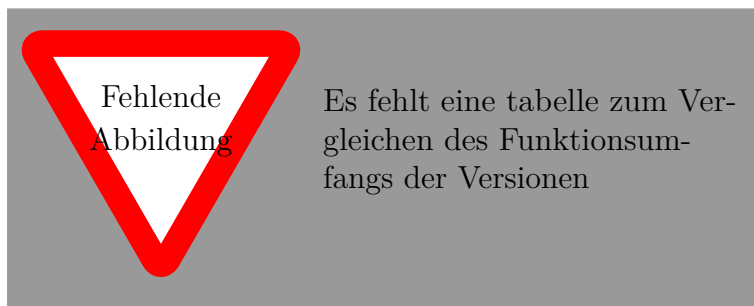
1 <?php
2
3 include_once './settings.php';
4 require_once 'classes/MysqliDb.php';
5 require_once 'classes/appgati.php';
6 // currently not used maybe later
7 } else {
8     $bahnapi = new bahnapi($apikey2);
9 }
10
11
12 // Using old querie here because limit dosnt seem to work in
    rawquery
13 $params = date("Y-m-d_H:i:s", time() - 3600);
14 $mysqislave = new mysqli(SETTING_DB_IP, SETTING_DB_USER,
    SETTING_DB_PASSWORD, SETTING_DB_NAME);
15
16 if($minute == 0 || $minute == "00" || $minute == "0") {
17     $stationsquery = $mysqislave->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 = $mysqislave->query("SELECT EVA_NR as nr,
    NAME FROM haltestellen2 WHERE fetchactive2=1 AND fetchtime <
    '$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 Miners 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)

64 Gigabyte statt 16 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.

3.3 Weatherminer

Um eine bessere Prognose der Verspätungen zu ermöglichen, sollte auch das Wetter miteinbezogen werden. Aus diesem Grund wurde entschieden, Wetterdaten Deutschlands abzuspeichern um sie später mit einbeziehen zu können. Dafür wurde eine Datenbankstruktur entwickelt, die das Auslesen der Wetterdaten früherer Zeitpunkte ermöglicht. Es ist allerdings zu beachten, dass, wenn der Wert nicht belegt ist, das Feld nicht übertragen wird. Hat es zum Beispiel in den letzten 3 Stunden nicht geregnet, wird das Feld Regen auch nicht übertragen. Um nicht zu viele Daten speichern, beziehungsweise abfragen zu müssen, wurden nicht die Postleitzahlen Postleitzahl (PLZ) genutzt um das Wetter zu speichern, sondern das nächst größere Gebiet Postleitregion Postleitregion (PLR) genutzt. Die Postleitregionen entsprechen den ersten beiden Zahlen der Postleitzahl und sind wie in Abbildung 3.2 in Gebieten zusammengefasst.

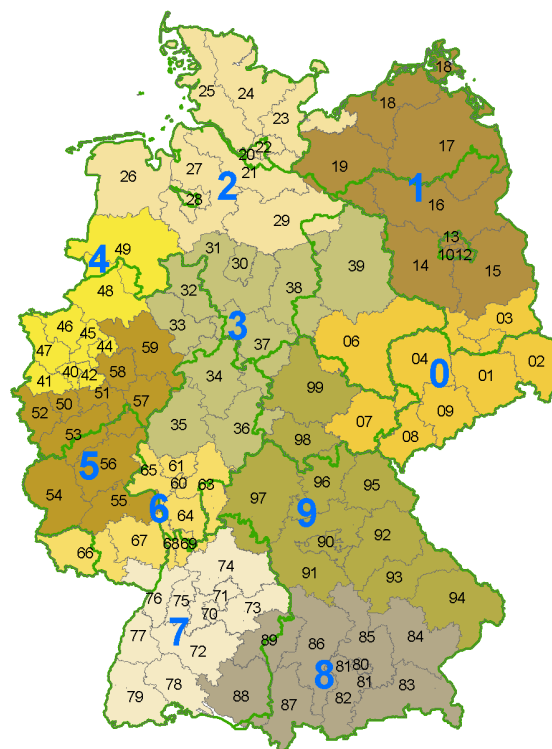


Abbildung 3.2: Übersicht der Postleitregionen von Deutschland. Erstellt von Stefan Kühn, vom Wikimedia Commons

3.3.1 OpenWeatherMap

Durch Angabe der Postleitzahl, Koordinaten oder des Namens der Stelle an der das Wetter abgefragt werden soll. In der Antwort können folgende Parameter ausgelesen werden:

Koordinaten Die geografische Lage der Stadt die angegeben wurde.

Wetter Ein oder mehrere Wetterlagen ID's. Diese geben Aufschluss über die Wetterlage an dem ausgewählten Ort. Eine Liste mit allen Wetterlagen ist in der Tabelle 3.1 auf Seite 25 zu finden. Diese enthält die Wetterlagen ID, die Bezeichnung der Gruppe der Wetterlage, die Beschreibung der Wetterlage und die Kennung des entsprechenden Bildes.

Main In diesem Teil der Antwort werden Werte wie die Temperatur in Kelvin, der atmosphärische Druck in hPa, die Luftfeuchtigkeit, die minimale und die maximale Temperatur sowie Druck auf Meereshöhe und Druck auf Normalhöhe angegeben.

Wind In diesem Abschnitt wird sowohl die Windgeschwindigkeit als auch die Richtung des Windes abgelegt.

Wolken In diesem Abschnitt wird abgelegt, wieviel Prozent des Himmels mit Wolken bedeckt sind.

Regen Unter diesem Key wird die Niederschlagsmenge in l/m² sowie die Niederschlagsmenge der letzten 3 Stunden abgelegt.

Schnee Sowie im Abschnitt Regen, wird auch im Abschnitt Schnee die Menge des momentan fallenden Schnees und die der letzten 3 Stunden abgelegt.

Zeitstempel Zeitpunkt der Datenerhebung

System Neben drei nicht weiter beschriebenen Parametern, wird der angegebene Ländercode nochmal zurückgegeben. Auch die Zeit des Sonnenaufgangs in der Region sowie die Zeit des Sonnenuntergangs wird in diesem Abschnitt in Unix Zeitstempelform abgelegt.

Stadt Hier wird die ID und der Name der abgefragten Stadt abgelegt.

Die OpenWeatherMap API hat allerdings momentan einen Bug, wodurch die Parameter Regen und Schnee in Deutschland nicht übermittelt werden. Um trotzdem abspeichern zu können ob es regnet, werden zusätzlich zu den Wind, Regen, Schnee Parametern die Wetterlagen abgespeichert.

ID	WCondition
200	thunderstorm with light rain
201	thunderstorm with rain
202	thunderstorm with heavy rain
210	light thunderstorm
211	thunderstorm
212	heavy thunderstorm
221	ragged thunderstorm
230	thunderstorm with light drizzle
231	thunderstorm with drizzle
232	thunderstorm with heavy drizzle
300	light intensity drizzle
301	drizzle
302	heavy intensity drizzle
310	light intensity drizzle rain
311	drizzle rain
312	heavy intensity drizzle rain
313	shower rain and drizzle
314	heavy shower rain and drizzle
321	shower drizzle
500	light rain
501	moderate rain
502	heavy intensity rain
503	very heavy rain
504	extreme rain
511	freezing rain

Tabelle 3.1: Tabelle mit allen Wetterverhältnisse

3.4 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 3.3 ist das Schema von der ersten Version abgebildet.

Hier etwas darüber erläutern

In Abbildung 3.4 dagegen ist 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.

Bei dem Umzug des Data Miners mitsamt der Datenbank auf einen neuen Server, mussten 10 Gigabyte an Datenbank migriert werden.

Diese Aufgabe war schwieriger als erwartet, da Im- und Export mehrere Stunden Zeit in Anspruch nehmen und nicht exportierte Einträge des Miners während des Umzuges mit dem neuen Server synchronisiert werden müssen. Dies ist bei einer Datenbanktabelle, welche dauerhaft mehrere Transaktionen des Miners erfährt, sehr mühsam umzusetzen (warum? vielleicht noch erläutern). Um den Prozess so schonend wie möglich zu gestalten, wurde ein Skript geschrieben, das nach der fertigen Migration der Datenbank die Tabellen miteinander synchronisiert. Ein MySQL Sharding mit Master- und Slave-Modus war aufgrund inkompatibler Versionen nicht möglich. Nach der Synchronisation der Tabellen durch das Skript wurde der alte Miner gestoppt und der Miner auf dem neuen Server gestartet. Die Downtime des Miners betrug nur circa 60 Sekunden. Ein Fehler in der Installation...

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,

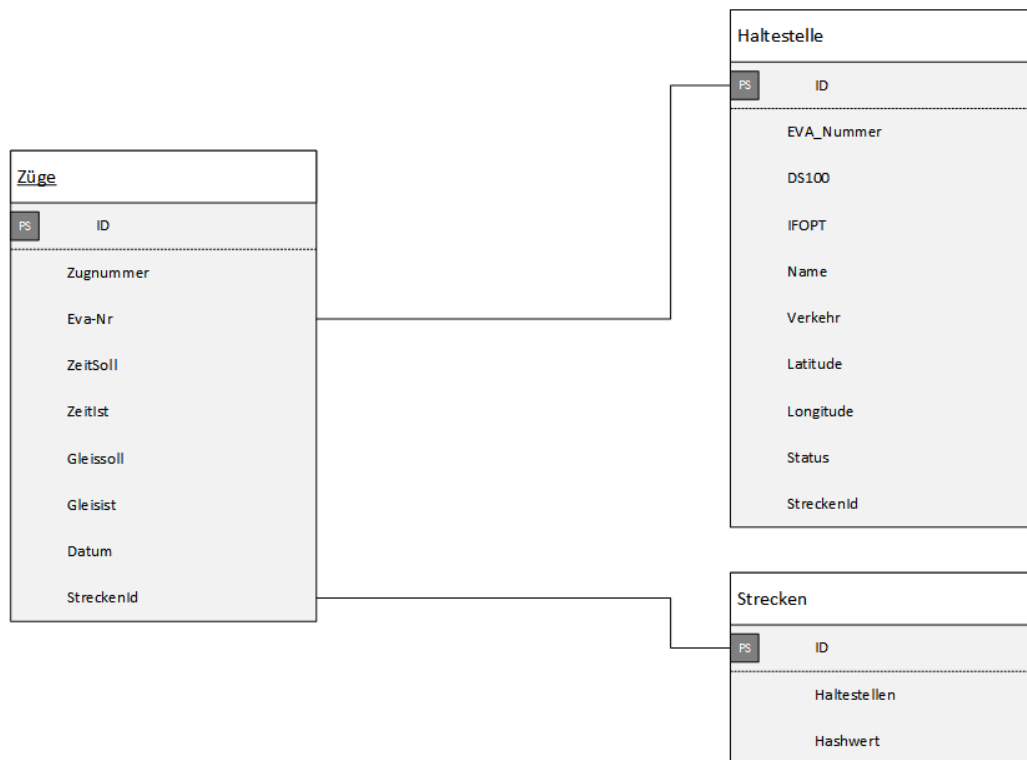


Abbildung 3.3: Datenbank Schema Version 1

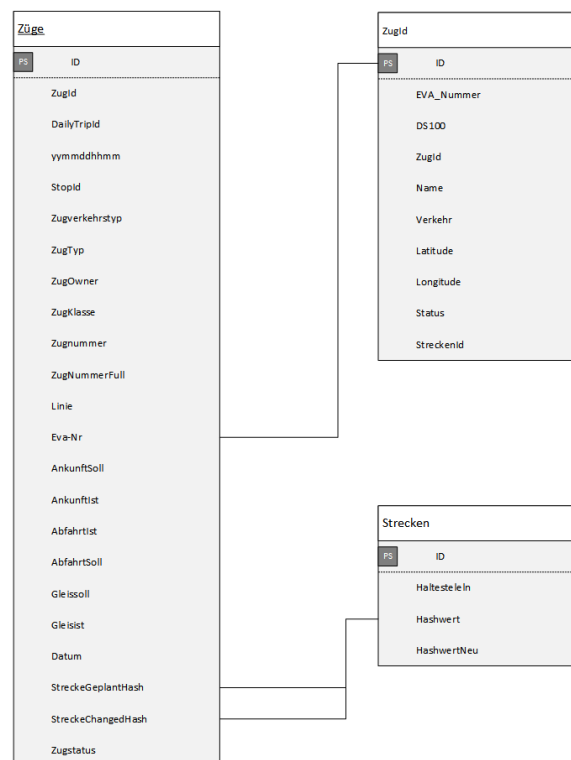


Abbildung 3.4: Datenbank Schema Version 1

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 .

3.4.1 Datenbank Schema des Wetterminers

Für das Datenbankschema wurden weitestgehend die Felder der Server Antwort als Spalten der Datenbank übernommen. In der Spalte Time wird ein Timestamp abgespeichert, an dem der Datensatz aufgenommen wurde. Die Spalte PLR wird genutzt, um die Postleitregion abzulegen, Temperatur, Luftfeuchtigkeit, Luftdruck, Windgeschwindigkeit und Windrichtung werden in je einer Spalte gespeichert. Da mehrere Wetterlagen in einem Datensatz möglich sind, musste für die Speicherung der Wetterlagen eine zusätzliche Tabelle angelegt werden, die die Verlinkung zwischen Datensatz und Wetterlage herstellt. Diese Speichert die ID des Datensatzes und die ID der Wetterlage.

3.5 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 Backups eine große Transaktion im Cache oder auf der Festplatte zwischengespeichert werden. Zum anderen ist durch die Menge der Datensätze 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 Datenbankimport auf funktionierende Constraints und geprüft werden.

watt?

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 den 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.

Um die Erstellten Dateien wieder in eine Datenbank einzufügen, wird ein kleines Script

implementiert, dass mit dem Befehl 3.2 die Dateien nach und nach wieder einfügt. Der Befehl muss aufgrund seiner langen Laufzeit auf dem Server laufen. Web Oberflächen würden den Befehl nach einiger Zeit abbrechen, da sich der Server längere Zeit nicht meldet.

```
1      mysql -u test Zuege < dump-split-$fn.sql || echo "  
      error_in_file  
      ***** "
```

Quellcode 3.2: Befehl zum Einfügen der SQL Dateien

Listing mit splitfile.php

Kapitel 4

Datenverarbeitung mit Data Mining

4.1 Grundlagen von Data Science und KDD

4.1.1 Data Mining

Knowledge Discovery

4.1.2 Knowledge Discovery in Databases (KDD)

Das Ziel von KDD ist es, das menschliche Vermögen, Daten zu analysieren und zu untersuchen, zu steigern, indem die Datenanalyse automatisiert wird. Die Automatisierung ist nötig, um mit den immer größer werdenden Datenmengen umgehen zu können. [USAMA FAYYAD 1996, S. 39]

Der KDD-Prozess ist ein mehrstufiger Prozess, der sich nach [] in folgende Schritte untergliedern lässt:

zitat leer

KDD verwendet zur Analyse von Daten auch das Data Mining. Jedoch wird ein Unterschied zwischen KDD und Data Mining gemacht: Data Mining ist das Anwenden von spezifischen Algorithmen auf Daten, um Muster in den Daten zu finden. KDD hingegen ist ein allumfassender Prozess, um Wissen aus den Daten zu gewinnen. [USAMA FAYYAD 1996, S. 39]

4.2 Grundlagen

Hier noch text

Begriffsdefinition

Statistische Methoden

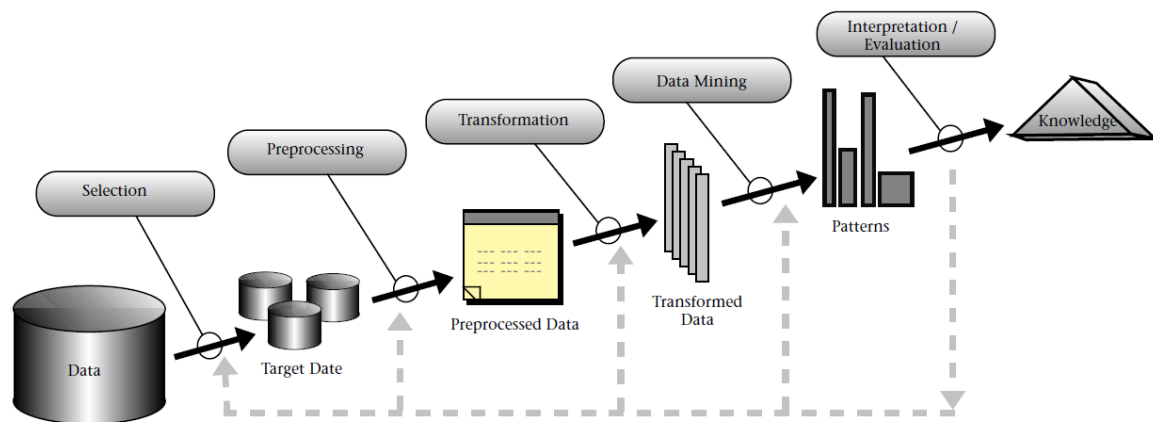


Abbildung 4.1: Übersicht des KDD-Prozesses [USAMA FAYYAD 1996, S. 41].

Machine Learning

Visualisierungsmethoden

Denkbare Auswertungen - Verspätungsarten pro Linie - davon der Durchschnitt über mehrere Züge die die gleiche Linie zu verschiedenen Zeiten befahren - damit können Heatmaps erzeugt werden

4.3 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

Dieses Kapitel bezieht sich auf den Tabellen cache vorgang, also generierung einer neuen besser select baren tabelle, das Quellcode listing ist noch nicht korrekt

```

1 import sys
2 import argparse
3 from glob import glob
4 import os
5 import datetime
6 import json
7 import re
8 FLAGS, unparsed = parser.parse_known_args()
9
10 def timetotimeint(input):
11     input = str(input)
12     if input == 'None':
13         input = "24:00:00"
14     hhmss = input
15     (h, m, s) = hhmss.split(':')
16     result = int(h) * 60 + int(m)
17     result = math.floor(result/5)
18     return result
19
20 def openvocalfile(name):
21     lines = []
22     filename = "./vocabfiles/" + str(name) + ".txt"
23     with open(filename, mode="w+", encoding="utf-8") as file:
24         for line in file:
25             line = line.rstrip('\n')
26             lines.append(line)
27     return lines
28

```

Quellcode 4.1: Some Python File

Ebenfalls für Abfragen ineffizient, ist der "Primary key" der Deutschen Bahn. Dieser besteht, wie in

da scheint es ein problem mit der Referenzierung zu geben

?? beschrieben aus drei teilen. Um eine Abfrage der Datenbank auf einen bestimmten Zug zu machen muss die Daily Trip ID sowie das Datum im "yymmddhhmm"Format angegeben werden. Diese Werte müssten dann allerdings von der Datenbank mit dem String der Zug Id verglichen werden. Das ist Natürlich nicht sehr performant. Deshalb wurde entschieden, dass die Datenbankstruktur um drei spalten erweitert wird. Da sie im Nachhinein hinzugefügt worden sind müssen alle schon vorhandenen Einträge bearbeitet werden. Dafür wurde ein Algorithmus geschrieben, der die Zugid aus der Datenbank ausliest und wie in 4.2 beschrieben gesplittet. Diese Komponenten werden dann in die jeweiligen Spalten inseriert.

```

1     if temp.empty:
2         print("actual_id_{}_{}_is_{}".format(actual_id))
3         missing_IDs.write('{}\n'.format(actual_id))
4         missing_IDs.flush()
5     else:
6         # check if id is already filled
7         if temp["stopid"][0] is None:
8             # split zugid in komponenten
9             zugid = temp["ttsid"][0]
10            print("actual_id_{}_{}_t_zugid:{}_{}\n\r".format(
11                actual_id, zugid))
12            zugid = zugid.split("-")
13            # if first komponent is empty dailytripid was
14            negative
15            try:
16                if zugid[0] == "":
17                    zugid[1] = int(zugid[1]) * (-1)
18                    qs.insert_3tuple_with_id(actual_id, zugid[
19                        1], zugid[2], zugid[3])
20                else:
21                    qs.insert_3tuple_with_id(actual_id, (zugid
22                        [0]), zugid[1], zugid[2])
23            except ConnectionResetError:

```

Quellcode 4.2: Zerlegen der Zug ID in seine Komponenten

4.4 Software-Architektur der Datenauswertung

In diesem Abschnitt wird kurz die Software-Architektur dargestellt, die bei dem Data Mining angewandt wird. In Abbildung 4.2 zeigt sich, wie der allgemeine Ablauf des Datenabrufes mit der Architektur zusammenhängt: Grundsätzlich müssen die Daten zuerst beschafft werden, bevor sie ausgewertet werden können. Aus diesem Grund werden zuerst die nötigen Daten aus der Datenbank abgerufen. Für die Datenbankabfrage wird die Python-Klasse “QuerySuite” eingeführt. Die “QuerySuite” gibt die Daten der Abfrage in Dataframes zurück, die geeignet für den Transport der Daten sind. Nach der Abfrage werden die Daten weiterverarbeitet. Zu diesem Zweck wird das Python-Package “ProcessingUtils” konzipiert, das Funktionen enthält, um Berechnungen mit den Dataframes ausführen zu können. Im Folgenden wird genauer auf die genannten Komponenten eingegangen.

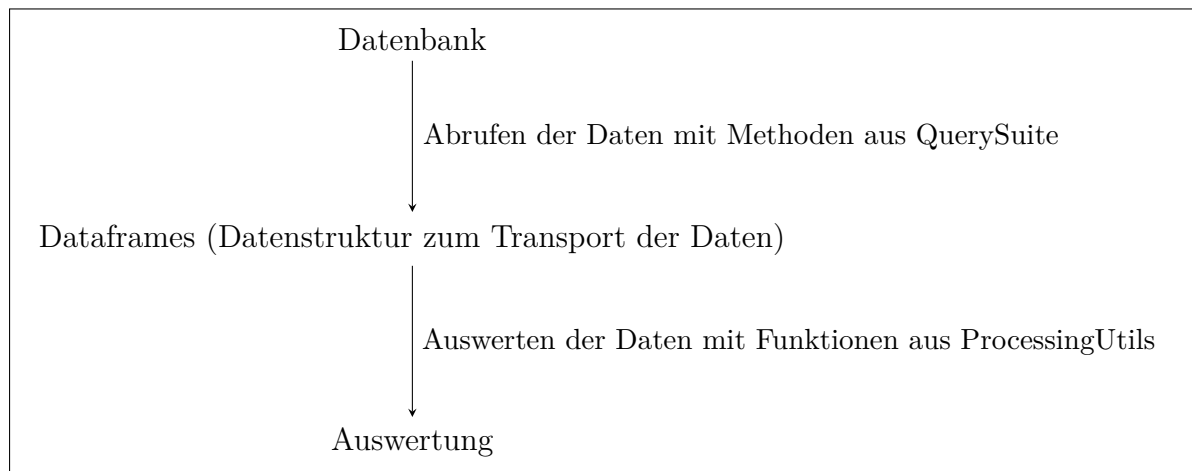


Abbildung 4.2: Grundablauf des Data Minings

4.4.1 Query Suite

Kapselung der SQL-Queries Da je nach Auswertung verschiedene Datensätze aus der Datenbank gebraucht werden, sind auch mehrere SQL-Queries notwendig, um diese Daten von der Datenbank abzufragen. Zu diesem Zweck werden die Abfragen von der Klasse “QuerySuite” durchgeführt. Die Klasse enthält für jede SQL-Abfrage eine spezielle Methode, die diese Abfrage durchführt. Die Abfragen werden in Methoden gekapselt, um das Programmieren übersichtlicher zu gestalten. Jede Methode kann zusätzliche Parameter bei Aufruf entgegennehmen, die dann in der SQL-Abfrage verwendet werden können. Auf diese Weise müssen die SQL-Abfragen nicht ständig kopiert und abgeändert werden, wenn sich nur die Parameter, nicht jedoch die Struktur der Abfrage ändern. Diese Maßnahme verhindert Code-Duplication, die in der Software-Entwicklung vermieden werden soll. Sollte sich während der Entwicklung beispielsweise das Datenbankmodell ändern, so müssen nur die einzelnen SQL-Abfragen, die in den jeweiligen Methoden der QuerySuite gekapselt sind, angepasst werden, anstatt nach jeder SQL-Abfrage in jeder Auswertung zu suchen und anzupassen.

Verwaltung der Verbindung Die Klasse ist neben den Abfragen auch verantwortlich für die dafür benötigte Verbindung zur Datenbank. Der Klasse wird einmalig die zu verwendende Datenbankverbindung übergeben, die für alle nachfolgenden Aufrufe der Abfrage-Methoden verwendet werden. Bevor die SQL-Queries an die Datenbank gestellt werden, fügt die Klasse noch eine “LIMIT”-Klausel an die auszuführende SQL-Query an, sodass die Anfragen implizit begrenzt werden. Während der Entwicklung ist das ein wichtiges Feature, um bei fehlerhaften Abfragen die Datenbank nicht unnötig zu überlasten. Werden von einer Query mehr Datensätze erwartet, als es die Begrenzung zulässt, so wird für diese spezielle Query die Begrenzung davor explizit als aufgehoben.

Umformatierung der Datensätze in ein Dataframe Nach der Query liefert die Datenbank die Ergebnisse. Die Python-Bibliothek PyMySQL, die für die Datenbankinteraktionen verwendet wird, speichert hierbei die Ergebnisse in einem zweidimensionalen Python-Tupel. Das Tupel enthält zeilenweise weitere Tupel, die die einzelnen Datensätze enthalten. Die inneren Tupel speichern jeweils die Elemente eines Datensatzes in der Reihenfolge der Tabellenspalten aus der Datenbank. Für die Weiterverarbeitung der Datensätze ist das Datenformat des zweidimensionalen Tupels jedoch nachteilig. Dies liegt daran, dass Tupel dafür gedacht sind, Folgen von Elementen mit einem Reihenfolgencharakter zu speichern. Um zu wissen, welches Datum in welcher Spalte und Zeile des Tupels welcher Tabellenspalte aus der Datenbank entspricht, muss die vorangegangene Query bekannt sein. Die nachfolgende Datenverarbeitung soll also nicht eine Reihenfolge von Elementen als Eingabedaten erhalten, sondern ein Datenformat erhalten, das dem Datenmodell der Datenbank nachempfunden wird.

Zu diesem Zweck wird die Python-Bibliothek Pandas verwendet. Pandas ermöglicht es, sogenannte Dataframes zu erzeugen. Ein Dataframe ist eine tabellenähnliche Datenstruktur, sodass die enthaltenen Daten nach Spalten und Zeilen organisiert werden. Die Spalten können hierbei mit Labels und die Spalten mit Indizes versehen werden.

hier und im ff Indizes? Spalten und Spalten?

Auf diese Weise können die Ergebnisse der Query in Dataframes eingefügt werden, sodass die Spalten der Dataframes auch die entsprechenden Namen der Datenbanktabellen tragen. Das Umformatieren der Query-Daten in ein Dataframe hat den Vorteil, dass die nachfolgende Datenverarbeitung die gewünschten Daten nach den Tabellennamen und den Indizeswerten auflösen kann und nicht auf eine feste Reihenfolge von Elementen in einem Tupel angewiesen ist. Aus diesem Grund werden Dataframes als Datenstruktur verwendet, um die Daten während der Verarbeitung zu transportieren.

Beispiel In Quellcode 4.3 ist als Beispiel “get_tts_by_ttsid” als eine der Query-Methoden in der QuerySuite-Klasse aufgelistet. An diesem Beispiel soll gezeigt werden, wie die oben genannten Punkte sich in dem Quellcode äußern. In Zeile 2 werden zunächst die Labels beschrieben, die für die Benennung der Dataframe-Spalten verwendet werden. In Zeile 9 steht die Methode “_get_tts_by_ttsid_query”. In Zeile 10 fügt die Methode, den übergebenen Parameter “ttsid” in die SQL-Query ein. In Zeile 12 findet über die Methode “_do_query” die Ausführung der Query statt.

In Zeile 15 wird die Methode “get_tts_by_ttsid” definiert. In Zeile 19 erhält die Methode das Query-Ergebnis als Tupel zurück. Anschließend wird in Zeile 20 das Tupel in ein Dataframe umformatiert. Wie zu sehen ist, wird vor der Erzeugung des Tupels das zweidimensionale Tupel mittels der Funktion “list” in eine Liste von eindimensionalen Tupeln konvertiert, damit Pandas die Struktur der Daten richtig deuten kann. Bei der Erzeugung werden der “DataFrame”-Funktion die zuvor definierten Labels für die Spalten mitgegeben.

```

1 # labels for table "time table stops" (named "zuege" in
   database)
2 TABLE_LABELS_TTS = ["id", "ttsid", "dailytripid",
3   "yymmddhhmm", "stopindex", "zugverkehrstyp", "zugtyp",
4   "zugowner", "zugklasse", "zugnummer", "zugnummerfull",
5   "linie", "evanr", "arzeitsoll", "arzeitist", "dpzeitsoll",
6   "dpzeitist", "gleissoll", "gleisist", "datum",
7   "streckengeplanthash", "streckenchangedhash", "zugstatus"]
8
9 def _get_tts_by_ttsid_query(self, ttsid):
10     query = "SELECT_*_FROM_zuege_WHERE_zuege.zugid_=_\"{}\" \" \
11         .format(ttsid)
12     result = self._do_query(query)
13     return result
14
15 def get_tts_by_ttsid(self, ttsid):
16     """
17     Retrieves full row of database table by given 'ttsid' (
       named 'zugid' in database).
18     """
19     result = self._get_tts_by_ttsid_query(ttsid)
20     result_df = pd.DataFrame(data=list(result), columns=
        TABLE_LABELS_TTS)
21     return result_df

```

Quellcode 4.3: Beispiel einer Query-Methode

4.4.2 Processing Utils

Das Package “ProcessingUtils” beinhaltet verschiedene Python-Funktionen, die bei den unterschiedlichen Datenauswertungen benutzt werden. Die Funktionen operieren auf den von der QuerySuite gelieferten Dataframes. Das Package dient nur dazu, die verschiedenen Funktionen zu gruppieren, die für die Datenauswertung geschrieben werden. Wie die Funktionen schließlich kombiniert werden, hängt von der jeweiligen Auswertung ab, die Gebrauch von diesen Funktionen machen kann.

4.5 Statistische Auswertungen

4.5.1 Definition der Verspätungen

Für die Auswertung der Daten ist die Verspätung eine interessante Größe. Hierbei können verschiedene Verspätungen definiert und in dem Datenbestand untersucht werden. In diesem Abschnitt werden verschiedene Verspätungsarten definiert und anschließend dargestellt, wie diese in dem Datenbestand analysiert werden.

Verspätung bei Ankunft Die Verspätung der Ankunft Δan eines Zuges zug_m im Bahnhof bhf_n ist definiert als

$$\Delta an(bhf_n, zug_m) := an_{real}(bhf_n, zug_m) - an_{plan}(bhf_n, zug_m) \quad (4.1)$$

Verspätung bei Abfahrt Die Verspätung der Abfahrt Δab eines Zuges zug_m im Bahnhof bhf_n ist definiert als

$$\Delta ab(bhf_n, zug_m) := ab_{real}(bhf_n, zug_m) - ab_{plan}(bhf_n, zug_m) \quad (4.2)$$

Geplante Haltedauer Die geplante Haltedauer lässt sich mit folgender Formel ermitteln:

$$halten_{plan}(bhf_n, zug_m) := ab_{plan}(bhf_n, zug_m) - an_{plan}(bhf_n, zug_m) \quad (4.3)$$

Reale Haltedauer Die reale Haltedauer lässt sich mit folgender Formel ermitteln:

$$halten_{real}(bhf_n, zug_m) := ab_{real}(bhf_n, zug_m) - an_{real}(bhf_n, zug_m) \quad (4.4)$$

Verspätung durch Haltedauer Hier werden die zwei zuvor aufgestellten Formeln für die geplante und reale Haltedauer wiederaufgegriffen und mit ihnen die Verspätung definiert, die durch eine zu lange außerplanmäßige Haltedauer entsteht.

$$\Delta halten(bhf_n, zug_m) := halten_{real}(bhf_n, zug_m) - halten_{plan}(bhf_n, zug_m) \quad (4.5)$$

Geplante Fahrtdauer Zum Ermitteln der geplanten Fahrtdauer werden nun zwei Bahnhöfe benötigt. Der eine ist der Start-Bahnhof (bhf_{n-1}) und der andere ist der Zielbahnhof (bhf_n), mit denen nun über die Ankunfts- und Abfahrtszeit die Fahrtdauer ermittelt werden kann:

$$fahren_{plan}(bhf_{n-1}, bhf_n, zug_m) := an_{plan}(bhf_n, zug_m) - ab_{plan}(bhf_{n-1}, zug_m) \quad (4.6)$$

Reale Fahrtdauer Die reale Fahrtdauer wird nach dem gleichen Schema wie die geplante Fahrtdauer berechnet:

$$fahren_{real}(bhf_{n-1}, bhf_n, zug_m) := an_{real}(bhf_n, zug_m) - ab_{real}(bhf_{n-1}, zug_m) \quad (4.7)$$

Verspätung durch Fahrtdauer Aus der geplanten und der realen Fahrtdauer, lässt sich nun die Verspätung, die sich durch eine außerplanmäßige Fahrtdauer ergibt, ermitteln:

$$\Delta fahren(bhf_{n-1}, bhf_n, zug_m) := fahren_{real}(bhf_{n-1}, bhf_n, zug_m) - fahren_{plan}(bhf_{n-1}, bhf_n, zug_m) \quad (4.8)$$

Beispiel Die Berechnungen der verschiedenen Verspätungen sind mit der Programmiersprache Python implementiert. Folgendes Beispiel 4.4 zeigt die Funktion, die die Verspätung bedingt durch die Fahrtzeit berechnet wird.

Zu Beginn erhält die Funktion Parameter, die bestimmen, welcher Abschnitt einer Route untersucht werden soll. Hierfür bekommt die Funktion zwei Dataframes übergeben. Im ersten Dataframe wird der Startpunkt des Abschnitts angegeben (`“train_stop_from_df”`) und im zweiten Dataframe wird der Zielpunkt des Abschnitts angegeben (`“train_stop_to_df”`).

Bevor mit der Berechnung der Verspätung begonnen wird, validiert die Funktion zunächst, ob die nötigen Parameter übergeben wurden. Es kann während der Analyse einer ganzen Route vorkommen, dass der Startpunkt oder Zielpunkt eines Abschnittes nicht angegeben werden und somit den Wert `“None”` haben.

Im nächsten Schritt wird das `“ttsid”`-Attribut, das einen Haltepunkt eindeutig identifiziert, von Start- und Zielpunkt ausgelesen. Die `“ttsid”`-Attribute von Start- und Zielpunkt wird zum Schluss im Ergebnis-Dataframe zusammen mit der berechneten Verspätung gespeichert, um die Verspätung dieser Strecke in späteren Auswertungen zuordnen zu können.

Im nächsten Schritt wird nun die Verspätung, die bedingt durch die Abweichung von der geplanten Fahrtzeit auftritt, berechnet. Auch hier wird überprüft, ob beide Punkte definiert wurden. Ist ein Punkt nicht angegeben worden (und besitzt somit den Wert `“None”`), so wird die Verspätung mit einem Wert von `“NaT”` (Not a Time) angegeben, um zu signalisieren, dass für den gegebenen Start- und Zielpunkt keine Verspätungsberechnung durchgeführt werden konnte. Die Repräsentation von nicht vorhandenen Verspätungen mittels `“NaT”` ist hierbei sinnvoll, da diese Werte bei der späteren Weiterverarbeitung oder Visualisierung ignoriert werden können.

Im letzten Schritt wird der Ergebnis-Dataframe konstruiert erstellt. In der Spalte `“travel_time_real”` speichert dieser, die ermittelte Verspätung. In der Spalte `“ttsid_from”` und `“ttsid_to”` werden die `“ttsid”`-Attribute des Start- und Zielpunktes gespeichert. Der befüllte Dataframe wird schließlich an den Aufrufenden der Funktion zurückgegeben.

Beispiel ausführen

```

1 def calc_delay_by_traveltime_df(train_stop_from_df,
2   train_stop_to_df):
3     """
4     Calculates the delay that has been caused by the travel of
5     the train.
6     Positive value means, that the travel time caused
7     additional delay.
8     Negative value means, that the travel time decreased the
9     delay.
10    :param train_stop_from_df: Pandas dataframe. Input for the
11    train stop the train comes from.
12    :param train_stop_to_df: Pandas dataframe. Input for the
13    train stop the train arrives at.
14    :return: Returns a pandas dataframe with columns '
15    delay_by_traveltime', 'ttsid_from', 'ttsid_to'.
16    """
17    if train_stop_from_df is None:
18        ttsid_from = None
19    else:
20        ttsid_from = train_stop_from_df["ttsid"].iloc[0]
21
22    if train_stop_to_df is None:
23        ttsid_to = None
24    else:
25        ttsid_to = train_stop_to_df["ttsid"].iloc[0]
26
27    if train_stop_from_df is None or train_stop_to_df is None:
28        delay = pd.NaT
29    else:
30        traveltime_real = calc_traveltime_real_df(
31            train_stop_from_df, train_stop_to_df)["
32            traveltime_real"].iloc[0]
33        traveltime_scheduled = calc_traveltime_scheduled_df(
34            train_stop_from_df, train_stop_to_df)["
35            traveltime_scheduled"].iloc[0]
36        delay = traveltime_real - traveltime_scheduled
37
38    result = pd.DataFrame(
39        data=[[delay, ttsid_from, ttsid_to]],
40        columns=["delay_by_traveltime", "ttsid_from", "
41            ttsid_to"])
42    return result

```

Quellcode 4.4: Berechnung der SARV

4.5.2 Analyse der Verspätungen eines Zuges

Mit den in Abschnitt 4.5.1 definierten Verzögerungen ist es bereits möglich, erste statistische Auswertungen auszuführen. In diesem Abschnitt wird dargestellt, wie die Verspätungsarten eines Zuges entlang seiner Route analysiert werden. Hierbei werden bei der Analyse diese Verspätungsarten berücksichtigt.

- Verspätung bei Ankunft
- Verspätung bei Abfahrt
- Verspätung durch Haltezeit
- Verspätung durch Fahrtzeit

eventuell noch absätze einfügen um Lesbarkeit zu verbessern, neue Absätze mit „/“
um Einrückungen zu vermeiden

Beispiel

Als Beispiel ist nun Abbildung 4.3 zu betrachten, die den RE4725 von Karlsruhe Hbf. nach Konstanz zeigt. Anhand des Beispiel wird die Bedeutung von Visualisierungen im Data Mining recht deutlich, da hier wichtige Fakten schnell erkannt werden können. Die blaue Kurve ist sehr auffällig und zeigt, dass die Verspätung bei Ankunft ab Offenburg kurz aber stark zunimmt und erst nach Donaueschingen wieder abnimmt. Das Maximum der Verspätung bei Ankunft liegt bei +5 Minuten, das bei den Stationen Haslach, Hausach, Hornberg und Donaueschingen erreicht wird. Wird die orangene Kurve betrachtet, wird auch deutlich, weshalb die Verspätung kurz nach Offenburg zunimmt: Der Zug wartet unplanmäßig lange in Offenburg (+4 Min.) und verursacht so die Verspätung. Ansonsten ist gut zu sehen, dass die orangene Kurve sehr wenig schwankt und die Verspätung durch Haltezeit für die weitere Fahrt des Zuges keine wesentliche Ursache für weitere Verspätungen darstellt. Wird nun die rote Kurve betrachtet, so ist auch hier zu erkennen, dass die Verspätung durch Fahrtzeit im Mittel keine Verspätungen verursacht. Zwischen Hornsberg und Allensbach ist die Fahrzeit dafür verantwortlich, dass die in Offenburg verursachte Verspätung wieder abgebaut wird. Jedoch zeigt sich noch eine Besonderheit: Zwischen Konstanz-Petershausen und Konstanz nimmt die Verspätung des Zuges (bei Ankunft) wieder deutlich zu, da der Zug für die Strecke außerplanmäßig +5 Minuten länger benötigt. Zuletzt ist die violette Kurve zu betrachten. Hier ist zu sehen, dass die Verspätung bei Abfahrt kaum Mehrwert liefert, da sie größtenteils Deckungsgleich mit der Verspätung bei Ankunft ist.

Nun sollen Interpretationen dargestellt werden, die aus den visualisierten Daten gewonnen werden können. Die erste Interpretation lautet, dass Offenburg womöglich ein Bahnhof ist, der oft Verspätungen durch zu lange Haltezeit aufbaut. In Offenburg werden viele Verbindungen zu anderen Zügen hergestellt, sodass Abweichungen von der geplanten Wartedauer plausibel erscheinen, wenn ein Zug auf einen anderen Verbindungszug warten muss. Um dieser Vermutung nachzugehen, wäre es eine Idee, den Durchschnitt der Verspätung durch Verspätung von mehreren Zügen zu bestimmen, die in Offenburg halten. Wenn

der Durchschnitt eine erhebliche Verspätung zeigt, könnte damit die Hypothese gestützt werden, dass das Warten auf Verbindungszüge ein möglicher Grund für Verspätungen ist.

Eine andere Interpretation ist, dass die Strecke zwischen Konstanz-Petershausen und Konstanz möglicherweise auch ein Grund für Verspätungen ist. Hier ist eine Durchschnittsberechnungen von mehreren Zügen auch eine gute Idee, um dieser Vermutung nachzugehen. Mögliche Gründe für Verspätung auf einer Strecke könnten sein, dass der Zug beispielsweise auf Signalfreigabe zur Fortsetzung der Fahrt warten muss.

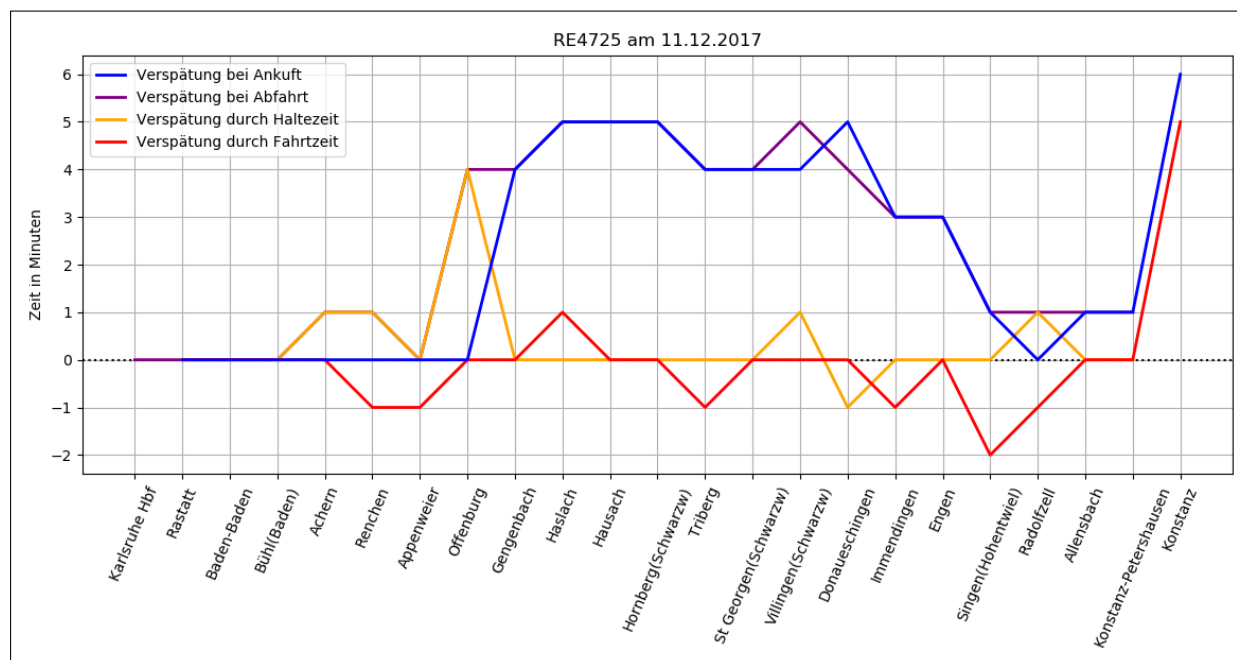


Abbildung 4.3: Verspätungsanalyse von RE4725 von Karlsruhe nach Konstanz

Verspätung bei Abfahrt liefert kaum Mehrwert Aus der zuvor gezeigten Abbildung 4.3 wurde die Beobachtung gemacht, dass die Verspätung bei Abfahrt größtenteils deckungsgleich mit der Verspätung bei Ankunft ist und deshalb kaum neue Informationen liefert. Auch in Abbildung 4.4 kann sehr gut beobachtet werden, dass sich der Verlauf der Verspätung bei Abfahrt nur geringfügig von dem Verlauf der Ankunft unterscheidet.

Die Verspätung bei Abfahrt trägt auch an sich kaum Bedeutung: Es ist bei der Analyse der Verspätungen vorwiegend die Verspätung bei Ankunft interessant. Schließlich ist es für einen Fahrgast wichtig, dass ein Zug pünktlich am Start- und Zielbahnhof ankommt. Ob der Zug verspätet aus einem Bahnhof abfährt ist dem Fahrgast gleichgültig, solange sich die Ankunft nicht verspätet.

Aus diesen Gründen wird die Verspätung bei Abfahrt nicht mehr in der Darstellung der Verspätungen berücksichtigt. Dies hat auch den Vorteil, dass die Visualisierungen der Verspätungen übersichtlicher wird. Dies kann in Abbildung 4.5 nachvollzogen werden, die

den gleichen Zug wie in Abbildung 4.4 zeigt, aber die Verspätung bei Abfahrt ausgelassen wird.

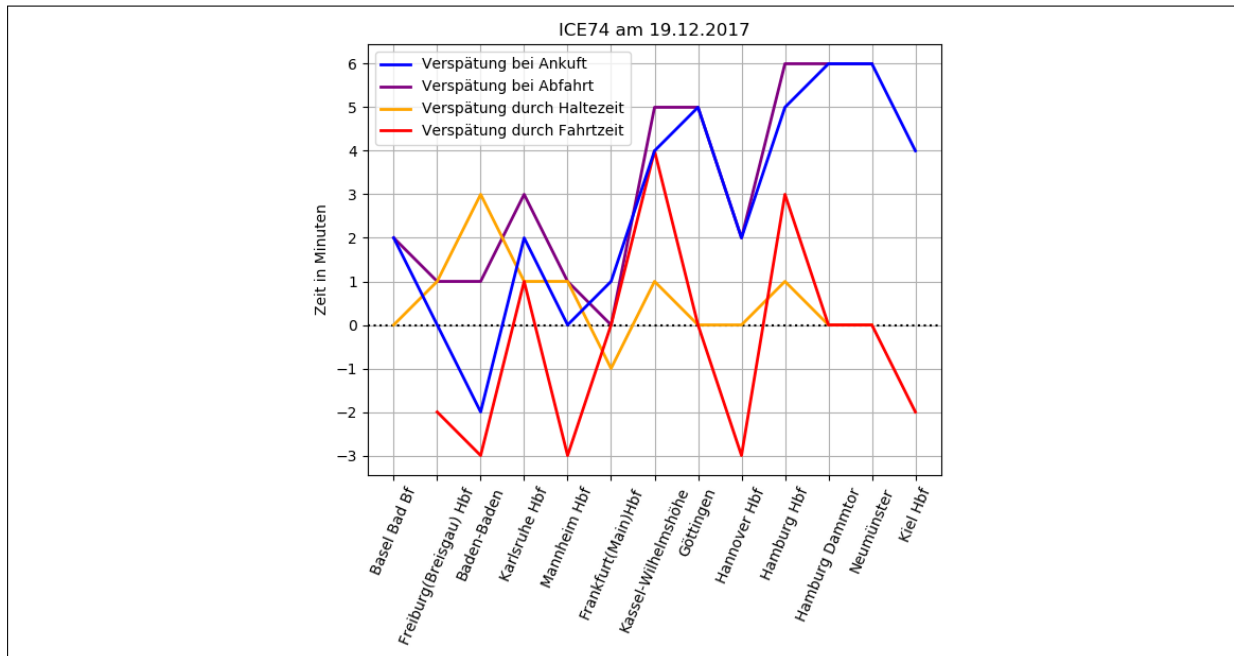


Abbildung 4.4: Verspätungsanalyse von ICE74 von Basel Bad. Bf. nach Kiel mit Verspätung bei Abfahrt

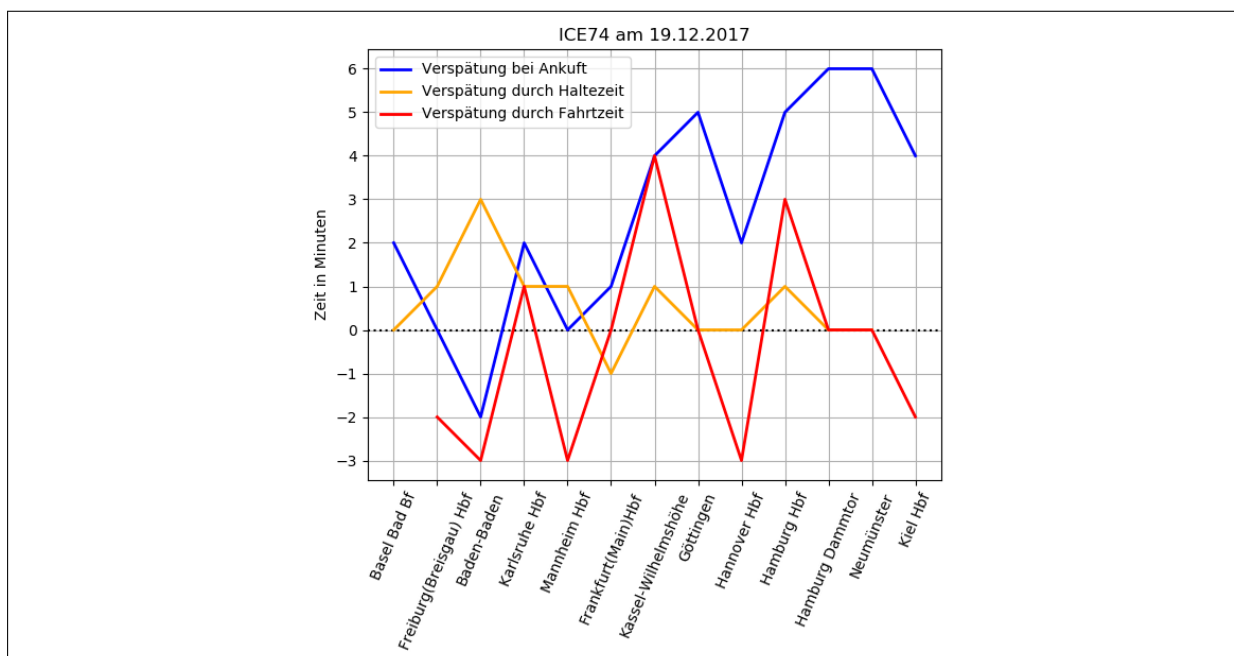


Abbildung 4.5: Verspätungsanalyse von ICE74 von Basel Bad. Bf. nach Kiel mit Verspätung bei Abfahrt

Zusammenhang der Verspätungen Somit verbleiben drei Kategorien von Verspätungen, die Verspätungen durch Haltezeit und Fahrzeit sind hierbei wichtig, um nachvollziehen zu können, ob der Grund der Verspätung bei Ankunft in der Haltezeit oder Fahrzeit liegen. Somit reduziert sich

$$\Delta an(b_n, z_m) = \Delta an(b_{n-1}, z_m) + \Delta halten(b_{n-1}, z_m) + \Delta fahren(b_{n-1}, b_n, z_m) \quad (4.9)$$

Zusammenhang der Verspätungen näher ausführen

4.6 Stochastische Analyse

Viele der Informationen können auch ohne Neuronale Netze gewonnen werden. Beispielsweise kann aus der wirklichen Abfahrtszeit des vorherigen Bahnhofes und der wirklichen Ankunftszeit des Nächsten Bahnhofes die Fahrzeit der Züge berechnet werden. Des Weiteren kann durch die Ankunftszeit und der Abfahrtszeit des Zuges auch die Zeit bestimmt werden, die der Zug benötigt, um eine bestimmte Strecke zu fahren.

4.6.1 Durchschnitt der Zeiten einer Haltestelle

eventuell noch absätze einfügen um Lesbarkeit zu verbessern hier und im ff

An einer Haltestelle können Verspätungen an einer Strecke festgemacht werden. In diesem Abschnitt werden die durchschnittliche Verspätung bei Ankunft, die durchschnittliche Verspätung bei Abfahrts und die durchschnittliche zusätzliche Standzeit der Züge in einer Haltestelle berechnet. In der Ersten Version wurde für jede Berechnung der Werte alle Werte in der Datenbank verwendet. Die Verspätungen werden addiert, und später durch die Anzahl der Summanden geteilt. Diese Berechnungsweise führt allerdings zu dem Problem, dass die Berechnung sehr lange dauert. Alleine für die Berechnung mit einem Drittel der Daten dauert 6-7h. Da das nicht praktikabel ist, muss diese Funktion performanter gemacht werden. Um dies zu erreichen, wird ein Speicher angelegt, in dem die Daten der Durchschnittsberechnung abgespeichert werden kann. Dadurch ist es nicht mehr nötig, immer alle Halte einer Haltestelle abzufragen, damit die Durchschnitte berechnet werden können. Stattdessen werden die in Tabelle 4.1 gezeigten Daten in der Datenbank abgespeichert. Die Spalte ID wird lediglich als Primary Key benutzt. Die EVA Nummer wird verwendet, um die Durchschnitte den Haltestellen zuzuweisen. Die Description ID wird verwendet, um die verschiedenen Durchschnitte der Stationen zu den Strings Ankunfts Verspätung, Abfahrts Verspätung und zusätzliche Standzeit zuzuweisen. Diese sind in einer weiteren Tabelle in der Datenbank abgelegt. Die Ziffer eins entspricht dementsprechend dem String Ankunfts Verspätung, die zwei Abfahrts Verspätung und die drei zusätzliche Standzeit. Dadurch kann die Homepage mit wenig Aufwand alle verfügbaren Durchschnitte anfordern und verarbeiten. In der Spalte Average der Tabelle 4.1 wird der aktuelle Durchschnitt abgelegt. Die aktuelle Anzahl der verwendeten Datensätze wird durch die Spalte Count abgebildet. In Last Value wird die Letzte ID abgelegt.

Diese wird verwendet, um alte Einträge der Datenbank von den neuen zu trennen. Um die Durchschnitte zu berechnen, werden die Zustände der jeweiligen Durchschnitte aus der Datenbank abgefragt. Die Werte Average und Count werden zu der letzten Summe multipliziert.

$$\text{Summe} = \text{Average} * \text{Count} \quad (4.10)$$

Count und Last Value werden lediglich abgespeichert. Nun werden die Datensätze aus der Datenbank abgefragt. Um nicht alle Datensätze der Datenbank abrufen zu müssen, wird die letzte ID (Last Value) verwendet. Sie wird als Bedingung in der Datenabfrage 4.5 eingesetzt. Dadurch werden nur die Datensätze aus der Datenbank geladen, die eine höhere ID besitzen als die der letzten ID. Die im Pandas Datenframe formatierten Daten werden nun zeilenweise durchlaufen, und die jeweiligen Summen berechnet sowie Counts erhöht. Sind alle Reihen durchlaufen, sollte die Summe sowie der Counter null sein, wird kein Durchschnitt berechnet. Sind dagegen Werte vorhanden, werden nacheinander die verschiedenen Durchschnitte berechnet. Die neuen Werte werden dann direkt wieder in die Datenbank geschrieben. Ein weiterer Vorteil dieser Implementierung, ist dass die Werte nicht aus einer Json oder CSV Datei eingelesen werden müssen, sondern direkt von der Homepage aus der Datenbank geladen werden können. Die Description ID gibt dabei Aufschluss über die Art des Durchschnittes.

Mit diesen Durchschnitten haben wir herausgefunden, dass Züge an 67,46% der Haltestellen "pünktlich" kommen. Zu Spät kommen dagegen 28,43% und 4,11% kommen zu früh. Bei den abfahrenden Zügen fahren 33,68% nicht pünktlich von den Haltestellen ab. 3,60% fahren zu früh und 62,72.% pünktlich von ihrer Haltestelle ab. Wobei pünktlich im Sinne der Deutschen Bahn berechnet wird. Also mit 6 Minuten Toleranz. Würde man pünktlich mit ± 0 Minuten rechnen, wären 62,2% tatsächlich pünktlich. Zu früh kommen Züge im Durchschnitt nur bei 4,61%. Zu spät kommen 33,19% der Züge. Bei den Abfahrenden Zügen dagegen kommen 4,08% zu früh. 39,00% der Züge kommen zu spät und 56,91% der Züge kommen pünktlich. zu sehen ist das nochmal in Abbildung 4.6 Bei der Analyse der Standzeiten wurde allerdings herausgefunden, dass 84,33% der Züge Genau so lange in den Haltestellen stehen, wie es geplant war. Allerdings ist nicht gesagt, dass die Züge pünktlich sind. Gesamt wurden 6274 Haltestellen im Zeitraum vom 03.12.2017 bis zum 01.05.18 beobachtet und analysiert.

```
1 SELECT zuege.arzeitist , zuege.arzeitsoll , zuege.dpzeitist ,
   zuege.dpzeitsoll , `zuege`.`ID` FROM `zuege` WHERE `zuege`.`evanr` = {} AND `zuege`.`ID` > {} ORDER BY `zuege`.`ID` ASC
```

Quellcode 4.5: SQL Query für neue Halte einer Haltestelle

4.6.2 Durchschnitt der Zeiten einer Strecke

Um die Durchschnitte der Fahrten zu berechnen, wird zu Beginn jeder Strecke einmal analysiert (siehe 4.5.2). Die daraus resultierenden Daten werden dann Bahnhof für Bahnhof summiert und danach, wie für einen Durchschnitt nötig, mit der Anzahl der Summanden geteilt. Problematisch sind vor allem Züge die frühzeitig ihre Fahrt beenden, da diese dann

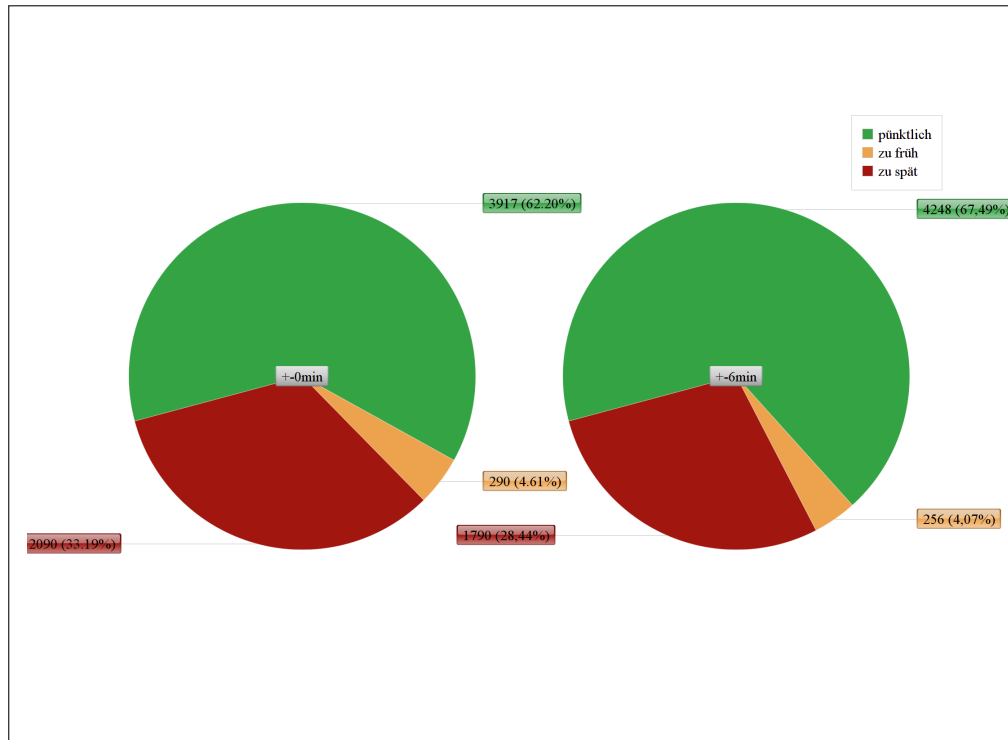


Abbildung 4.6: Prozentuale Anteile der Durchschnittlichen Verspätung bei ankommenden Zügen

Name der Zelle	Datentyp
ID	INT
EVA_nummer	INT
DescriptionID	INT
Average	double
Count	INT
lastValue	INT

Tabelle 4.1: Struktur der Durchschnitts Tabelle

natürlich einige Bahnhöfe nicht anfahren. Das kann dazu führen, dass bei der Berechnung der Durchschnitte Fehler passieren, die das Ergebnis verfälschen.

4.7 Visualisierung

Hier eventuell über die beiden libraries was schreiben oder einfach in den jeweiligen chapter bisschen erklären

Visualisierung mit

- matplotlib pyplot (Python)
- d3.js (Web basiert)

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

Neuronale Netze benötigen zum trainieren der Neuronen des Netzes viele Trainingsdatensätze. Diese Datensätze werden als lokale Dateien auf dem Dateisystem der Computer, welche das Netz trainieren, zwischengespeichert. Dies ist notwendig, um die Performance beim Trainieren zu erhöhen, da beim Trainieren sehr viele Aufrufe auf die Datensätze über eine Schnittstelle durchgeführt werden. Diese Aufrufe würden bei einzelnen Datenbankabfragen deutlich länger dauern und ein schnelles Anlernen des Netzes verhindern. Alleine die Latenzzeiten eines Aufrufs übersteigt die Dauer einer lokalen Ladezeit um den Faktor x.

Hier wert und verweis latenzzeiten mit quelle

Daher wird für die Datenverarbeitung ein Skript programmiert, welches die Daten vorverarbeitet und auf dem lokalen Dateisystem ablegt. Dabei werden auch bereits die drei verschiedenen Modi (Training, Test, Vorhersage) beachtet. Dies ist wichtig, da in den Test Datensätzen keine Trainingsdatensätze vorkommen sollten. Denn dann kann es passieren, dass die Neuronen gezielt nur diese Datensätze beachten und das Training nur auf die gegebenen Testdatensätze ausgerichtet ist. Diese Faktoren müssen bei der automatischen Datenverarbeitung beachtet werden.

In diesem gesamten Chapter/Kapitel fehlen noch Quellenangaben und weitere Literaturhinweise

5.2 Vorverarbeitung der Datensätze

Kurze Einführung schreiben

formatierung einrückung d. absätze

Bei der automatischen Vorverarbeitung werden die Datensätze aus der Datenbank in einzelne .csv-Dateien geschrieben. Hierbei werden je nach Modus Trainings-, Test-, Vorhersagedatensätze in einer anderen Struktur generiert. Bei einem Vorhersagedatensatz werden die unbekannten Spalten mit None aufgefüllt. Bei Trainings- und Testdatensätzen werden die abgefragten Datenbankdatensätze in zwei Gruppen aufgeteilt, da diese sich nicht überschneiden sollen. In Abbildung 5.1 ist die Verzeichnisstruktur zu sehen, diese wird bei der Vorverarbeitung automatisch angelegt. Des Weiteren ist es notwendig, die Datentypen in Zahlen zu konvertieren, da Tensorflow nur mit numerischen Typen sinnvoll umgehen kann. Die allgemeine Lösung führt zu einem Encoding der Strings als Integer oder Float Datenwert. Diese Konvertierung wird anhand von sogenannten Vocabfiles vorgenommen. Diese beinhalten alle möglichen Strings der Datensätze mit je einem String pro Zeile. Die Zeilennummer wird dabei von Tensorflow automatisch als Wert für den String verwendet. In Abbildung 5.2 ist ein Ausschnitt des Vocabfile für die Gleisbelegung zu sehen.

```

1 def timetotimeint(input):
2     input = str(input)
3     if input == 'None':
4         input = "24:00:00"
5     hhmmss = input
6     (h, m, s) = hhmmss.split(':')
7     result = int(h) * 60 + int(m)
8     result = math.floor(result/5)
9     return result
10
11 def openvocalfile(name):
12     lines = []
13     filename = "./vocabfiles/" + str(name) + ".txt"
14     with open(filename, mode="w+", encoding="utf-8") as file:
15         for line in file:
16             line = line.rstrip('\n')
17             lines.append(line)
18     return lines
19
20 def writevocalfile(name, vocab):
21     lines = vocab
22     filename = "./vocabfiles/" + str(name) + ".txt"
23     with open(filename, mode="w+", encoding="utf-8") as file:
24         for item in lines:
25             if item == "":
26                 print("No item found, dont save")
27             else:
28                 file.write("%s\n" % item)
29

```

Quellcode 5.1: Ausschnitt aus der Datei generate_csv.py

In der Abbildung 5.1 ist ein Ausschnitt aus der Datei generate_csv.py zu sehen. In dem Ausschnitt befinden sich die wichtigsten Funktionen zur Vorverarbeitung der Datensätze und zur Erstellung der einzelnen Vocabfiles.

generate csv noch genauer beschreiben und eventuell ablaufdiagramm oder so

Hier die einzelnen Spalten noch definieren, beschreiben

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.

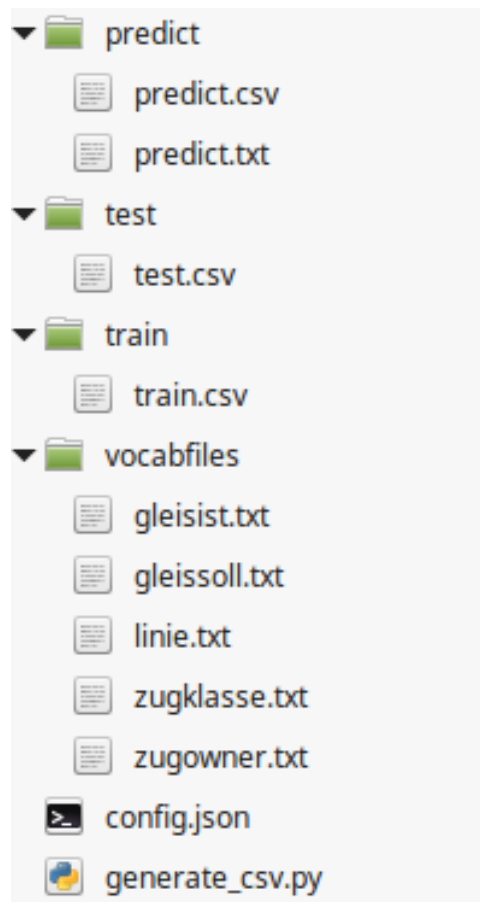


Abbildung 5.1: Der Verzeichnisbaum mit der Aufteilung der Datensätze

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.

1	4
2	7
3	1
4	9
5	3
6	2
7	6
8	23
9	21
10	22
11	19
12	18
13	14
14	13
15	5
16	8
17	2a/b

Abbildung 5.2: Ausschnitt aus einem Vocabfile, hier zu sehen Gleis (Soll)

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.

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

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.

Stimmt diese Tabelle noch?

5.3 Einrichten der Tensorflow Umgebung

Was wird alles für Tensorflow benötigt

Eventuell how to install tf verlinken

Bevor ein neuronales Netz mit Tensorflow realisiert werden kann, 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 Paket 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 in etwa verzehnfacht.

5.4 Begriffsdefinitionen für ein neuronales Netz

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 und genauere Definitionen können in den Quellen nachgelesen werden.

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 dritten Layer.

Loss ist ein FLoat-Wert, welcher durch eine Funktion bestimmt wird, welche die falschen Vorhersagen gewichtet, je geringer der Loss-Wert, desto geringer die Lerndauer.

Accuracy ist die Genauigkeit der Vorhersagen, wenn zum Beispiel von 100 Vorhersagen 40 korrekt sind, beträgt die Genauigkeit 0,4 oder 40 Prozent.

Optimizer ist die Funktion, welche für die Einstellungen der Neuronen beim Training verwendet wird. Mit dieser Funktion kann auch die Lernrate angepasst werden. Wenn diese zu hoch eingestellt ist, beginnt die Loss Funktion stark zu schwingen.

Estimator ist die Grundlage für das Modell und beinhaltet alle Einstellungen der Parametern. Hier wird auch die Anzahl an Layern, Neuronen und der Optimizer angegeben.

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.

Activation Function ist eine Funktion, welche am Ende jedes Layer angewendet wird, um die Neuronen in dem Layer zu aktivieren. Dies kann mit einem Join in der parallelen Programmierung verglichen werden. Je nachdem welche Aktivierungsfunktion gewählt wird kann es zu verschiedenen Gewichtungen der Neuronen kommen.

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 sind die mathematische Abbildung eines Vektorraumes. Als Vereinfachung kann in diesem Modell die Tensoren als Matrix oder Vektor angesehen werden, welche je nach Parameter sich in ihrer Dimension unterscheiden.

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.

5.5 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 einem späteren Zeitpunkt von der Eingabefunktion echte Werte eingesetzt werden. Deshalb befinden sich in einem Modell des neuronalen Netzes auch niemals echte Datensätze sondern nur die Parameter, die durch das Trainieren erstellt wurden.

```

1 def input_fn_mode(mode):
2     filenames = ""
3     if mode == "train":
4         filenames = glob(os.path.join('./train', '*.csv'))
5     elif mode == "test":
6         filenames = glob(os.path.join('./test', '*.csv'))
7     else:
8         filenames = glob(os.path.join('./predict', '*.csv'))
9     # get all filenames for datasets in this mode, shuffle
       them
10    random.shuffle(filenames)
11    # select one file
12    filename = filenames[0]
13    # Extract lines from input files using the Dataset API.
14    dataset = tf.data.TextLineDataset(filename)
15    if mode == "predict":
16        dataset = dataset.map(parse_csv, num_parallel_calls=5)
17        # do only one prediction
18        dataset = dataset.batch(1)
19    else:
20        shuffle = True
21        num_epochs = 4000
22        batch_size = 1
23        # shuffle if wanted
24        if shuffle:
25            dataset = dataset.shuffle(buffer_size=500000)
26
27        dataset = dataset.map(parse_csv, num_parallel_calls=5)
28        # repeat for epoch count
29        dataset = dataset.repeat(num_epochs)
30        # generate batches of datasets
31        dataset = dataset.batch(batch_size)
32
33    return dataset

```

Quellcode 5.2: Ausschnitt von der Input Funktion aus der Datei train_test_predict.py

Input FN beschreiben siehe 5.2

5.6 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

5.7 Anlernen des Netzes

Beim Anlernen eines neuronalen Netzes sind sich viele der Quellen einig.

ein paar quellen querverweise/belege hier einfügen

Je mehr Daten vorhanden zum Anlernen, desto genauer das daraus entstehenden Modell und die erzielten Resultate.

Für dieses Projekt wurde das selbe Verhalten erwartet?

In diesem Fall sollte dies sich ebenfalls so verhalten. Da jedoch bei der Einarbeitung in Tensorflow und dessen Verwendung sehr viel Zeit geflossen ist, kann diese These nicht belegt werden. Aus Zeitmangel beim Trainieren des neuronalen Netzes auf eigener Hardware muss auf ein Großteil der Datensätze aus Zeitgründen vorerst verzichtet werden. Vorab gilt es die Genauigkeit für eine kleinere Testregion zu testen und verifizieren. Da diese Region nicht die Situationen in ganz Deutschland widerspiegeln kann ist im Vorhinein klar. Nichtsdestotrotz soll eine Vorhersage in kleinem Rahmen ermöglicht werden. Die Weiterentwicklung des neuronalen Netzes muss in die Zeit nach dem Abschluss verschoben werden oder von einer Gruppe Studenten aus dem nachfolgenden Jahr übernommen werden, da hierfür schlicht und ergreifend die Ressourcen zu knapp sind. Vorhersagen sollen trotzdem ermöglicht werden, auch wenn diese eventuell mit dem jetzigen Modell nicht genau sind.

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.

5.8 Verifizieren des Netzes

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

Ein grundlegendes Problem bei neuronalen Netzen ist die Ungewissheit, ob die angenommenen Parameter des Netzes überhaupt zu einem passenden Ergebnis führen können. Als Beispiel dient die Klassifizierung in 1441 Klassen. Hierbei wurde bei den Tests eine sehr niedrige Genauigkeit festgestellt. Dies hat jedoch nicht direkt etwas zu bedeuten, denn es könnte sein, dass durch die hohe Anzahl an Klassen die Lerndauer der Neuronen ansteigt. Dies ist im Vergleich zu den vereinfachten 24 Klassen deutlich zu sehen, denn dort kann das neuronale Netz innerhalb einer Stunde bereits über 90% Genauigkeit erreichen, sofern die passenden Parameter gewählt wurden. Dieses Problem ist vor allem kritisch, wenn noch keine Erfahrungen gemacht wurden, welche Parameter Entscheidend für eine korrekte Vorhersage sind, da die Lerndauer bei 288 Klassen auf weit über 6 Stunden ansteigt und das zuvor erprobte Netz nur noch eine Genauigkeit von 35-45% erreicht hat. Alles in allem lässt sich aus diesen Tests die Erkenntnis gewinnen, dass ein neuronales Netz bei mehr Klassen nicht nur länger braucht um genauere Ergebnisse zu liefern, sondern auch deutlich mehr Trainings- und Testdatensätze benötigt.

5.9 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 übergeben. Dort wird die Vorhersage durchgeführt und liefert einen Tensor als Antwort zurück. In diesem Falle besitzt der Tensor jeweils 24, 288 oder 1441 Klassen, welche jeweils eine Uhrzeit oder einen Zeitintervall von einer Stunde oder fünf Minuten darstellt. Jeder Uhrzeit wird über eine Softmax Funktion eine relative Wahrscheinlichkeit zugeordnet. Dies bedeutet, dass die Summe aller Klassen gleich 100% entsprechen. Ein erwartetes Ergebnis einer Vorhersage wäre also eine Normalverteilung über einen bestimmten Wert. Das würde zum Beispiel bedeuten, dass ein Zug 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.

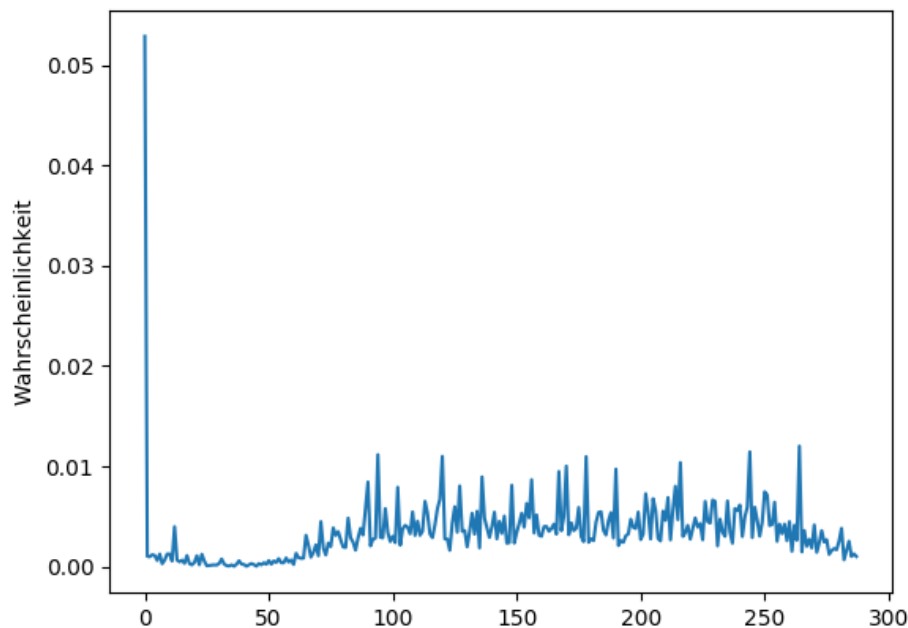


Abbildung 5.3: Verteilung der Wahrscheinlichkeiten bei 288 Klassen und allen Spalten als Eingabeparameter

Da durch die fehlende Lerndauer und das nicht optimale neuronale Netz keine hohe Genauigkeit beim Trainieren erreicht werden konnte, ist das Ergebnis bei der Vorhersage mit Tensorflow sehr ungenau. Ein Problem hierbei ist die nicht optimierte Loss Funktion, welche das Anlernen von Datumswechseln nicht berücksichtigt. Des Weiteren sollte eine vorbereitende Gewichtung der einzelnen Spalten in den Datensätzen vorgenommen werden vor der Eingabe in das neuronale Netz vorgenommen werden. In Abbildung 5.3 ist eine Vorhersage mit 288 Klassen zu sehen, jede falsche Abweichung in der Vorhersage bedeutet also direkt fünf Minuten Fehler für den Nutzer. Das direkt aus der Abbildung erkennbare Problem ist das Rauschen der Vorhersage um mehrere zeitlich weit auseinander liegenden Klassen. Das Rauschen kommt durch die mangelnde Lerndauer und die daraus resultierende Unsicherheit der Vorhersage. Als Vergleich kann Abbildung 5.4 herangezogen werden, dieses Modell nutzt jedoch nur die Ankunftszeit ohne die Information über die Zugklassen oder den betreffenden Bahnhof. Dort kann nach etwa drei Stunden Lerndauer die erwartete Vorhersage mit einer Genauigkeit von 50-75% erreicht werden. Da jedoch durch diese Vereinfachung der Sinn hinter der Vorhersage verloren geht, ist dieses Ergebnis im Endeffekt nicht nützlich.

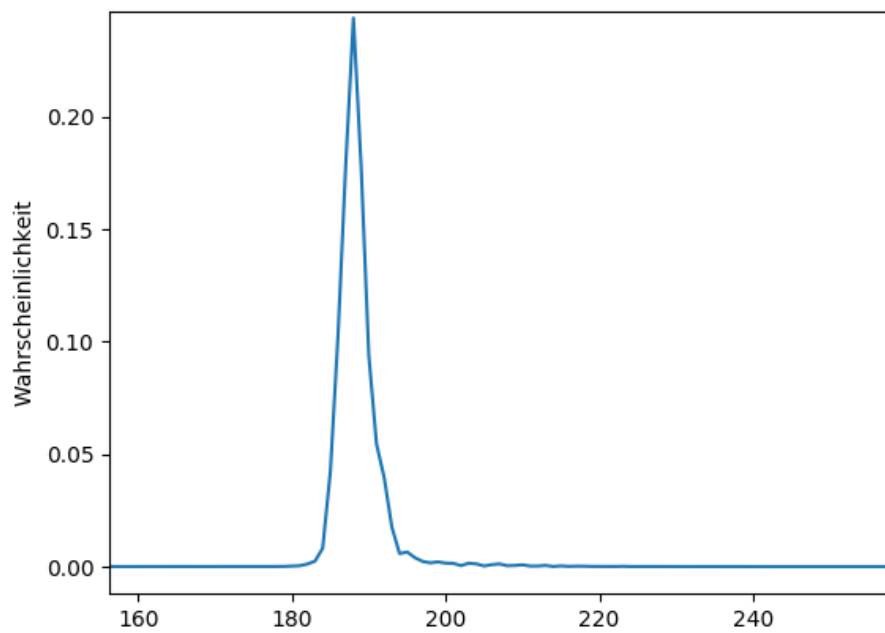


Abbildung 5.4: Verteilung der Wahrscheinlichkeiten bei 288 Klassen mit nur der Ankunftszeit als Eingabeparameter

5.10 Bewertung der Ergebnisse

Die Bewertung der erzielten Vorhersagen muss, je nach Modell und Aufbau betrachtet werden. Als Hauptmerkmal diene die Anzahl der Klassen und Neuronen, so benötigt ein Modell mit vielen Klassen deutlich mehr Neuronen und damit Lerndauer, um die gleichen Ergebnisse zu erzielen. Außerdem sind die verwendeten Eingabeparameter von hoher Bedeutung für die Lerndauer und die Genauigkeit. Je mehr Datenspalten in das Modell übergeben werden, desto länger dauert ein Einschwingen auf die gewünschte Vorhersage. Als kurzes Fazit zum Thema Tensorflow kann gesagt werden, dass das Vorhaben die Vorhersage genau zu machen leider kein Erfolg hatte. Jedoch konnte die Vorverarbeitung und eine Grundoptimierung vorgenommen werden. Dies ist in Anbetracht der Zeit und der Komplexität des Modells bereits als ein Teilerfolg zu sehen.

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

In diesem gesamten Kapitel fehlen noch Quellenangaben und weitere Literaturhinweise

Heutzutage ist die Bereitstellung einer Website eine einfache Methode Daten mit anderen Menschen zu teilen. Da die Datenbank und Website einen gewissen Sicherheitsstandard erfüllen soll, wird sich für ein Framework entschieden, welches bereits integrierte Sicherheitsfunktionen bietet. Der Name des Frameworks lautet Laravel¹. 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 6.1 kann die Struktur der einzelnen Ansichten erkannt werden.

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

in anderen Kapiteln paragraph anstatt item, formatierung schöner

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 Anbetracht der mobilen Nutzung der Website wichtig.

Toplist

Hier die Punkte der Website updaten wenn sich etwas ändert.

¹Siehe: <https://laravel.com/>

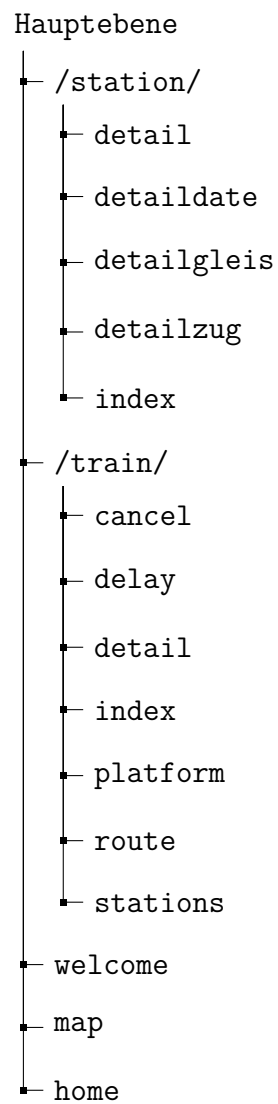


Abbildung 6.1: Struktur der einzelnen Views der Website

Map ist eine Karte, welche als Basisoberfläche Kartenmaterial von OpenStreetMaps² verwendet. Darauf werden mithilfe von Leaflet³ einzelne Schichten gezeichnet, wie zum Beispiel die Bahnhöfe der Deutschen Bahn.

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 Erstellung der Webrouuten

Mittlerweile gibt es in beinahe allen Webframeworks eine native...

In Mittlerweile fast allen Webframeworks gibt es eine native Unterstützung für Restful basierte Routen. In Laravel werden hier die Routen nochmals in vier Kategorien je nach Anwendungsfeld aufgeteilt. Diese Routen sind nach deren Zuständigkeit benannt und heißen `api`, `channels`, `console` und `web`

Namen kursiv?

. Im Normalfall reichen die Webrouuten für das Vorhaben aus, falls es eine komplette API für alle Datensätze geben soll, kann diese über die API Routen definiert werden. Der Aufbau einer Webroute ist relativ simpel, wie in Listing 6.1 zu sehen ist. Zuerst wird die Route ausgehend vom Startpunkt der Webseite angegeben. In der Route können Parameter mit geschweiften Klammern als Platzhalter dargestellt werden. So ist es möglich Routen für alle Stationen anzulegen, ohne diese einzeln programmieren zu müssen. In der Route wird dann der Parameter aus der URL genommen und anhand dessen der Inhalt der entsprechenden Seite angezeigt.

Beim Erstellen der Routen gibt es jedoch auch Fallstricke, die zu Beginn nicht direkt erkennbar sind. So muss zum Beispiel die längste Route zuerst angegeben werden, da die Routen nach dem First Match Prinzip abgearbeitet werden. Sollte unter der Hauptroute noch eine Subroute mit Parameter stehen, wird trotz Parameter nur die Hauptroute angezeigt.

Event Abbildung wie die Reihenfolge richtig und wie falsch aussieht als Listing

```
1 Route::get('/station/{id}/timetable/{date}', '
    StationController@timetable')->name('station.detaildate');
```

Quellcode 6.1: Beispiel einer Webrouutendefinition

²Siehe: <https://www.openstreetmap.org/>

³Siehe: <https://leafletjs.com/>

6.3 Erstellung der Seiten

Hier was zu Mockups und usability einbringen mit Beispielen anhand der Website

Bei der Erstellung der einzelnen Ansichten der Website wird zuvor eine grundlegende Strukturierung anhand von Mockups erstellt. Diese dienen dazu schnell Änderungen vorzunehmen und diese anschließend nach verschiedenen Faktoren wie Ordnung und verständliche Anordnungen zu bewerten. Ein Mockup der Stationsseite ist in Abbildung x.y zu sehen.

Hier Abbildung von Mockup der Stationsübersicht einfügen

Dort wird bereits zu Beginn auf die verschiedenen Subseiten geachtet. Diese sollen die Datenmenge in für die Nutzer besser verständliche kleinere Teile aufspalten und ordnen. Des Weiteren gilt es zu beachten, dass durch die Struktur von Templates in Laravel eine einheitliche Ansicht für alle Stationen gegeben ist. Nur der Inhalt der Seiten unterscheidet sich von Station zu Station. Ein weiterer Vorteil ist die Nutzung von Bootstrap. Dieses Webframework nutzt CSS und Javascript, um je nach Webbrowser und Auflösung die Website trotzdem anschaulich darzustellen. So soll die Website auf dem Smartphone ohne spezielle App genauso gut benutzbar sein, wie auf dem heimischen Computer der Nutzer. Dabei ist die Ladedauer und die Größe der ausgelieferten Webseiten bereits beachtet. Die Größe ist immernoch von Relevanz, da die Nutzer noch mit geringen Bandbreiten auf Edge oder GPRS Geschwindigkeiten unterwegs sein können. In Tabelle x.y ist ein Vergleich der Ladezeit zwischen zwei Webseiten aufgezeigt.

Tabelle mit Ladezeit pro Website und Netz anlegen und füllen, eventuell Erklärung was ist gprs und edge siehe Gespeicherten bookmark

6.3.1 Idee des dynamischen Nachladen

Hier was zum Gedanken nicht alle Statistiken direkt zu laden = langsam und viele Daten + Serverlast

Bei der Erstellung der Übersichten und Statistiken für die Züge und die Stationen, sollen immer nur die dem Nutzer sichtbaren Elemente erstellt und geladen werden. Dies spart Ressourcen auf dem Server und gleichzeitig Bandbreite beim Nutzer. Des Weiteren ist die Webseite dadurch deutlich performanter, da die Menge an Quellcode im Hintergrund besser aufgeteilt wird. Diese Unterteilung der Statistiken sorgt also nicht nur für eine bessere Übersichtlichkeit, sondern auch für einen schnelleren Seitenaufbau beim Nutzer. In Abbildung x.y wird das Laden der Subressourcen aufgezeigt. Dieses erfolgt via Javascript code, welche die nicht sichtbaren Elemente gleichzeitig im Hintergrund aus dem DOM entfernt.

Hier Abbildung einfügen mit nachladenden Subseiten

6.3.2 Die Stationsübersicht

Auf der Seite der Stationsübersicht wird dem Nutzer eine Übersicht über die vorher ausgewählte Station angezeigt. Da es viele verschiedene Statistiken gibt, wird die Navigation durch Tabs realisiert. Die folgenden Elemente sind auf der Stationsübersicht wählbar:

Die Elemente anpassen sollte sich was ändern am Namen oder Inhalt. DER NAME/INHALT STIMMT NICHT

Fahrplan Sollte Fahrplan heißen

Gleisstatistiken Sollte Gleisstatistiken werden. Hier Gleiswechsel auswerten

Stundenstatistiken Wie oft fährt im Schnitt ein Zugklasse x ab, wie oft fährt ein Zug auf Gleis x pro Stunden

Tägliche Statistiken Hier Verspätung und Gleiswechsel pro Tag

Haltestellenstatistik Hier was zur Haltestelle allgemein, wv Züge gesamt recorded und wv Ausfall in Prozent, wv Verspätungen Barschart wie damals ?

6.3.3 Die Zugübersicht

Auf der Seite der Zugübersicht werden dem Nutzer allherhand Informationen zu dem ausgewählten Zug angezeigt. Um die Informationen besser zu ordnen wird eine Navigation mit Tabs erstellt, welche die folgenden Elemente enthält:

Die Elemente anpassen sollte sich was ändern am Namen oder Inhalt

Haltestellen Hier wird dem Nutzer die Route des Zuges angezeigt. Gleichzeitig gibt es einen Querverweis auf die angefahrenen Haltestellen, um deren Statistiken anzuschauen.

Verspätung Hier wird dem Nutzer eine Statistik zur Verspätung über den Verlauf der Strecke angezeigt. So sollen etwaige Engpässe aufgedeckt und erkennbar werden.

Ausfallstatistik Hier soll dem Nutzer eine Statistik zur Ausfallwahrscheinlichkeit angezeigt werden.

Gleiswechsel Hier kann der Nutzer sehen, ob der Zug in einem Bahnhof häufiger von einem anderen Gleis als dem Sollgleis abfährt.

Streckenwechsel Hier kann der Nutzer die verschiedenen Strecken sehen, im Falle einer Umleitung in der Vergangenheit

Verlauf Hier werden dem Nutzer alte Daten angezeigt und diese ausgewertet.

6.4 Testen der Seiten mit Unit Test

Tests sind immer wichtig um bei Änderungen am Code zu merken, ob was schief läuft.

Mit einer steigenden Komplexität der Website wird das manuelle Testen immer aufwändiger. Um bestehende Seiten auf Fehler durch eine Änderung schnell zu überprüfen werden Unit Tests eingesetzt. Neben den Unit Tests werden Integration Tests durchgeführt, um ein fehlerfreies Zusammenarbeiten der Komponenten als Gesamtes sicherzustellen. Als einfachste Testart lässt sich eine Überprüfung von HTTP Status Codes realisieren. So kann geprüft werden, ob eine Route den Code 200 (OK) oder einen Fehlercode zurück gibt. Im Falle eines internen Fehlers in Laravel, wird dem Nutzer eine benutzerdefinierte Fehlerseite angezeigt. Diese wird mit dem HTTP Code 500 (Internal Server Error) an den Nutzer gesendet. In der Entwicklungsumgebung werden Debug Informationen dem Entwickler ausgegeben. Diese werden im Produktiveinsatz aus Gründen der Sicherheit nicht an die Nutzer ausgegeben. In Abbildung x.y ist eine Fehlermeldung aus Entwicklersicht zu sehen. Der Fehler wird zuvor bereits von einem Integrationstest erkannt und in einer Logdatei mit weiteren Details vermerkt. Diese Logdatei kann durch das Ausführen des Testframeworks angezeigt werden. Die Ausgabe ist in Abbildung x.y zu sehen. Die Vorteile von automatischen Tests ist die schnelle Erkennung, ob eine Änderung im Quellcode ungewollte Effekte verursacht.

Hier weiter Schreiben.....

Abbildung eines Fehlers auf der Website

Abbildung eines Testdurchlaufs Assert 5 Success 4 Error 1 oder so

6.5 Visualisierung der Datensätze

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

Absätze setzen für bessere Lesbarkeit hier und im ff

Zur Visualisierung der Datensätze werden vorberechnete und LiveDaten verwendet. Je nachdem, ob dem Nutzer eine Interaktionsmöglichkeit gegeben werden soll, wird 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 Datensätze 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.

Alle Skripts werden vor dem Hintergrund der Wiederverwendbarkeit geschrieben.

Ein wichtiger Schritt 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.

Um die Datenmenge für den Nutzer zu verringern, wird die Datenaufbreitung serverseitig durchgeführt. Der Nutzer bekommt daraufhin vorverarbeitete Datensätze, welche als simples Datenformat oder JSON in eine Grafik eingebettet werden. Die Formate unterscheiden sich abhängig von den anzuzeigenden Statistiken. Für die einfachere Wiederverwendbarkeit wird ein Controller für die Datenaufbereitung entwickelt. Dieser wird GraphController.php genannt und soll intern die Datenvorverarbeitung im Webserver übernehmen. Je nachdem wie die Python Scripts aufgebaut sind, können sie entweder vom Nutzer per WebCGI, oder durch die Backendschnittstelle des Laravel Frameworks ausgeführt werden. Ein Vorteil bei der Ausführung durch laravel ist der Cache, welche in diesem Fall problemlos mit anderen Nutzern geteilt werden kann, da keine persönlichen Nutzerdaten darin enthalten sind.

Grafik von CGI zu User mit und ohne Cache.

Da die Website komplett dynamisch generiert wird, müssen alle Querverweise durch die in Laravel vorgesehenen Routen ersetzt werden.

Diese im ersten Moment etwas ungewohnte Arbeitsweise/Programmierung/... hat den entscheidenden Vorteil, dass die Verlinkung durch das Framework vorgenommen wird und nur noch die reale Adresse, nicht die Datei abgeändert werden muss

Dies sieht im ersten Moment etwas ungewohnt aus hat aber den Vorteil, dass beim Ändern der realen Adresse die Datei nicht mehr bearbeitet werden muss, da die Verlinkung durch das Framework vorgenommen wird.

Grafik einer Verlinkung , deren dynamische ersetzung.

Für jede Haltestelle sollen den Nutzern verschiedene Statistiken zu Verfügung gestellt werden. Hierzu ist es notwendig sich Gedanken über die möglichen relevanten Themen zu machen. Eine interessante Betrachtung ist zum Beispiel die Verteilung von verschiedenen Zugklassen (ICE, RB, ...) auf die im Bahnhof vorhandenen Gleise. Am Beispiel Karlsruhe kann man erkennen, dass die Gleise 101 und 102 nicht für den Fernverkehr verwendet werden. Gleichzeitig kann die Verteilung der Zugklassen pro Gleis relativ zueinander erkannt werden. So gibt es Gleise welche hauptsächlich vom Fernverkehr bedient werden und Gleise, die häufig für S-Bahnen benutzt werden. Das sich daraus weitere Informationen gewinnen lassen, ist deutlich beim Berliner Ostbahnhof zu erkennen. Dort fahren die S-Bahnen auf den ausgebauten Gleisen, die über eine Stromschiene verfügen. Züge in der Nacht wie der NightJet verkehren dabei hauptsächlich auf den Gleisen 1 bis 3. Als besondere Herausforderung beim Programmieren der Anzeige der Statistik kann das noch relativ unbekannte Framework c3js gesehen werden. Die Datensätze aus der Datenbank müssen bevor sie an den Nutzer gesendet werden als JSON formatiert werden. Diese Aufgabe übernimmt der GraphController des Backends. Dieser liefert für die verschiedenen

Statistiken die jeweiligen Ausgaben als JSON. Eine Problematik kann die Begrenzung des PHP Memory Limits sein, da bei großen Datenabfragen dieses leicht überschritten werden kann. Weitere Probleme treten in Verbindung mit Offset Bugs auf, diese sind durch die von extern kommende Programmteile vorprogrammiert. Oftmals ist ein Index eines Gleises nicht sichtbar, da die Ausführung der Javascript Funktion einen Index zu früh aufhört und somit den letzten Datensatz verschluckt. Dieses Problem kann mithilfe von weiteren Datensätzen am Ende behoben werden, diese werden, da die Zugklasse auf NONE gesetzt ist nicht im Frontend angezeigt. Die Funktion des Backends für die Gleisbelegungsstatistik wird in Quellcode Listing x.y dargestellt. Die darin verwendete MySQL Abfrage ist belastet den Server kaum, da dieser alle Einträge der Station durch eine vorherige Query bereits zwischengespeichert hat und nur eine neue Aggregatfunktion über diesen Zwischenspeicher laufen lassen. Der auf die Abfrage folgende Quellcode sorgt für eine Formatierung der Datensätze in einem für c3js günstigen Ausgangsformat. Die generierte JSON Datei ist exemplarisch in Listing x.y abgebildet.

Um die Daten in der JSON nicht dauerhaft erneut generieren zu müssen, wird der in Laravel bereits integrierte Cache benutzt. Dieser ist sehr mächtig und verfügt über verschiedene Routinen, welche diverse Speichermethoden des Caches unterstützen. Entschieden wurde sich für einen Dateibasierten Cache ohne weitere Software. Dieser ist in der Regel ausreichend schnell und wird vom Webserver im Arbeitsspeicher gehalten. Die Ladezeiten einer Station nachdem diese im Cache vorhanden ist, fällt von über 250 Millisekunden auf unter fünf Millisekunden. Dies entspricht dem Faktor 50. Gerade bei den Daten der Stationen ist ein Cache ohne größere Probleme umsetzbar. Da der Miner jede Station maximal einmal die Stunde aufruft, werden dem Nutzer auch keine Daten längerfristig vorenthalten. Zudem verändern sich die bereits gespeicherten Datensätze nicht mehr. Ein Nachteil des Caches ist bei der Entwicklung ebenfalls nicht zu merken, da hier die Konfigurationsdatei auf geringer oder gar keine Cachenutzung global eingestellt werden kann.

Listing von der Funktion `GraphController@getTrainclassPerPlatformStatistic`

Listing von der Ausgabe JSON `GraphController@getTrainclassPerPlatformStatistic`

Hier Grafik von Gleisbelegung Karlsruhe HBF einfügen

Eine weitere interessante Statistik könnte die Verspätung eines Zuges an verschiedenen Tagen sein. So kommt ein Zug zum Beispiel chronisch zu spät oder es gibt häufig ungewollte Gleiswechsel eines Zuges. Vor allem Muster sollten am Ende von den Nutzern erkannt werden können. Die Beziehungen und Muster von Zügen untereinander ist auch bei einer Vorhersage mithilfe eines neuronalen Netzes wichtig. Diese Muster erstmals zu erkennen ist die Grundlage für eine spätere Aufbereitung der Daten für das neuronale Netz. In der Theorie müsste das neuronale Netz diese Muster selbständig erkennen und erlernen können, dies dauert aber einige Zeit und benötigt viele Datensätze, daher soll vorab eine Sortierung der Datensätze vorgenommen werden.

Da der Nutzer eine hübsche Statistik sehen will, werden verschiedenen Grafiken und Diagramme je nach Anwendungsbereich benutzt. Eine Idee für ein Diagramm wäre ein dreidimensionales Histogramm über die Zeit mit dem Streckenverlauf und der Verspätung.

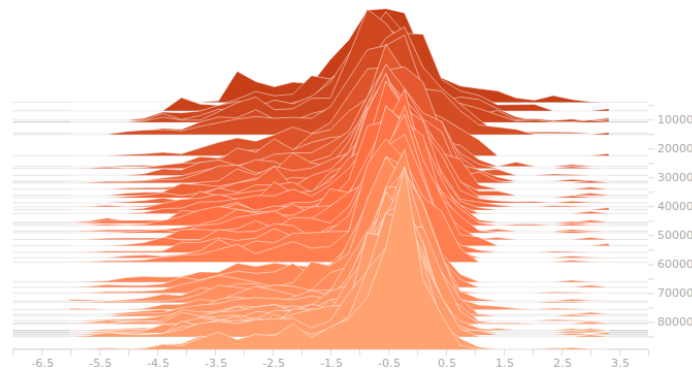


Abbildung 6.2: Dreidimensionales Histogramm aus dem TensorBoard

Ein solches Histogramm wird derzeit im Tensorflow eigenen Tensorboard verwendet, um die Verteilungen der einzelnen Schichten im Netzwerk über die Lerdauer zu visualisieren. In Abbildung 6.2 ist ein solches Histogramm zu sehen. Die dritte Dimension stellt den zeitlichen Verlauf deutlich dar.

6.6 Darstellung der Datensätze

Hier was zum View Stations schreiben, un deren details

In der Bahnhofsübersicht der Website werden den Nutzern alle Informationen und Statistiken zu diesem Bahnhof angezeigt. In der Fahrplanübersicht zu jedem Tag soll der Nutzer die Möglichkeit erhalten, zu jedem Zug Statistiken zu dessen Verspätungen über einen Zeitraum anzeigen zu lassen. Die Auswahl des Zuges findet entweder über den Fahrplan auf der Seite oder über die eigenständige Seite für die Zugsuche statt. Die Statistiken für jeden Zug werden, sofern nicht vorhanden, automatisch generiert und dann für 60 Minuten gecached. Der Einsatz eines Caches ist sinnvoll, da die Generierung der Statistiken einige Ressourcen benötigt und sich die Datensätze eines Zuges nur selten ändern. Somit wird das doppelte Senden von Anfragen an den MySQL Server verhindert. Die Routen der Website sind darauf ausgelegt möglichst kleine Teile der Website auszutauschen oder dynamisch nachzuladen. Dies soll vor allem bei dem mobilen Nutzen der Website die Ladezeiten gering halten und den Server entlasten.

Grafik Fahrplan bzw. Suche Zug mit zugnummerfull

Erstes Laden der Seite, Weiteres laden der seite (wie tcp syn,act,etc. Diagramm)

Kapitel 7

Schlussfolgerung

7.1 Rückblick

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

Rückblickend haben wir einiges geschafft was wir schaffen wollten. Leider aber nicht alles. Das ist vor allem dem Unterschätzen der Arbeit geschuldet, die notwendig war, um dieses Thema zu bearbeiten. Natürlich gab es auch Probleme innerhalb der Gruppe. So war die Kommunikation zwischen den Mitgliedern meist spärlich oder wenig hilfreich. Ein wichtiger Schritt waren vor allem die Datenerfassung aus der API der Deutschen Bahn. Mithilfe der Daten aus dem Miner konnten die Daten um einiges einfacher analysiert werden als die Daten die wir als Response der DBAPI erhalten haben. Müssten wir eine weitere Studienarbeit schreiben, würden wir bereits vor der Suche nach einem Dozenten für das Thema eine Gliederung für das Thema erstellen. Dadurch ist es für die Studenten und den Dozenten einfacher einen ordentlichen Umfang der Studienarbeit festzulegen. Auch sollte festgelegt werden welche Themen im Vordergrund stehen sollen.

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).

Insgesamt können wir sagen, dass „Analyse und Auswertung von Echtzeit-Fahrplänen der Deutschen Bahn“ ein faszinierendes Thema für eine Studienarbeit war. Problematisch war, dass wir keine klare Definition des Problems festgelegt hatten. Zu Beginn war uns nicht klar was wir innerhalb unserer Studienarbeit bearbeiten wollten. Dies wurde erst nach und nach im 6. Semester deutlich. Dadurch wurde es schwierig, die Parameter in Bezug zu setzen und zeitnah auch praktisch umzusetzen. Eine weitere Schwierigkeit war die schlechte Sichtbarkeit unseres Erfolgs. Trotz der ganzen Arbeit die wir betrieben haben, sind unsere Ergebnisse nur schwer sichtbar. Die visualisierbaren Anteile sind in Form von Diagrammen und Abbildungen in die Studienarbeit mit eingeflossen, schlecht darstellbar bleibt der Aufwand für die Abfrage und Vorverarbeitung der Daten. Die

Berechnungen eines Durchschnitts zum Beispiel sind simpel, problematisch war jedoch die Beschaffung der dazu benötigten Daten. So ist es zum Beispiel notwendig, für Berechnung der Durchschnittlichen Verspätung nicht nur die Informationen zur aktuellen Haltestelle abzufragen, sondern auch die der vorherigen Haltestelle. Werden diese in zu vielen Abfragen abgerufen, kann das zu Performanceproblemen führen.

Formulierung und Bezüge unklar evtl rewrite

Für Datenbeschaffung von der Deutschen Bahn API, muss die API in regelmäßigen Abständen abgefragt werden. Die dadurch erhaltenen Daten müssen nun überprüft und in die Datenbank geschrieben werden. Andere Arbeitsschritte wie die Auswertung der Daten oder die Prognose der Ankunftszeit sind komplexere Unterfangen gewesen.

7.3 Ausblick

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

Als ungeklärt kann man die Prognose der Ankunftszeit betrachten. Diese benötigt noch sehr viel Arbeit um eine Prognose mit einer besseren Genauigkeit zu erreichen. Dafür ist aber noch sehr viel Arbeit und Rechenleistung nötig. Eine Erweiterungsmöglichkeit sehen wir in der Publizierung unserer Rohdaten. Damit Andere, die sich mit der Materie beschäftigen wollen, nicht wie wir auf der grünen Wiese starten müssen. Eine Idee wäre, die historischen Daten der Deutschen Bahn zu puffern, um Neueinsteigern einen Grundbestand an Daten zu liefern. Durch ein API könnte jeder auf diese Daten zugreifen. Die Daten könnten verändert werden, um andere Algorithmen, die auf viele Daten zurückgreifen müssen, direkt testen zu können. Dann müsste nicht monatelang auf neue Daten gewartet werden. Auch gibt es bestimmt noch Möglichkeiten der Analyse und des Dataminings, die wir nicht berücksichtigt haben. Diese könnten noch entworfen und implementiert werden. Außerdem könnte die Nutzerinteraktion auf der Homepage verbessert werden.

Weitere Arbeit an dem Model+ vorhersage, weitere visualisierungen, bessere nutzerinteraktion

Literatur

- JIAWEI HAN, Micheline Kamber [2000]. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, S. 7 [siehe S. 10].
- USAMA FAYYAD Gregory Piatetsky-Shapiro, Padhraic Smyth [1996]. „From Data Mining to Knowledge Discovery in Databases“. In: *AI Magazine Volume 17 Number 3*. Association for the Advancement of Artificial Intelligence [siehe S. 30, 31].













Liste der ToDo's

■	Dieses Abstract besser schreiben und eventuell eine englische Übersetzung anfertigen	1
■	TODO	7
■	TODO	7
■	Wieso wollen wir das machen und warum ist das für uns wichtig.	7
■	Planbarkeit soll verbessert werden, dadurch verringerte Reisedauer	7
■	Hier das Ziel aus der Anmeldung schön definieren und klar Abgrenzen was Ziel und was optional nice to have ist.	8
■	Hier etwas zum Stand der Technik schreiben, neuronale Netze, Tensorflow, KI, Datamining, OpenData, etc	10
■	Hier Zitat aus Buch Definition zu datamining und datenbeschaffung einfügen	10
■	Eine wichtige Änderung in den letzten Jahren war die Entscheidung einiger Großunternehmen, Daten über API Schnittstellen für Entwickler verfügbar/-nutzbar zu machen.	10
■	Zeitliche Einteilung, beachten 5. Semester ist weniger Zeit, Hauptteil wird im 6. Semester passieren	11
■	schwerer Fehler	11
■	Dieses gesamte Kapitel aufsplitten und in die Grundlagen der einzelnen Chapter einordnen	13
■	Data Mining Einführung und dessen Bedeutung für das Projekt	13
■	Das Data Mining ist ein essentieller Bestandteil des Projektes. Ohne genügend Daten als Grundlage kann dieses Projekt nicht funktionieren, da ein Neuronales Netz nur mithilfe von Datenmaterial trainiert werden kann. Hierbei gilt: Je mehr Datenmaterial zur Verfügung steht, desto genauer das Neuronale Netz. Um die weitere Automatisierung des Datenflusses zu ermöglichen, werden die Datensätze in einem offenen und weiterverwendbaren Format gespeichert.	13
■	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
■	Datenmodell erläutern, welche Rohdaten aus der DB-API	14
■	Schauen, ob Kapitel noch Sinn macht	15
■	Was bekommen wir eigentlich alles über die API geliefert	16
■	Hier noch besser formatieren und nochmal lesen Kapitel x.1.1 -x.1.4	16
■	In diesem gesamten Kapitel fehlen noch Quellenangaben und weitere Literaturhinweise	17

■ Durch die geringere Datenmenge (Reduzierung der Anzahl abgerufener Stationen von 6600 auf 1200), konnte die Umsetzung schnell realisiert werden. Da es sehr viele Optionen und Probleme gab, wurde die erste Version nach etwa x Wochen durch die zweite Version des Miners ersetzt. Diese besitzt neben neuen Funktionen auch die Erweiterung zur vollständigen Abfrage der API. Außerdem konnten die Probleme des Miners minimiert werden. Die zweite Version kann zudem alle Daten abfragen und nutzt deutlich mehr Informationen, die in der API der Bahn bereitgestellt werden. Die wichtigste Änderung ist die Fehlererkennung in der Abfrage der Datensätze. Hierdurch wird vermieden, dass zu viele Datensätze fehlen. 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 denen aus der API Dokumentation erhalten Datenstrukturen. So sollte zum Beispiel 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.	19
■ Anzahl Wochen	19
■ Hier noch verfeinern und grafiken anpassen	19
■ Hier Quellcode updaten und anzeigen, beschreiben	20
Abbildung: Es fehlen Abbildungen von elementaren Abläufen	21
Abbildung: Es fehlt eine tabelle zum Vergleichen des Funktionsumfangs der Versionen	21
■ Hier noch was bedeutet 503 und eventuell zitat aus RFC https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html	22
■ 64 Gigabyte statt 16 Gigabyte	22
■ Wie werden die Datensätze abgespeichert und verwaltet? Das Schema der Datenbank befindet sich im Ressourcen Ordner	26
■ Hier etwas darüber erläutern	26
■ Literatur verweise einfügen	26
■ x.y	26
■ Diese Aufgabe war schwieriger als erwartet, da Im- und Export mehrere Stunden Zeit in Anspruch nehmen und nicht exportierte Einträge des Miners während des Umzuges mit dem neuen Server synchronisiert werden müssen. Dies ist bei einer Datenbanktabelle, welche dauerhaft mehrere Transaktionen des Miners erfährt, sehr mühsam umzusetzen (warum? vielleicht noch erläuterung). Um den Prozess so schonend wie möglich zu gestalten, wurde ein Skript geschrieben, das nach der fertigen Migration der Datenbank die Tabellen miteinander synchronisiert. Ein MySQL Sharding mit Master- und Slave-Modus war aufgrund inkompatibler Versionen nicht möglich. Nach der Synchronisation der Tabellen durch das Skript wurde der alte Miner gestoppt und der Miner auf dem neuen Server gestartet. Die Downtime des Miners betrug nur circa 60 Sekunden. Ein Fehler in der Installation...	26

■ Datenbanken sind toll, aber es muss bei einer kritischen Stelle ein Backup vorhanden sein.	28
■ watt?	28
■ Listing mit splitfile.php	29
■ zitat leer	30
■ Hier noch Text	30
■ Strecken eines Zuges werden in langen Zeichenketten statt EVA-Nummern abgelegt	31
■ Dieses Kapitel bezieht sich auf den Tabellen Cache Vorgang, also Generierung einer neuen besser selectbaren Tabelle, das Quellcode Listing ist noch nicht korrekt	31
■ da scheint es ein Problem mit der Referenzierung zu geben	32
■ hier und im ff Indizes? Spalten und Spalten?	35
■ Beispiel ausführen	38
■ eventuell noch Absätze einfügen um Lesbarkeit zu verbessern, neue Absätze mit „//“ um Einrückungen zu vermeiden	40
■ Zusammenhang der Verspätungen näher ausführen	43
■ eventuell noch Absätze einfügen um Lesbarkeit zu verbessern hier und im ff . .	43
■ Hier eventuell über die beiden Libraries was schreiben oder einfach in den jeweiligen Chapter bisschen erklären	46
■ Wie kommen die Datensätze aus der DB zum neuronalen Netzwerk, wie wird die Formatierung vorgenommen	47
■ Hier Wert und Verweis Latenzzeiten mit Quelle	47
■ In diesem gesamten Chapter/Kapitel fehlen noch Quellenangaben und weitere Literaturhinweise	47
■ Kurze Einführung schreiben	48
■ Formatierung Einrückung d. Absätze	48
■ Generate CSV noch genauer beschreiben und eventuell Ablaufdiagramm oder so .	49
■ Hier die einzelnen Spalten noch definieren, beschreiben	49
■ Stimmt diese Tabelle noch?	52
■ Was wird alles für Tensorflow benötigt	52
■ Eventuell how to install TF verlinken	52
■ Verweis einfügen	53
■ Quelle mit weiterführenden Definitionen angeben	53
■ Input Funktion beschreiben	54
■ Input FN beschreiben siehe 5.2	55
■ Erläuterung welche Informationen in das Neuronale Netz eingegeben werden und welche Daten von dem Netz ausgegeben werden.	56
■ Achtung eventuell doppelter Eintrag siehe spätere Kapitel.	56
■ ein paar Quellen querverweise/belege hier einfügen	56
■ Für dieses Projekt wurde das selbe Verhalten erwartet?	56
■ 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	56

■ Aufzeigen wie sich die Menge an Daten auf die Genauigkeit auswirkt	56
■ Welche Optionen und Parameter können optimiert werden, wie ändert sich dadurch das Ergebnis.	56
■ Hier Tabellen mit Vergleich der Methoden und Genauigkeit, Geschwindigkeit, Erläuterungen weshalb das Ergebnis so ist.	57
■ Testen des neuronalen Netzes, Verifikation der Genauigkeit und deren Steigerung durch Training oder Anpassungen des Netzes	57
■ Vorhersagen aus Daten treffen und anschauen wie gut sie sind, wo gibt es Probleme, welche Probleme treten auf.	57
■ Wie wird die Website bereitgestellt, was kann sie und welche Views existieren für die Nutzer	61
■ In diesem gesamten Kapitel fehlen noch Quellenangaben und weitere Literaturhinweise	61
■ in anderen Kapiteln paragraph anstatt item, formatierung schöner	61
■ Hier die Punkte der Website updaten wenn sich etwas ändert.	61
■ Mittlerweile gibt es in beinahe allen Webframeworks eine native...	63
■ Namen kursiv?	63
■ Event Abbildung wie die Reihenfolge richtig und wie falsch aussieht als Listing	63
■ Hier was zu Mockups und usability einbringen mit Beispielen anhand der Website	64
■ Hier abbildung von mockup der Stationsübersicht einfügen	64
■ Tabelle mit Ladezeit pro Website und Netz anlegen und füllen, eventuell erklärung was ist gprs und edge siehe Gespeicherten bookmark	64
■ Hier was zum gedanken nicht alle statistiken direkt zu laden = langsam und viele daten + serverlast	64
■ Hier Abbildung einfügen mit nachladenden Subseiten	64
■ Die Elemente anpassen sollte sich was ändern am Namen oder Inhalt. DER NAME/INHALT STIMMT NICHT	65
■ Die Elemente anpassen sollte sich was ändern am Namen oder Inhalt	65
■ Tests sind immer wichtig um bei änderungen am code zu merken, ob was schief läuft.	66
■ Hier weiter Schreiben.....	66
■ Abbildung eines Fehlers auf der Website	66
■ Abbildung eines Testdurchlaufs Asser 5 Success 4 Error 1 oder so	66
■ Dimensionen der Daten, ORT; ZEIT; NAME; STRECKE; etc.	66
■ Absätze setzen für bessere Lesbarkeit hier und im ff	66
■ Verlinken auf mathplot und d3js	66
■ Alle Skripts werdenvor dem Hintergrund der Wiederverwendbarkeit geschrieben.	66
■ Verlinkung auf Explain bzw. erklären mit Screenshot	67
■ Grafik von CGI zu User mit und ohne Cache.	67
■ Diese im ersten Moment etwas ungewohnte Arbeitsweise/Programmierung/... hat den entscheidenden Vorteil, dass die Verlinkung durch das Framework vorgenommen wird und nur noch die reale Adresse, nicht die Datei abgeändert werden muss	67

	Grafik einer Verlinkung , deren dynamische ersetzung.	67
	Listing von der Funktion GraphController@getTrainclassPerPlatformStatistic .	68
	Listing von der Ausgabe JSON GraphController@getTrainclassPerPlatformStatistic	68
	Hier Grafik von Gleisbelegung Karlsruhe HBF einfügen	68
	Hier was zum View Stations schreiben, un deren details	69
	Grafik Fahrplan bzw. Suche Zug mit zugnummerfull	69
	Erstes Laden der Seite, Weiteres laden der seite (wie tcp syn,act,etc. Diagramm)	69
	Was ist geschehen, was würden wir anders machen, was waren wichtige Schritte	70
	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.	70
	Formulierung und Bezüge unklar evtl rewrite	71
	Wie geht es weiter, könnte es weiter gehen, was sollte verbessert werden, wo befinden sich Schwachstellen, event. ungelöste Probleme	71
	Weitere Arbeit an dem Model+ vorhersage, weitere visualisierungen, bessere nutzerinteraktion	71