



Hochschule **RheinMain**  
Fachbereich Design Informatik Medien  
Studiengang Medieninformatik

## Abschlussarbeit

zur Erlangung des akademischen Grades

Master of Science

---

# Data Mining komplexer Datenstrukturen aus PDF-Dokumenten

---

Vorgelegt von	Deniz Aydar
am	25. Juli 2022
Referent	Prof. Dr. Dirk Krechel
Korreferent	Prof. Dr. Philipp Schaible

## Erklärung gemäß ABPO

Ich erkläre hiermit, dass ich

- die vorliegende Abschlussarbeit selbstständig angefertigt,
- keine anderen als die angegebenen Quellen benutzt,
- die wörtlich oder dem Inhalt nach aus fremden Arbeiten entnommenen Stellen, bildlichen Darstellungen und dergleichen als solche genau kenntlich gemacht und
- keine unerlaubte fremde Hilfe in Anspruch genommen habe.

Wiesbaden, 25. Juli 2022

---

Deniz Aydar

## Erklärung zur Verwendung der Masterthesis

Hiermit erkläre ich mein Einverständnis mit den im folgenden aufgeführten Verbreitungsformen dieser Abschlussarbeit:

Verbreitungsform	Ja	Nein
Einstellung der Arbeit in die Hochschulbibliothek mit Datenträger	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Einstellung der Arbeit in die Hochschulbibliothek ohne Datenträger	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Veröffentlichung des Titels der Arbeit im Internet	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Veröffentlichung der Arbeit im Internet	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Wiesbaden, 25. Juli 2022

---

Deniz Aydar

## **Zusammenfassung**

PDF-Dokumente besitzen viele Informationen, aus denen sich neue Daten generieren lassen. Doch das Extrahieren von solchen Daten ist heutzutage immer noch mit Hürden verbunden. Dies gilt auch für die Dokumente, die für die einzelnen Arbeitsprozesse von Autohäusern und deren Kfz-Werkstätten verwendet und anschließend gelagert werden. Um die Frage zu beantworten, inwiefern neue Daten aus dieser Art von Dokumenten verarbeitet werden können, wird im Rahmen dieser Arbeit ein System konzipiert und entwickelt, welches es ermöglichen soll, weiteres Wissen zu gestalten und zu generieren. Dieses System will dabei zwei verschiedene Ansätze aus der künstlichen Intelligenz nutzen, um einen automatisierten Lösungsweg zu kreieren, und eine grafische Schnittstelle für die Möglichkeit der menschlichen Nutzung anbieten. Der erste Ansatz handelt um die Nutzung der symbolischen künstlichen Intelligenz, während demgegenüber der zweite Ansatz die modernen Techniken des maschinellen Lernens für das Lösen dieser Problematik nutzt.

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>3</b>
1.1	Zielsetzung . . . . .	4
1.2	Ablauf . . . . .	6
<b>2</b>	<b>Hintergrund</b>	<b>9</b>
2.1	Kontext . . . . .	9
2.2	Analyse . . . . .	10
2.2.1	Extraktionsbibliotheken . . . . .	11
2.2.2	Dokumentarten . . . . .	13
2.2.3	Frameworks . . . . .	18
<b>3</b>	<b>Konzeption und Umsetzung</b>	<b>21</b>
3.1	Aufbau . . . . .	21
3.2	Symbolische KI . . . . .	23
3.2.1	Extraktionsvorgang . . . . .	24
3.2.2	Webserver . . . . .	26
3.2.3	Syntax und Grammatik der Fakten und Regeln . . . . .	26
3.2.4	Regelmaschine . . . . .	30
3.3	Entwicklungsumgebung . . . . .	32
3.3.1	Editor . . . . .	34
3.3.2	Dokumentenanzeige . . . . .	35
3.3.3	Output . . . . .	36
3.4	Machine Learning . . . . .	36

---

3.4.1 Pipeline . . . . .	37
3.4.2 Webserver . . . . .	41
<b>4 Ergebnisse</b>	<b>42</b>
4.1 Testing . . . . .	43
<b>5 Konklusion und Ausblick</b>	<b>45</b>

# Kapitel 1

## Einleitung

„Digitalisierung im Alltag voranbringen“ – Das war einer der Wahlslogans während der Bundestagswahl 2021. Gemeint war damit die Forderung nach einer zunehmenden Digitalisierung im privaten Alltag vieler Bürger:innen, aber auch in der Wirtschaft machte sich wachsend der Wunsch nach mehr digitalen Alternativen breit (vgl. Bundesregierung 2021). Dieser Wunsch beinhaltete vor allem einen Wechsel von gängigen Papierformen verschiedenster Dokumente hin zu denselben in digitaler Ausprägung.

Genau jenes Bedürfnis nach Digitalisierung betrifft auch die vielen Arbeitnehmer:innen und Händler:innen in Autohäusern und deren Kfz-Werkstätten, die durch ihre beruflichen Tätigkeiten mit einer Vielzahl an Unterlagen, Dokumenten oder Belegen arbeiten müssen. Unter dieser Vielzahl fallen Dokumente wie Werkstatt-, Kauf- und Mietverträge sowie Rechnungen oder Diagnoseberichte.

Jene Beschäftigte in einigen Autohäusern und deren Kfz-Werkstätten sind Kund:innen bei ilexius GmbH. Ilexius bietet Enterprise-Resource-Planning (ERP) Systeme an, mit Hilfe derer Arbeitsaufträge und Arbeitsschritte innerhalb des täglichen Arbeitsprozesses in Autohäusern und deren Werkstätten durch Digitalisierung vereinfacht werden. In Kooperation von ilexius GmbH werde ich daher die Problematik meiner Thesis, die ich im folgenden noch genauer beschreiben werde, lösen und in die Tat umsetzen.

Durch eine Digitalisierung jener Dokumente eröffnen sich Vorteile wie bessere Zugänglichkeit, größere Langlebigkeit und vor allem eine angepasste und leichtere Nutzung, die zu einer höheren Effektivität innerhalb täglicher Arbeitsschritte führt. Durch bisherige erste Schritte der Digitalisierung sind diese benötigten Dokumente bereits elektronisch aufgearbeitet und den Beschäftigten in Autohäusern und deren Kfz-Werkstätten zur Verfügung gestellt worden.

Die Folge dessen ist, dass alle Dokumente standardisiert sind und dadurch die in den Dokumenten beinhalteten Informationen neu verarbeitet werden müssen. Eine Extraktion der Daten der einzelnen Dokumente ist jedoch nur eingeschränkt möglich, da die Struktur im gängigen Gebrauch im Dateiformat Portable Document Format (PDF) festgelegt ist.

Jene Limitierung der Datenextraktion wirkt sich gleichermaßen auf die Dealer-Management-Systeme (DMS), mit solchen die Mitarbeiter:innen ihre Prozesse abwickeln, aus. Der daraus resultierende Arbeitsaufwand, welcher dabei entsteht, um die Daten wiederverwendbar zu konstruieren, ist derzeit enorm.

## 1.1 Zielsetzung

Aus diesem Grund möchte ich innerhalb dieser Master-Thesis ein System entwickeln, das genau jene Problematik erleichtert und löst.

Innerhalb üblicher Methoden werden heutzutage die Daten zunächst über eine grafische Anzeige mithilfe einer bereits existierenden Benutzeranwendung via gängiges Kopieren und Einfügen entnommen. Dieser Schritt funktioniert zwar im Regelfall, ist jedoch oftmals stark fehleranfällig und kann lückenhaft sein.

In so einem Fall muss das Einfügen und Kopieren manuell durchgeführt werden, was bei einer enormen Menge an Dokumente zu einer monotonen Arbeit für die Händler:innen der Autohäuser führt. Andernfalls können die Daten eigenhändig abgeschrieben werden – jener Vorgang benötigt jedoch viele Ressourcen. Eher geeignet ist es, einen Datenkonverter für die PDF Dateien zu nutzen, um so die Dokumente beispielsweise in Bilder umzuwandeln, wodurch die Daten zugänglicher sind, aber immer noch verarbeitet werden müssen. Die letztgenannte Möglichkeit wirkt allerdings eher wie eine Übergangslösung als eine professionelle endgültige Durchführung.

Aufgrund dieser bisherigen teils aufwendigen und mit Fehlern verbundenen Möglichkeiten möchte ich mich in meiner Thesis von diesen Optionen abwenden und das Data-Mining nutzen. Das Data-Mining soll eine Automatisierung unterstützen, mit der Daten aus dem PDF-Dokument extrahiert werden. Allein über das Data-Mining ist es möglich, Inhalte zu extrahieren und für andere Systeme bereit zu stellen.

Der Unterschied zwischen meiner Verwendung des Data-Minings und zum gängigen Data-Mining liegt dabei in den Einschränkungen des Dateiformats: der Quelltext einer solchen Datei stellt erstens keine klare Hierarchie der Daten dar und zweitens besitzt es durch das Fehlen von Markierungen keine

Informationen darüber, was die Daten an sich darstellen sollen [28].

Mit Hilfe dieser Technik des Data-Minings soll es ermöglicht werden, aktuelle Prozesse detaillierter zu steuern und zu überwachen. Des Weiteren können unter anderem Abgleiche von Rechnungen mit dem System durchgeführt werden und weitere Datenanalysen und Reporting kreiert werden. Mit dem derzeitigen Stand der Datenextraktion können jedoch lediglich Agierende aus dem technischen Bereich arbeiten, da allein sie das dafür technisch notwendige Know-how besitzen. Die Technologien der Datenextraktion benutzen unterschiedliche Ansätze und müssen daher zunächst für diesen Fall ausgewählt werden.

Damit jedoch auch für Agierende innerhalb des technischen Bereichs die Nutzung nicht zu abstrakt bleibt, soll durch die Entwicklung meines Systems der Zugriff greifbarer gestaltet werden, welches wiederum allein durch eine benutzungsfreundliche Anwendung ermöglicht wird. Die Benutzer:innen sollen eine Möglichkeit bekommen und über eine Eingabe der Steuerung bestimmen können, welche Informationen extrahiert werden sollen.

Darüber hinaus muss eine Automatisierung vorhanden sein, um auch eine große Datenmenge verarbeiten zu können; zudem sollen die Vorgänge eine niedrigere Fehlerquote aufzeigen. Die Prozesse müssen außerdem während des gesamten Ablaufs über eine Steuerung und Regelung kontrollierbar sein. Solche Prozesse können zum Beispiel über Regeln festgelegt beziehungsweise gesteuert werden, die hingegen bei einer Erfüllung weitere Aktionen oder Regeln auslösen können.

Diese im vorherigen genannten Absatz beschriebene Eigenschaft kann durch eine symbolische künstlichen Intelligenz (KI) abgedeckt werden, welche eine vorgegebene Verarbeitung zulässt. Hiermit bezeichnet man einen unzeitgemäßen Ansatz der KI, bei der über das Festlegen von Symbolen menschliches Wissen in einer Logik gehalten wird und diese benutzt wird, um weiteres Wissen sodann zu generieren [27]. Durch eine Parametrisierung einer solchen KI kann dann die Unschärfe der ausgewählten Daten festgelegt werden, damit eine Feineinstellung erfolgen kann. Das heißt, die Benutzer:innen eines solchen Systems sollen sich beginnend von groben Definitionen für die Verarbeitung zu einer detaillierten und angepassteren Definition über Feedback des Systems hinarbeiten.

Mit Hilfe dieser Annäherung an Definitionen und Regeln kann so für eine bessere Erfolgsquote gesorgt werden. Zudem kann der Ansatz des Machine Learnings (ML), mit jenem erfolgreiche Verarbeitungen einem System antrainiert werden, weitere Dokumente aus Daten extrahieren. Hierbei werde ich im Laufe der Entwicklung feststellen, welcher Typ des Machine Learnings sich für diesen Fall am optimalsten eignet.



Beide Ansätze – der der symbolischen KI und der des Machine Learnings – bieten als Option die Entwicklung einer grafischen interaktiven Entwicklungsumgebung (IDE) an. Die IDE soll das Festlegen von Definitionen und Regeln erlauben, mit welchen die Dokumente verarbeitet werden können. Außerdem soll diese Oberfläche verschiedene Funktionalitäten offerieren und auch Feedback sowohl bei einem problemlosen Lauf, als auch bei einem fehlerhaften Lauf, zurückgeben.

Des Weiteren soll es das System ermöglichen, Änderungen der Definitionen anzumerken und Unterschiede zu erstellen. Mit dieser Funktion sollen auch bisherige Ergebnisse angezeigt werden. Die Festlegung von Definitionen und Regeln soll außerdem durch eine Ansicht der Dokumente unterstützt werden. Insgesamt wird das System die zwei Ansätze als Subsysteme aufteilen, um die Umsetzung zu modularisieren und einen Vergleich zu ermöglichen.

## 1.2 Ablauf

Um genau dieses optimale System für das beschriebene Szenario entwickeln zu können, werde ich in dieser Arbeit wie folgt vorgehen:

Für eine vollständige Auseinandersetzung soll eine weitere Analyse stattfinden, damit die vorhandenen Grundlagen für den Anwendungsfall, um den es in der Arbeit geht, evaluiert werden können. Das heißt, es werden mit Hilfe der Funktionalitäten der Bibliotheken Ergebnisse auf Basis der Datensätze erzeugt, um die Erfolgsquoten zu überprüfen. Diese werden dann verglichen und ausgewertet. Parallel dazu soll die Umgebung für die Entwicklung eingerichtet werden, mit der die Implementierung der Systeme stattfindet.

Danach widme ich mich der Konzeption und der Entwicklung des Systems. Diese beginnt mit der symbolischen KI als ein Subsystem des gesamten Projekts, durch selbige ein direkterer Ansatz ermöglicht wird. Die Entwicklung hierbei wird im Wasserfall-Ansatz, einem Ansatz bei der phasenweise die Eigenschaften einer solchen KI durchgeführt werden, um eine Grundlage für die nächste Komponente zur Verfügung zu stellen.

Bei dieser Komponente handelt es sich um die interaktive Entwicklungsumgebung, die den Zugang für die Benutzer:innen zum System darstellt. Deswegen folgt im nächsten Schritt die Konzeptionierung und Entwicklung der interaktiven Entwicklungsumgebung, welche sich während der Entwicklung der symbolischen KI parallelisieren lässt. Sobald das Gerüst der Benutzungsoberfläche fertiggestellt ist, möchte ich dies mit der symbolischen KI verbinden.

Als Gegenstück wird sich dann mit dem Machine Learning Ansatz und dem dazugehörigen Subsystem, bei der gleichermaßen eine Konzeptionierung und Entwicklung stattfindet, gewidmet. Hier wird das bisherige Gesamtkonstrukt in einem iterativen Ansatz umgesetzt, sodass der Hauptteil so früh wie möglich verfügbar ist, damit das Training des ML-Systems durch die Dokumente auch zeitnah starten kann. Auf das Auswerten dieses Subsystems folgt eine Anpassung dessen. Die Komponenten der symbolischen KI und des Machine Learnings erlauben damit einen Vergleich zwischen zwei verschiedenen Ansätze aus der Künstlichen Intelligenz.

Im Anschluss daran wird das Software Testing für die beiden Subsysteme eingeführt. Mit Unit Testing, Integration Testing und Functional Testing wird das erfolgreiche Ausführen gewisser Eigenschaften abgedeckt. Dies dient wiederum als Grundlage dafür, die Continuous Integration (CI) für das System einzuführen, mit dieser eine sichere und konsistente Entwicklung nach dem Prototyping angeboten wird.

Wenn hierauf die Systeme gegeben sind, kann das grafische Tool weiter angepasst werden, welches sich hiernach mit den Systemen verbinden soll, um infolgedessen die Ausführbarkeit der Funktionalität überprüfen zu können. Um ein gemeinsames System zu ermöglichen, welches dann gegebenenfalls weitere Testschritte bedarf, werden anschließend die Komponenten zusammengeführt. Sobald jenes eine erfolgreiche Form annimmt, wird der Teil der Continuous Integration mit einbezogen, mit jener das System einen Zustand der Instandhaltung annehmen kann. Über Testing sollen so mögliche Fehlerquellen entdeckt werden, bevor diese überhaupt auftreten können.

Unter die Zielsetzung für meine beiden eben genannten Ansätze fallen die Kundendaten mit den Eigenschaften wie Name, Firma, Adresse und Kundennummer. Zu Eigenschaften der Fahrzeugdaten zählen Angaben der Fahrgestellnummer, Modell, Farbe, Motor, Erstzulassung, Letzter/nächster Service und TÜV sowie die Abrechnungsdaten wie die Jobs mit den dazugehörigen Arbeitspositionen, der Reparaturteile, den Preisen etc.

Dazu stehen als weitere Grundlage mehr als 10.000 annotierte unterschiedliche Dokumente mit genau den zu extrahierenden Daten als Datenbank zur Verfügung. In Zukunft sollen für das Mengengerüst mehr als 100.000 Dokumente pro Jahr bearbeitet werden. Hierbei ist es aber auch möglich, aus den bereits bestehenden Daten einen Pool für den Lernprozess beziehungsweise für die Verifikation zu bilden.

Zum Schluss werde ich die Ergebnisse des gesamten Systems betrachten und einen Ausblick zu der Thematik geben.

Aus Datenschutzgründen werde ich in meiner Arbeit persönliche Informationen in den gezeigten Dokumenten anonymisieren und verpixeln.

# Kapitel 2

## Hintergrund

Um die Problematik detaillierter darzustellen und um eine Übersicht zu ermöglichen, wird in diesem Kapitel der Rahmen der Problematik aus technischer Sicht genauer dargestellt. Hierbei werden die möglichen Technologien abgewogen und evaluiert. Zunächst jedoch wird der Kontext ausführlicher erklärt.

### 2.1 Kontext

Im Automobilbereich wird der Verkauf von Kraftfahrzeugen in zwei Segmente aufgeteilt: dem Sales-Bereich, bei dem es sich um den Verkauf von Neu- und Gebrauchtwagen handelt und dem After-Sales-Bereich, bei jenem eine Bindung zum Kunden über den Verkauf von Verschleiß- und Ersatzteilen geschaffen wird. [54] Diese Aufteilung der Segmente wird ebenfalls von den Autohäusern angewendet, zugleich werden für diese Aufteilung jeweils gängige Abläufe als Prozesse benutzt.

Für den After-Sales-Bereich bietet ilexiu GmbH zwei Arten von ERP-Systemen an. Die erste Art ist das von der Automobilmarke Jaguar Land Rover finanzierte *vTab*, welches eine Plattform für elektronische Fahrzeugkontrollen und ein Backend für Jaguar Land Rover Mitarbeiter:innen, mit dem die Autohäuser bewertet werden können, anbietet. Die zweite Art mit dem Namen *ATT* ist eine Plattform, mit der Autohäuser ihre Arbeitsprozesse in Schrift, Sprache und Bild dokumentieren. Die Dokumentation erfolgt mittels mobiler Endgeräte und wird an das ERP-System weitergeleitet, sodass die Dokumentation archiviert werden kann.

Die ERP-Systeme, die bereits im Kapitel 1.1 beschrieben worden sind, unterstützen damit die Kunden:innen von ilexiu GmbH in den Autohäusern und

deren Werkstätten bei der Planung, der Steuerung sowie der Verwaltung von verschiedenen Aufgaben in ebenjenem After-Sales-Bereich.

Eine Art der Aufgaben sind Arbeitsaufträge, bei der es sich in den meisten Fällen um die Handhabung und Abfertigung eines Autos inklusive Kundendaten handelt. Der Auftrag besteht hierbei aus einer Liste an Arbeitsabläufen, die als Jobs bezeichnet werden. Die Jobs wiederum unterteilen sich in einzelne Arbeitsschritte, die Joblines genannt werden, welche die konkreten Verrichtungen der Arbeit darstellen und somit den gesamten Arbeitsauftrag in einzelne Teilschritte vervollständigen.

Sind diese Aufträge abgeschlossen, benutzen die Mitarbeiter:innen der Autohäuser ihre Dealer-Management-Systeme, um eine Rechnung zu erstellen und diese im ERP-System zu archivieren. Das ERP bietet des Weiteren eine Schnittstelle für diese Dokumente, wie z.B die Rechnungen oder die Garantieforderungen an, die allein über Scanner die Dokumente übertragen. Dort werden die Dokumente verarbeitet und mit Hilfe eines OCR Scanners digitalisiert.

Derzeit ist allerdings die Digitalisierung nicht in der Lage, einen Kontext beziehungsweise eine Semantik für diese Dokumente zu erstellen. Idealerweise kann dies genutzt werden, um aus den Dokumenten Arbeitsaufträge zu generieren und somit zu digitalisieren. Demzufolge ergibt sich hier an der Stelle die Möglichkeit, einen neuen Ansatz zu testen, infolgedessen die Informationen aus diesen Dokumenten als Daten erhaltbar sind.

## 2.2 Analyse

Das Extrahieren von Daten ist eine gängige Methode, um Informationen aus verschiedenen Systemen wie Datenbanken oder Software as a Service (SaaS) Plattformen zu erhalten. Durch diese Beschaffung können die Daten weiter verarbeitet werden, wodurch neue Lösungsmöglichkeiten für das eigene System bereitzustellen. Die Arten des Extrahierens unterscheiden sich hierbei im Zeitverlauf und der Herangehensweise. Sie können extrahiert werden, sobald Änderungen der Daten beobachtet oder mitgeteilt worden sind. Durch diese Herangehensweise werden so die Daten schrittweise herangeschafft [50].

Des Weiteren gibt es die Option, eine komplette Extraktion durchzuführen. Dieser Ansatz kann sich in vielen Fällen als nachteilig erweisen, da die Daten jedes Mal bei Anpassungen der ursprünglichen Information neu verarbeitet werden müssen. Aufgrund des Bezugs der Problematik im Rahmen der Thesis auf Dokumente, die bereits im ERP-System archiviert wurden, trifft dieser Fall hier nicht zu.

### 2.2.1 Extraktionsbibliotheken

Für das Data Mining von PDF Dateien existieren bereits Lösungen, die von großen Unternehmen wie Amazon Web Services (AWS) oder Adobe angeboten werden. Zum Beispiel nutzen Produkte wie Textractor [22] von AWS oder Adobes Extractor API [26] künstliche Intelligenz, Machine Learning und Optical Character Recognition (OCR), um die Extraktion der Daten zu ermöglichen. Unternehmen wie ABBYY [21] haben sich bereits auf diesem Gebiet spezialisiert und sind erfolgreich im Umgang mit der Datenextraktion. Auch existieren bereits Open-Source Bibliotheken, die sich mit dem Entnehmen der Daten aus PDF Dokumenten auseinandersetzen.

Die meisten der Open-Source Bibliotheken lassen sich hinsichtlich der Aufteilung der Aufgaben gruppieren. So erlauben unter anderem PDFMiner [55] oder Apache Tika [23], welches vor der Extraktion der Texte und Metadaten noch die Klassifizierung und Identifizierung des Medientyps von Dateien anhand einer umfangreichen Hierarchie besitzt, Texte aus PDF Dokumenten zu gewinnen [12]. Funktionalitäten für das Extrahieren von Daten aus Tabellen bieten Projekte wie Tabula [52] oder Camelot [25] an. Diese Projekten decken ferner einen weiteren Aspekt durch die Nutzung der Funktionalitäten ab.

Bei den beiden zuletzt genannten Bibliotheken weisen sich niedrige Fehlerquoten [24] und die Nutzung von einer Fuzzy Logik auf. Camelot ermöglicht überdies ein Web Interface, mit dessen Hilfe eine simple Extraktion mittels einen Browser lokal stattfinden kann. Weitere Bibliotheken, wie Texttricator [37] sind umfangreicher gestaltet und ergeben für die Anfragen spezifischer Inhalte geeignete Schnittstellen.

Doch auch wenn dies mit textbasierten Dokumenten funktioniert und der Export flexibel ist, können zunächst weder nicht-textbasierte noch komplexere PDF Dokumente verarbeitet werden. Des Weiteren sind die Verarbeitungen nicht automatisierbar und die Nutzer:innengruppe wird durch eine fehlende grafische Komponente eingeschränkt. Aus dieser Problematik resultierend zeigt sich, dass community-angetriebene Projekte wie PdfMiner.six [42] (das eine Abspaltung des bereits erwähnten PdfMiner ist) und PdfPlumber [49] (welches wiederum auf PdfMiner.six basiert) geeignete Alternativen darstellen.

PdfPlumber stellt im Vergleich zu seinem Ausgangsprojekt überdies eine Tabellenextraktion bereit und bietet für beide Aspekte der Extraktion eine Schnittstelle an. Es stellt eine Kommandozeile, die Möglichkeit, visuelles Debugging zu nutzen, und Funktionalitäten wie das Entfernen von Duplikaten in Texten zur Verfügung.

Das in Abbildung 2.1 zu sehende Debugging ermöglicht, ermöglicht es unter anderem, die Texte des Dokuments hervorzuheben, was wiederum die Extraktion greifbarer darlegt und die Entwicklung mit den Funktionalitäten unterstützt. Auch die Funktionalität des Entfernens von Duplikaten kann dann zum Einsatz kommen, wenn Dokumente bei ihrer Erstellung oder Verarbeitung versehentlich dieselben Einträge erneut im Quellcode des PDFs besitzen. In dieser Abbildung sind die persönlichen Daten des Auftrags anonymisiert und deshalb verpixelt.

**JAGUAR**  
 Robert-Bosch-Str. 7  
 55129 Mainz-Hechtsheim  
 Telefon: +49 (0)31 - 30 37 50  
 Telefax: +49 (0)31 - 30 37 510  
 E-Mail: service@robertbosch.de

**INTERN-AUFTRAG**

Rechnung an: 10000

AW / Teile-Nr.	Bezeichnung	Menge	Preis	Betrag
Intern: Sonstiges				
Durchzuführende Arbeiten:				
keine SA offen 25.02.2022 TM				
Nach der Probefahrt war die Motorkontrollleuchte an und es hat vorne aussen nach Kühlwasser gerochen				
JAG1	Motorkontrollleuchte brennt im Kombi FC auslesen	0,20	66,00	13,20
	Zwischensumme Job : 1			13,20
JAG2	Kühlwassersystem überprüfen ggf. nachfüllen	0,50	66,00	33,00
	Zwischensumme Job : 2			33,00
Endsumme				46,20 €

Die Erteilung des Auftrages erfolgt unter ausdrücklicher Anerkennung unserer Allgemeinen Geschäftsbedingungen, die hier zur Einsichtnahme ausliegen. Auf Verlangen wird ein Exemplar ausgehändigt.

Unterschrift (Kunde): \_\_\_\_\_ Name in Druckbuchstaben: \_\_\_\_\_

(a) ohne visuellem Debugging

(b) mit visuellem Debugging

Abbildung 2.1: visuelles Debugging bei einem Dokument

Für die Auswahl einer Extraktionsbibliothek ist auch die dazugehörige Programmiersprache für das System der Datenextraktion, in der die Bibliothek umgesetzt worden ist, relevant. Darunter fallen die bereits genannten Bibliotheken, welche sich in zwei Sprachen gruppieren lassen: Projekte wie Tika, Tabula und Textricator sind mit Hilfe von Java implementiert, während die restlichen aufgezählten Bibliotheken in Python geschrieben sind [15]. Für Tabula gilt hier hingegen die Ausnahme, dass für das Projekt unterschiedliche Wrapper für weitere Programmiersprachen wie Python existieren. Hieraus ist zu beobachten, dass die Nutzung von Python viele mögliche Extraktionswerkzeuge erlaubt, die sich für die spätere Umsetzung als noch bedeutsam erweisen.

Die Hürden für eine komplett automatische Extraktion bestehen jedoch nach

wie vor. So ist nämlich nicht ersichtlich, wie das zu erwartende Verhalten der Logik der Extraktion für den generellen Fall auszusehen hat. Eine differenzierte Extraktion für weitere Eigenschaften eines Dokuments, wie zum Beispiel Zeilenumbrüche von Absätzen, die Seitennummern, die Kopf- und Fußzeilen, die Formatierung et cetera (etc.) ist daher nötig. Die Frage, die sich dabei stellt, ist, welche Eigenschaften eines Dokuments konkret zusätzlich extrahiert werden sollen. Um eine Auswahl zu treffen, müssen die Arten der Dokumente, die in dem ERP-System auftreten, präziser durchleuchtet werden.

### **2.2.2 Dokumentarten**

Bei den Dokumenten, die in das ERP-System eingescannt werden, handelt es sich in der Mehrheit um Garantie- und Werkstattaufträge und Rechnungen. Jedes Autohaus benutzt hierbei ein individuelles Muster beziehungsweise Format für die jeweilige Art des Dokuments. Dabei ist jedoch der Aufbau dieser Dokumente zu Teilen indirekt durch die von den Autohaus genutzten Dealer-Management-Systeme vorgegeben.

Ebendaher passiert es, dass es Autohäuser gibt, welche das gleiche DMS verwenden und zudem darüber hinaus auch einen detailliert angepassten Dokumentenaufbau für ihre Aufträge benutzen. Dem Umfang dieser Thesis geschuldet werde ich mich auf drei Dokumententypen von verschiedenen Autohäusern mit ihren jeweils genutzten DMS beschränken und anhand jener drei Typen mein Vorgehen exemplarisch demonstrieren. Die Abbildungen der Dokumente in diesem Abschnitt zeigen in einigen Fällen persönliche Daten und wurden deshalb, wie in der Einleitung beschrieben, anonymisiert.

In Abbildung 2.2 wird die erste Art eines Auftrags dargestellt. Bei diesem digitalen Schriftstück handelt es sich zunächst von der Form her um eine Garantierechnung des Dealer-Management-Systems Care. Das Dokument besitzt drei Bereiche, die relevante Informationen für einen Arbeitsauftrag enthalten. In der oberen Hälfte befindet sich zunächst das Adressfeld, welches die Kundeninformationen im genormten Adressformat enthält. Unter dem Adressfeld befindet sich ein weiterer Abschnitt, in dem die Fahrzeugdaten festgehalten werden.

Hierbei ist zu beobachten, dass dieser Abschnitt sich mit einer Tabelle vergleichen lässt. Die einzelnen Einträge werden mit jeweils unterschiedlich großen Abständen verteilt dargestellt, gleichzeitig erfüllen sie eine Gesamthöhe pro Spalte beziehungsweise eine Gesamtbreite pro Höhe einer Tabelle. Gleichwohl fällt auf, dass sich — im Gegensatz zur Anzahl der Zeileneinträge — die Anzahl der Spalteneinträge pro Zeile unterscheidet.



#!GRHMPF!#		10	70	DE
Jaguar Care Garantie				
Fahrzeughalter: Kunden-Nr.: 151365				
RECHNUNG GWL				
R70 700004693 000993		18.03.22	1	
		18.03.22		
F-Type		17649	GJC	47822300
		18.03.22		
AW/T-Nr.	Bezeichnung	AW/Menge	Preis	Betrag
00010	JAGUAR CARE WARTUNG			
CCC	WARTUNG DURCHFÜHREN			
102111	52.000 KM (32.000 MEILEN) ROUTINEWART	1,10	165,00	181,50
ZZZ001	PREMIUM SERVICE PAUSCHALE			40,00
KBEND	-----			
	TERMINIERT (X) JA: NEIN: _____			
	DATUM AUFTRAGSANNAHME: _____			
	DATUM REPARATURBEGINN: _____			
	DATUM REPARATUR ENDE: _____			
	DATUM FAHRZG. ABGEHOLT: _____			
FRE ESRVPLNOIL	MOTORÖL	7	25,00	175,00
JAG C2D3670	OLFILTER	1	33,92	33,92
JAG C2P2410	AIR FILTER	1	99,67	99,67
JAG C2A102010	SCHEIBENWASCHFLUESSIGK	1	7,13	7,13
JAG AJ813202	ÖLABLA SCHRAUBE	1	5,50	5,50
008325		18.03.22	9:55 10:65	1:10
Summe Arbeit: 221,50				
Summe Fremdl: 321,22				
Summe Teile :				
542,72	0,00%		EUR	542,72

Abbildung 2.2: Exemplar eines Garantiauftrags

Zudem fehlen jegliche Einrahmungen oder jegliche Trennlinien für Zeilen und Spalten. Bei den Einträgen in diesem Abschnitt geht es um die Abrechnungsnummer, das Datum der Reparatur, den Fahrzeugmodellnamen, das Kfz-Kennzeichen, den zuletzt abgelesenen Kilometerzählerstand, die Arbeitsauftragsnummer, der Fahrzeug-Identifikationsnummer beziehungsweise der Vehicle Identification Number (VIN), die/die zuständige/n Autohausmitarbeiter:in und das Datum der Erstellung des Arbeitsauftrags. Verpixelt und damit anonymisiert wurden die Felder des Kfz-Kennzeichens, der VIN und der/des zuständigen Mitarbeiter:in.

Im unteren Bereich wird eine Tabelle durch die Aufzählung der einzelnen Jobs und Joblines dargestellt. Die Tabelle ist an der Kopfzeile zu erkennen, die die

Jobs und Joblines in die Arbeitsnummer/Teilenummer, in die Bezeichnung, in die Menge des Arbeitszeitwertes, in den Preis und in den Betrag übersichtlich aufteilt. Die Teilenummern beziehen sich hierbei auf die Identifikationsnummer der Reparaturteile. Der Arbeitszeitwert ist dabei eine Einheit für die genaue Abrechnung der Arbeitszeiten, um eine Aufschlüsselung der Stunden zu ermöglichen. [20] Die Menge wird hierbei als Reparaturzeit oder als Anzahl der Elemente angegeben. Daneben existiert bei dieser Tabelle nur für die Kopfzeile eine untere Trennlinie. Auch hier sind Probleme bezüglich (bzgl.) der Form der Tabelle zu erkennen.

Während die ersten vier Zeilen noch dezent erscheinen, besitzt die fünfte Zeile im Abschnitt mit der Bezeichnung einen längeren Text, sodass der restliche Text in weiteren Zeilen in der gleichen Spalte darunter geschrieben steht. Die Abstände zwischen den Spalten insgesamt scheinen konsistenter zu sein, sind jedoch in der Abbildung 2.2 in mehreren Zeileneinträgen als Überläufe bei den Arbeitsnummern beziehungsweise den Teilenummern zu sehen. In der letzte Zeile ist darüber hinaus zu erkennen, dass die Einträge zu Teilen auch verschoben dargestellt werden, was eine eindeutige Erkennung der Positionierung noch weiter erschwert. Inhaltlich werden Jobs mit den dazugehörigen Joblines darunter angereiht dargestellt.

Dies ist durch die Codierung der Arbeitsnummern beziehungsweise der Teilenummern zu erkennen, die Hierarchie ist durch das jeweilige Dealer-Management-System vorgegeben. In der ersten Spalte können daher Codes von Jobs, Reparaturzeiten, Ersatzteilen, Fremdleistungen, Stempelzeiten und Kundenbeschwerden auftreten.

Für die Codierung ist keine standardisierte Spezifikation vorhanden, weshalb das für jedes Autohaus und seine Nutzung der DMS individuell definiert ist. So sind die Codes der Jobs als eine fünfstellige Zeichenkette aus Ziffern bestimmt, die mit den Ziffern 00 anfangen. Die Reparaturzeiten dagegen besitzen eine sechsstellige Zeichenkette aus Ziffern. Unterhalb der Tabelle werden am Ende des Dokuments schließlich die gesamten Kosten aus den Betragseinträgen summiert dargestellt. Diese Summe lässt sich wiederum in Kosten der Arbeit und der Teile aufteilen.

In Abbildung 2.3 wird eine Rechnung der Formel 1 DMS gezeigt. Auch hier sind sämtliche personenbezogenen Daten verpixelt und anonymisiert worden. Im Vergleich zur vorherigen Abbildung 2.2 ist ein ähnlicher Aufbau zu sehen. Während das Adressfeld geläufig aussieht, sind jedoch die Fahrzeug- und Kundendaten diesmal in zwei unterschiedlichen Tabellen aufgeteilt.

Die erste Tabelle befindet sich im oberen Teil der rechten Seite des Dokuments,

BMW Vertragshändler **ahg Autohandelsgesellschaft**

ahg Autohandelsgesellschaft mbH - Senefelderstr. 2 - D-73760 Ostfildern

**BMW Service**

**TESTRECHNUNG**

Kunden-Nr. : 444515-33  
Rechnungs-Nr. : 133  
Rechnungsdatum :  
Annehmer :  
Auftrag/LS-Nr. : 115025  
Auftragsdatum : 24.02.2022  
Hauptauftrag-Nr.: 112985  
Seite : 1  
Leistungsdatum : 24.02.2022

Filiale : E Ostfildern

Fahrz.-Nr.	Modell/Typ	Fahrz.-Nr./EZ/HU	km	Ein/Aus	angen./ausgef.
	VU31		177393	02321	
	320d Touring	20.09.2006/ 09/23	177584	02019	

Nummer	Bezeichnung	Menge/AW/St	E-Preis	Gesamt
0000556	Fahrzeugtest durchführen	2	10,66	21,32
1223505	Alle Glühstifte ersetzen	16	13,63	218,08
6100006	Prüfplan für Vorglühanlage abgearbeitet	1	13,63	13,63
6121528	nicht i.O. Bordnetzspannung unterstützen	1	13,63	13,63
11612246945	Bordnetzbatterie nachladen	4	4,27	17,08
11617790198	Profildichtung	4	4,27	17,08
12237786869	Glühstift	4	14,94	59,76
	<b>Zwischensumme</b>			<b>360,58</b>

0,00% MwSt von 360,58 = 0,00

Gesamt Lohn	EUR	266,66
Gesamt Teile	EUR	93,92
<b>Endsumme</b>	<b>EUR</b>	<b>360,58</b>

Zahlbar sofort!

**Hausanschrift**  
ahg Entenmann - Eine Niederlassung  
der ahg Autohandelsgesellschaft mbH  
Senefelderstraße 2  
73760 Ostfildern

**Kontakt**  
Telefon: +49 (0) 711/44983-0  
Fax: +49 (0) 711/44983-83  
E-Mail: ostfildern@ahg-entenmann.de  
www.ahg-entenmann.de

**Bankverbindung**  
Kreissparkasse Freudenstadt  
IBAN DE06 6425 1000 0000 5435 43  
BIC SOLADES1FDS


**Geschäftsführer**  
Alexander Krüner  
Thomas Linderich

**Registergericht**  
Stuttgart HRG 440 304  
Sitz: Hrb a. N.  
USt-ID-Nr.  
DE311162775

Abbildung 2.3: Exemplar einer Rechnung

wo die einzelnen Kundeninformationen aufgelistet sind. Die Tabelle für die Fahrzeuginformationen befindet sich im unteren Bereich der oberen Hälfte des Dokuments und ist aufgrund der vorhandenen Trennlinien eine klassische Tabelle. Im letzten Teil der Formel 1 Rechnung befindet sich die Liste der Job und Joblines.

Hier fallen ähnliche Problematiken, wie die unterschiedlichen Abstände zwischen den Spalten der Einträge im Vergleich zu dem Care Garantieauftrag auf. Auch existieren Überläufe der Einträge, allerdings treten diese nur in den Feldern der Bezeichnung auf. Die Gesamtkosten werden dann am Ende des Dokuments in den Teilsummen der Beträge bekannt gegeben.



# JAGUAR


Fuhrmeister - Robert-Bosch-Straße 7 - 55129 Mainz

**Fuhrmeister Exklusive Automobile GmbH & Co. KG**

Robert-Bosch-Str. 7  
55129 Mainz-Hechtsheim

Robert-Bosch-Straße 7  
55129 Mainz-Hechtsheim  
Telefon: +49 (0)6131 - 60 37 50  
Telefax: +49 (0)6131 - 60 37 559  
E-mail: service@fuhrmeister.de

**INTERN-AUFTRAG**



Rechnung an: 10000 „

Kundennummer 23552	Belegnummer 47061	Belegdatum 25.02.2022	Seitenbenutzer 1 / tmu
Anschließend Kennzeichen [redacted]	Typ E-PACE D240 AWD aut. 1590 AHK	Kfz-Nummer [redacted]	Eintragsung 11.06.2018
Motor 177 / 1999	KW / Hubraum 06.2021 / 06.2021	Fahrgestellnummer [redacted]	Tachostand *59.205 km
Telefon (privat) [redacted]	Telefon (geschäftlich) [redacted]	Telefon (mobil) [redacted]	Termin (unverbindlich) [redacted]
Auftragsdatum 25.02.2022			

AW / Teile-Nr.	Bezeichnung	Menge	Preis	*)	Betrag
Intern : Sonstiges					
<b>Durchzuführende Arbeiten:</b>					
<b>keine SA offen 25.02.2022 TM</b>					
Nach der Probefahrt war die Motorkontrollleuchte an und es hat vorne aussen nach Kühlwasser gerochen					
JAG1	Motorkontrollleuchte brennt im Kombi FC auslesen	0,20	66,00		13,20
Zwischensumme Job : 1					13,20
JAG2	Kühlwassersystem überprüfen ggf. abdrücken	0,50	66,00		33,00
Zwischensumme Job : 2					33,00

Endsumme  
**46,20 €**

Die Erteilung des Auftrages erfolgt unter ausdrücklicher Anerkennung unserer Allgemeinen Geschäftsbedingungen, die hier zur Einsichtnahme ausliegen. Auf Verlangen wird ein Exemplar ausgehändigt.

Unterschrift (Kunde): \_\_\_\_\_ Name in Druckbuchstaben: \_\_\_\_\_

Abbildung 2.4: Exemplar eines Werkstattauftrags

Bei der dritten und damit letzten Dokumentart (Abbildung 2.4) handelt es sich um einen Werkstattauftrag von dem Dealer-Management-System KFZ3000. Auch dieses hier abgebildete Dokument wurde anonymisiert. Dieser hier gezeigte Auftrag lässt sich in Bezug auf die Aufteilung der einzelnen Segmente mit dem Care-Warranty-Dokument vergleichen. Während auch das Adressfeld erneut normiert abgebildet ist, werden Kunden- und Fahrzeugdaten über Schlüssel-Wert Anordnungen in einem größeren Feld dargestellt.

Die untere Hälfte des Werkstattauftrags besitzt gleicherweise wie die bereits anderen genannten Dokumente eine tabellenähnliche Auflistung der Job und Joblines. Ebenfalls sind hier Überläufe und unterschiedliche Abstände in dem Bereich zu erkennen. Hinzu kommt, dass, im Gegensatz zu den anderen Doku-

mentarten, zwischen den Zeilen ein Art Summenstrich hinzugefügt wird, der die Anzahl der Jobs anzeigt. Außerdem wird im unteren Bereich des Dokuments lediglich die gesamte Betragssumme gezeigt.

Insgesamt lässt sich feststellen, dass alle hier vorgestellten Dokumentarten und ihre Hürden durch eine gemeinsame Problematik miteinander verbunden sind. Die Darstellung der Kunden- und Fahrzeuginformationen wird entweder in einer Tabellenform gemeinsam (siehe Abbildung 2.2) oder in zwei verschiedenen Tabellen (siehe Abbildung 2.3) abgebildet. Diese Tabellenform besitzt außerdem selten Trennlinien, welche essentiell sind, um eine derartige Tabelle als solche definieren zu können. Für die Liste der Job und Joblines tritt das gleiche Problem auf.

Die Auflistung der Job und Joblines als Tabelle wird allein durch ihre Anordnung der Einträge und deren Abstände dazwischen als solche erkenntlich, jedoch bricht sie in den prototypischen Dokumenten aber mindestens einmal die Form durch Überläufe oder Verschiebungen der Texte. Ein weiterer relevanter Aspekt, welcher übersehen werden kann, ist, dass die Aufträge mehr als eine Seite haben können.

Dies tritt für Dokumente auf, welche eine längere Auflistung der Jobs und Joblines besitzen und somit auf weiteren Seiten dargestellt werden müssen. Für die Extraktion der Liste muss daher auch überprüft werden, ob sich diese über weitere Seiten erstreckt, um sie zusammenzuführen und zu verarbeiten. Es ist festzustellen, dass verschiedene Tabellenstrukturen für unterschiedliche Dokumentarten existieren. Dieses Anliegen besteht, da bisher für Tabellen keine universelle Standardisierung definiert ist.

Damit diese Formen durch eine Logik des Extrahierens abgedeckt werden, besteht die Möglichkeit der Fallunterscheidung. Bei dieser Unterscheidung sollen die Dokumente anhand ihrer Arten konzipiert und definiert werden, sodass die Entnahme detaillierter stattfinden kann. Für andere Bereiche eines Dokuments wie dem Adressfeld werden gewöhnliche Funktionalitäten der Extraktion genutzt und zur Verfügung gestellt. Diese können dann von allen weiteren Dokumentarten geteilt und genutzt werden, sodass dadurch eine Verallgemeinerung verschiedener Dokumententypen ermöglicht wird.

### **2.2.3 Frameworks**

Damit das System für den Prozess der Extraktion als solches sich zeitnah im Rahmen einer Thesis umsetzen lässt, können Frameworks dem ganzen Zeitfenster entgegenkommen. Im Zuge dessen werden für die Teilsysteme verschiedene

Technologien betrachtet und in Erwägung gezogen. So wird zunächst die symbolische künstliche Intelligenz dem ersten Teilsystem zugeordnet und in seiner Art unterschieden. Zudem müssen die Realisierungsmöglichkeiten der symbolischen KI betrachtet und aus den verschiedenen Arten eines wissensbasierten Systems muss ein Modell ausgewählt werden.

Ein solches wissensbasiertes System ist hierbei der Oberbegriff für Programme, die mit Hilfe einer Wissensbasis Mechanismen der Schlussfolgerung nutzen, um Lösungen für Probleme im dafür gewählten Anwendungsbereich zu finden [5]. Es existieren verschiedene Modelle eines wissensbasierten Systems, die abhängig vom Anwendungsfall passend sind. Folgendermaßen gibt es den Ansatz des regelbasierten Systems beziehungsweise des Produktionssystems, bei dem die Schlussfolgerung über Regeln und Fakten, die von Benutzer:innen definiert werden, konstruiert wird.

Bekannt für die Erstellung eines solchen Systems ist das in der Programmiersprache C geschriebene Werkzeug *CLIPS*[3]. Dieses Werkzeug verfügt über eine eigene Syntax, welche von der Syntax des von der Inference Corporation entwickelten Expertensystems *ART* für die Regeln und Fakten abgeleitet wurde und damit zahlreiche Gemeinsamkeiten und Ähnlichkeiten aufweisen. [33]. Die Umsetzung wird ebenso mit dem Wrapper *clipsy* für die Programmiersprache Python zur Nutzung bereitgestellt.

Eine Alternative zu *CLIPS*, die zeitlich danach entwickelt wurde, ist das Expertensystem *NEXPERT OBJECT*, das sich bis auf die Datenbankintegration nur marginal in der Nutzung der Regel- und Faktendefinitionen unterscheidet [1] und für das keine aktive Entwicklung und Wartung existiert. Andere moderne Umsetzungen konzentrieren sich im Vergleich eher auf die als Regelinterpretierer verstandene Inferenzmaschine. Projekte wie *Pyke* [31], *Durable Rules* [48] oder *Experta* [43], dem Nachfolger von einem Expertensystem Framework namens *pyKnow*, erlauben es, eine Regelmenge zu definieren und die einzelnen Regeln mit selbstdefinierten Funktionen zu koppeln.

Für die Entwicklungsumgebung, die den Zugang zum System darstellen soll, eignen sich sowohl Nutzungsschnittstellen beziehungsweise User Interface (UI) Bibliotheken für den Desktop als auch für den Webbereich. Die Priorität liegt hierbei bei der Erstellung der einzelnen Komponenten der graphischen Benutzeroberfläche beziehungsweise der Graphical User Interface (GUI). Folglich können Projekte wie *Qt* [47] und das in C++ geschriebene *ImGui* [41] Bibliotheken bereitstellen, die auf einer nativen Ebene funktionieren und demgemäß in verschiedenen Anwendungsfällen genutzt werden. Die letztere genannte Bibliothek benutzt intern einen Zustandsautomaten, um die Render-Pipeline

darzustellen (diese erfordert mehr Erfahrung mit der Grafikprogrammierung), dass jenes verwendet werden kann [7].

Dem gegenüber stehen abstraktere im Front-End Bereich verwendete UI-Lösungen wie das Benutzen eines Web Frameworks. Bekannte Bibliotheken wie *Vue* und *React* ermöglichen eine schnelle und gängige Umsetzung einer Oberfläche, die auch die Zwecke einer Entwicklungsumgebung erfüllen können. Weitere Kandidaten wie *Svelte* [51] versuchen durch das Nutzen eines Compilers, die Arbeit von dem Browser auf die Build Zeit umzulagern.

Für das Teilsystem, welches den Ansatz des maschinellen Lernens nutzen soll, sind hier gleicherweise ML-Frameworks eine geeignete Lösung. Kandidaten sind die zurzeit konventionellen Frameworks wie *TensorFlow* [53], *pyTorch* [46] und *spaCy* [29]. Jene bieten für die Anwendungsgebiete unter anderem in der Bilderkennung, der Computer Vision und des Neuro-Linguistische Programmieren (NLP) eine automatisierte Lösung an. Weitere neue Projekte wie *Fonduer* [17] fokussieren sich mit Hilfe von schwachem überwachten Lernen (weak supervised learning) auf die Extraktion von umfangreichen formatierten Daten, worunter auch PDF Dokumente fallen. Dieser Lernansatz findet auch bei nicht vollständig beschrifteten Trainingsdaten einen Anwendungszweck [18], weswegen sich weak supervised learning für einfache Dokumente als praktisch erweisen kann.

## Kapitel 3

# Konzeption und Umsetzung

Für den Sprung zur Umsetzung bedarf es als nächstes die Konzeption des technischen Rahmens. Zunächst wird der Aufbau des Projekts vorgestellt, wie die beschriebene Problematik gelöst werden soll. Danach wird der Verlauf der Entwicklung über die Aufteilung des Systems in seine Komponenten dargestellt. Jene beschreiben im Detail ihre Eigenschaften und Zusammenhänge.

### 3.1 Aufbau

Für die Struktur des Systems ist vorerst die Untergliederung des Projekts relevant. Sie kann je nach einem ausgewählten Architekturmuster unterschiedlich gestaltet werden. Da die Teilsysteme wiederum als eigene Systeme dargestellt werden können und für die einzelnen Anwendungsfälle kommunizieren beziehungsweise interagieren müssen, ist hier der Ansatz der verteilten Systeme geeignet. Die Kategorie der verteilten Systeme ist in weitere unterschiedliche Architekturmuster gegliedert. Eines dieser Muster ist die Client-Server Architektur, die es erlaubt, über ein Netzwerk die Kommunikation bereitzustellen und die Teilsysteme und ihre Aufgaben zu fraktionieren.

Dahin gehend gibt es drei Komponenten mit unterschiedlichen Aufgaben, die in den eben beschreibenden Teilsystemen abgebildet werden: Die erste Komponente befasst sich mit der Extraktion über eine regelbasierte Logik, mit jener die Daten von mehreren Dokumenten eines Dokumententyps automatisiert erhalten werden. Die Zweite setzt die Entwicklungsumgebung um und ermöglicht dadurch auch jegliche Erstellung und Bearbeitung von Regeln über Textdateien. Die dritte und letzte Komponente setzt für die Extraktion der Dokumente den Lösungsansatz des Machine Learnings um.



Innerhalb des Architekturmusters fungieren die erste und dritte Komponente als Server, die die Logik der Extraktion ausführen sollen. Die Entwicklungsumgebung als zweite Komponente ist im Gegensatz dazu als Client zu betrachten, der den Benutzer:innen Änderungen in Bezug auf Regeldefinitionen erlaubt. Damit dieser Netzwerkansatz für die Architektur umgesetzt werden kann, müssen die Komponenten Schnittstellen auf der Netzwerkebene bereitstellen.

Für die jeweiligen Komponenten, die sich mit der Extraktionslogik beschäftigen, dient hier der Lösungsansatz der Benutzung eines Webserver, welcher die Anfragen verarbeitet und die verschiedenen Zustände als Antwort den Anfragen zurückgibt. Bei der Komponente der Entwicklungsumgebung eignet sich der Ansatz einer Webanwendung, mit welcher weitere Technologien aus dem Webbereich genutzt werden können. Somit wird jegliche Bearbeitung der Definitionen durch die IDE als Anfrage verschickt und die daraus folgenden Antworten als Ergebnisse präsentiert.

Für die praktische Umsetzung oder für die Entwicklung wird der Lösungsentwurf als ein Softwareprojekt in Form eines Monoliths konstruiert, welches das System als ganzes darstellt. Das Projekt trägt den Namen *Smart Document Extractor* und wird mit dem gängigen Versionsverwaltungssystem Git aufgesetzt. Die darin gegliederten Komponenten werden *Rule System*, *IDE* und *ML* genannt.

Demgemäß wird das Projekt als lokales Repository und als Remote Repository auf dem GitLab Server des Unternehmens festgehalten. Im Wurzelverzeichnis wird zudem mit der Container Virtualisierung von Docker eine betriebssystem-agnostische Umgebung für die Teilsysteme definiert. Dies ist nicht nur der Softwareentwicklung, sondern auch der Continuous Integration und deren Testschritte dienlich und ermöglicht einen Ausbau der Skalierung der Teilsysteme.

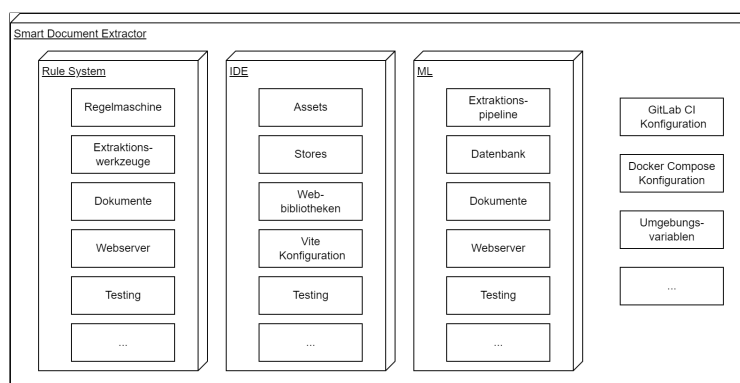


Abbildung 3.1: Struktur des Softwareprojekts

Die einzelnen Teilsysteme werden mit Hilfe einer Docker Compose Datei anschließend orchestriert, sodass die Container gemeinsam gestartet werden und ihre Kommunikation über ein gemeinsames Netzwerk geführt werden kann. Die zusätzliche Konfiguration von Umgebungsvariablen in dem Projekt erlaubt es außerdem, sensible Daten, wie die Anmeldeinformationen der Zugänge der Teilsysteme oder weitere Dienste für den Arbeitsbereich dynamisch anzupassen. Die Konfiguration wird hierbei von dem Versionsverwaltungssystem ignoriert.

Abbildung 3.1 zeigt eine Übersicht des Projekts und seine Eigenschaften, auf die im weiteren Verlauf dieser Arbeit detaillierter eingegangen wird. Dort sind neben den beschriebenen Konfigurationsdateien die Teilsysteme und deren Module als Komponenten abgebildet.

## 3.2 Symbolische KI

Für das Teilsystem der symbolischen künstlichen Intelligenz habe ich den Ansatz beziehungsweise die Nutzung eines Wissensbasierten Systems ausgewählt. Dieser Ansatz ermöglicht die Steuerung der Extraktion über definierte Fakten und Regeln. Wie in Kapitel 2.2.3 beschrieben, gibt es viele Frameworks, die ein solches System bereitstellen können. Doch zeigt sich auch bei jenen, dass die Nutzung eines dieser Frameworks nicht für den Fall geeignet ist, da mehr Funktionalitäten als benötigt abgedeckt werden.

Ein regelbasiertes System zeichnet sich durch eine Anzahl von Regeln (sogenannte Produktionsregeln), Fakten, welche sich während des Inferenzvorgangs verändern, und durch den Regelinterpreter aus. Der Interpreter kann mit Hilfe dieser Menge die Schlussfolgerungen in Form einer Vorwärtsverkettung oder Rückwärtsverkettung durchführen [11].

Für den Kontext der Dokumentenextraktion bedarf es allein die Ausführung von Extraktionen und die Generierung neuer Fakten, sobald die benutzerdefinierten Regeln mit den bereits vorhandenen Fakten erfüllt worden sind, als Funktionalitäten.

Die in dieser Arbeit vorgestellten Bibliotheken gehen mit einem größeren Aufwand der Implementierung einher, da jene Bibliotheken weitere Softwareabhängigkeiten zur Folge haben. Daher wird hier der Lösungsweg über ein eigens implementiertes Regelsystem gewählt. Der Regelinterpreter wird speziell für den Zweck der Datenextraktion entwickelt und passt sich damit an die Nutzung des gesamten Projekts an. Für die Implementierung dieses Teilsystems wird sich zudem für die Benutzung der Programmiersprache Python entschieden.

### 3.2.1 Extraktionsvorgang

Bevor auf die Regelmanchine genauer eingegangen werden kann, muss zunächst die konkrete Extraktion beleuchtet werden. Im Abschnitt 2.2.2 wurde gezeigt, dass die Hauptproblematik der Dokumente in der Extraktion der Tabellen liegt. Diesbezüglich definieren wir Tabellen, die für jede Zeile und Spalte Trennlinien besitzen, als konventionelle Tabellen. Ein Beispiel hierfür ist die Tabelle, die in Abbildung 2.3 die Fahrzeugdaten beinhaltet. Tabellen, die keine oder nur zu Teilen Trennlinien für die Zeilen und Spalten nutzen, werden als unkonventionelle Tabellen definiert. Diese Art der Tabelle tritt in den drei vorgestellten Dokumentarten in Form der Auflistung von den Jobs und Joblines auf. Andere Tabellen, die eine beliebige Form und somit die restlichen Tabellenarten darstellen, werden als benutzerdefinierte Tabelle bestimmt.

Textfelder können im Vergleich zu den Tabellen leichter extrahiert werden, weswegen die Funktionalität für alle Dokumentarten beziehungsweise Dokumententypen benutzt werden kann. Für eine Verallgemeinerung der Extraktionsfunktionalitäten wird daher eine gemeinsame Klasse namens *DocumentExtractionType* definiert. Diese gemeinsame Klasse fasst als Obermenge die Methoden, die für alle Dokumentenarten als Unterklasse genutzt werden können, zusammen. Sie bietet die Möglichkeit an durch weitere Klassendefinitionen, die den Mechanismus der Vererbung nutzen, Abbildungen von unterschiedlichen Dokumententypen zu kreieren. So werden daraus folgend die Dokumentarten, die im Kapitel 2.2.2 vorgestellt wurden, durch die Klassen *CareWarranty*, *Formel1Invoice* und *KFZ300Invoice* chronologisch dargestellt.

Die gemeinsamen Funktionalitäten, die in der Oberklasse definiert sind, bestehen aus der Textextraktion von Textfeldern, der Tabellenextraktion von konventionellen Tabellen, der Textextraktion von unkonventionellen Tabellen und der Textextraktion von benutzerdefinierten Tabellen. Damit diese Funktionalitäten ausgeführt werden können, müssen Eigenschaften der Extraktionsbereiche als Parameter angegeben werden. Für alle beschriebenen Ziele existiert mindestens die Positionierung als Eigenschaft. Sie kann als Bounding Box (bbox) mit den X und Y Koordinaten für die jeweiligen Extraktionsziele aufgeführt werden.

Eine weitere Eigenschaft, die sich aus der Positionierung schließen lässt, ist die Dokumentenseite. Bei den selektierten Dokumenten treten jeweils mehrere Seiten auf, wenn die Liste der Jobs und Joblines so viele Einträge besitzt, dass diese nicht auf eine Seite passen. Aus diesem Grund ist die Seitennummer, die den Funktionen neben den Bounding Boxen mitgegeben werden, relevant für die Beschreibung des Zielbereiches. Diese Seitennummern können den Funktionen neben den Bounding Boxen mitgegeben werden. Daraus folgend werden diese

zwei Eigenschaften — die der Positionierung und die der Seitennummer — als notwendige Bedingung für jegliche Extraktionsbereiche behandelt und damit für die Logik der Methoden durch die Parametereingabe als gegeben gesehen.

Für die Umsetzung der einzelnen Funktionen wird die vorgestellte Bibliothek PdfPlumber benutzt, um für die jeweiligen Extraktionsziele mit verschiedenen Methoden die Daten zu erhalten. Die Methoden nutzen dann die Angaben der Seitennummer und der Bounding Box, sodass der Bereich der Extraktion zunächst limitiert werden kann. Mit dem ausgewählten Bereich werden dann die Funktionalitäten für die Text- oder Tabellenextraktion aufgerufen.

Bei den prototypischen Dokumenten fällt vorerst die hinderliche Tabellenextraktion von unkonventionellen Tabellen und indirekt von benutzerdefinierten Tabellen auf: Die Bibliotheken können diese Tabellen mit den dafür konzipierten Funktionen nicht erkennen und dadurch von diesen keine strukturellen Daten, die die Tabellen abbilden sollen, erhalten. Um dies zu umgehen, werden als Lösungswege die unkonventionellen und die benutzerdefinierten Tabellen zunächst als reine Texte extrahiert und danach die Einträge der jeweiligen Felder mit Hilfe von regulären Ausdrücken gefiltert.

Auch wenn der Bereich durch die Bounding Box umrahmt ist, kann der Bereich nicht relevante Texte besitzen, die die Verarbeitung erschweren. Demzufolge werden weitere Eigenschaften wie die Angabe eines Start- und Endankers benötigt, um lediglich die Zeilen aus dem Rohtext zu erhalten. Da bei den elektronisch verarbeiteten Dokumenten die Tabellen als Texte mit Zeilenumbrüchen verfasst sind, können die Zeilenumbrüche genutzt werden, sodass aus den Zeilen eine Liste generiert wird. Die Liste wiederum wird mit regulären Ausdrücken weiterverarbeitet. In Folge dieser Verarbeitung erhalten die Daten durch Zuweisungen eine strukturierte Form und stellen folglich die Semantik der Tabellen dar.

Da auch Dokumente mit mehreren Seiten existieren, bedarf es neben den Start- und Endanker eine Menge von Zwischenankern, die als Brücke für die längeren Listen dienen. Für die unkonventionellen und benutzerdefinierten Tabellen werden daher neben der Positionierung und der Seitennummern die Start- und Endanker als die Voraussetzung für jene Extraktionsfunktionalitäten benutzt. Optional bleiben die Angabe der Zwischenanker.

Die Klassen der einzelnen Dokumentarten sind für weitere Funktionalitäten, die spezifischere Zwecke für den jeweiligen Dokumenttyp erfüllen und eine weitere Semantik für die extrahierten Texte und Tabellen bilden. Hier sind weitere Eigenschaften wie die Reihenfolge der Jobs für die Zuweisung der Werte relevant. Diese Methoden erhalten mit Hilfe von regulären Ausdrücken die einzelnen Werte für die gewünschten Felder und weisen jene mit ähnlichen

Algorithmen den Feldern zu. So werden die Daten je nach Dokumentenart detaillierter erfasst und als komplexe Objekte verarbeitet.

### **3.2.2 Webserver**

In Kapitel 3.1 wurde die Struktur des gesamten Systems definiert und die Client-Server Architektur festgelegt. Damit die Komponente der symbolischen künstlichen Intelligenz die Kommunikation über Netzwerke erlaubt, bedarf es die Implementierung eines Webserver. Insgesamt gibt es viele Ansätze, wie ein solcher Webserver aufgesetzt werden kann. Da aber die benötigten Funktionalitäten für die prototypische Entwicklung sich auf die Verarbeitung von gewöhnliche Hypertext Transfer Protocol (HTTP) Anfragen und Antworten beschränken, reicht die Nutzung eines Web Server Gateway Interface (WSGI) Servers aus.

Für diesen Zweck eignet sich das Mikro-Framework Flask, welches eine zeitnahe Umsetzung und Integration in das System ermöglicht [9]. Über die Umgebungsvariablen erhält daraufhin die Flask Instanz ihre Daten bezüglich der Konfiguration für das Hosting. Damit diese Webserver Anwendung auch das Testing und die Continuous Integration unterstützt, wird hier das von der Dokumentation vorgegebene Muster für die Application Factory angewendet.

Bei Verfolgung dieses eben genannten Ansatzes wird eine Factory Funktion geschrieben, die neben der Kreierung einer Instanz die weitere Konfigurationen je nach Anwendungsfall erlaubt. Für den hauptsächlichen Verwendungszweck wird hier zunächst die Regelmaschine instanziiert. Anschließend werden die Route Funktionalitäten deklariert [30]. Die Routes wiederum filtern die Anfragen nach den relevanten Daten, um diese an die Regelmaschinen-Instanz und ihren Funktionen weiterzugeben.

### **3.2.3 Syntax und Grammatik der Fakten und Regeln**

Die relevanten Daten der Dokumente können nun durch den Extraktionsvorgang erhalten werden. Damit die Steuerung beziehungsweise die Anfrage für die Datenextraktion den Benutzer:innen zugänglich gemacht wird, muss eine geeignete Syntax für die Regel- und Faktendefinition festgelegt werden.

Die Fakten werden deshalb zunächst als eine Menge der zu extrahierenden Objekte definiert. So wird eine Obermenge abgebildet, die als Hierarchie eine Teilmenge von Textfeldern, eine Teilmenge von konventionellen Tabellen, eine Teilmenge von unkonventionellen Tabellen und eine Teilmenge von benutzer-

definierten Tabellen besitzt. Die Teilmengen inkludieren jeweils auch die leere Menge und besitzen als Elemente die Extraktionsziele, welche sich an der Namenskonvention vom Python Style Guide orientieren, und somit einen beliebigen Namen tragen können [56]. Für jedes Extraktionsobjekt gibt es benötigte und optionale Eigenschaften, denen Werte zugewiesen werden. Ein Fakt ist daher eine Wertezuweisung einer Eigenschaft eines Extraktionsziel.

Die Regeln werden neben den Fakten als weitere Obermenge definiert. Diese Menge lässt sich als eine Liste von Regeln darstellen, welche nach den selben Namenskonventionen der Fakten benannt werden. Als Element in der Regelmenge besitzt jede Regel einen Regelausdruck, eine Regelaktion und eine Liste von Regelparametern, jene abhängig von der Regelaktion angegeben werden. Die Namen der Regelaktionen müssen dabei mit den Namen der gegebenen Funktionalitäten übereinstimmen, damit diese später von der Regelmaschine erkannt und mit den angegebenen Regelparametern ausgeführt werden können. Die Regelausdrücke werden als ein String angegeben und müssen zunächst syntaktisch verarbeitet werden. Hierfür existieren verschiedene Syntaxansätze.

Regelbasierte Systeme wie *CLIPS* nutzen, wie bereits beschrieben, eine angepasste Version der Syntax aus der Programmiersprache Lisp. Diese Syntax bildet mit Hilfe von Schlüsselwörtern die Ausdrücke für Regeln und Fakten. Python ermöglicht es Texte mittels der integrierten *eval()* Funktion zu evaluieren und, falls es sich um eine Python Aussage handelt, diese auch auszuführen. Die Funktion *literal\_eval()* aus dem Python-Modul *Abstract Syntax Tree* (*ast*) besitzt dieselbe Funktionalität mit einer eingeschränkten Untermenge für die Python Ausdrücke. Hierbei werden bei beiden Funktionalitäten die Texteingaben mit Hilfe des Parsings in abstrakte Syntaxbäume abgebildet, die die daraus generierten Terminale und Operatoren beinhalten. Bibliotheken, wie die von Google entwickelte Software Common Expression Language (CEL) erlauben durch das Auswerten von Ausdrücken die Bildung einer Semantik [32]. Hierzu wird eine Grammatik benutzt, dergleichen die Ausdrücke mit Hilfe von Regeln in ihre Typen und Namen als Bestandteile aufsplittet.

Das für dieses Projekt entwickelte regelbasierte System für die Komponente der symbolischen KI bedarf Regeln in Form von Booleschen Ausdrücken. Damit das regelbasierte System die Ausdrücke für die Regeln und Fakten gleichermaßen auswerten kann, braucht es eine für den Anwendungsfall angepasste Definition einer Grammatik. Zu diesem Zweck nutze ich die Parsing Bibliothek Lark. Jene erlaubt die Erstellung einer kontextfreien Grammatik in der Erweiterten-Backus-Naur-Form (EBNF). Basierend auf dieser Bibliothek werden unter der Nutzung eines LALR-Parsers abstrakte Syntaxbäume aus Texteingaben generiert [36]. Im Listing 3.1 wird die implementierte Grammatik für das regelba-

```
1      rule_grammar = r"""
2      ?start: disjunction
3
4      ?disjunction: conjunction
5                    | disjunction "or" conjunction -> or_
6
7      ?conjunction: expression
8                    | conjunction "and" expression -> and_
9
10     ?expression: expression operator expression
11                  | value
12
13     ?operator: comp_operator
14                | ident_operator -> ident_operator
15                | memb_operator -> memb_operator
16
17     ?comp_operator: "<" -> lt
18                    | ">" -> gt
19                    | "==" -> eq
20                    | ">=" -> ge
21                    | "<=" -> le
22                    | "!=" -> ne
23
24     ?ident_operator: "is" -> is_
25                     | "is" "not" -> is_not
26
27     ?memb_operator: "in" -> in_
28                     | "not" "in" -> not_in
29
30     OBJECT_NAME: CNAME
31     PROPERTY_NAME: OBJECT_NAME ( "." OBJECT_NAME | NUMBER ) *
32     CHAR: /[\\x27].[\\x27]/
33
34     ?value: PROPERTY_NAME -> property_name
35            | "True" -> true
36            | "False" -> false
37            | "null" -> null
38            | NUMBER -> number
39            | CHAR -> char
40            | ESCAPED_STRING -> string
41
42     %import common.CNAME
43     %import common.NUMBER
44     %import common.ESCAPED_STRING
45     %import common.WS
46     %ignore WS
47     """
```

Listing 3.1: kontextfreie Grammatik für die Regelevaluation

sierte System als Rohstring dargestellt. Dieser Rohstring wird dann dem Parser übermittelt, sodass jeglicher Regel- und Faktenausdruck durch einen abstrakten Syntaxbaum der Logik für die Evaluation dieser Regel bereitgestellt wird. Die booleschen Ausdrücke werden in der Syntax von Python genutzt, infolgedessen die Evaluation direkt stattfinden kann.

Der Aufbau der umgesetzten Grammatik basiert auf der klassischen Aussagenlogik und ihrer Definitionen[10]. Auf der Syntaxebene werden daher in EBNF die Produktionsregeln kleingeschrieben und die Terminale großgeschrieben. Die Regel *start* wird zunächst als Startsymbol definiert. Von ihr lässt sich jedes Wort ableiten, welches die *disjunction* Regel erfüllt. Jene Regel stellt eine aussagenlogische Oder-Verknüpfung beziehungsweise eine Disjunktion dar, welche wiederum aus der Regel für die Konjunktion oder für die Oder-Verknüpfung aus Disjunktion und Konjunktion ersetzt werden kann. Die Konjunktionsregel *conjunction* kann entweder durch die *expression* Regel ersetzt werden oder durch eine Und-Verknüpfung der *conjunction* Regel und der *expression* Regel. Die *expression* Regel stellt einen booleschen Ausdruck dar, welcher sich entweder mit der *value* Regel oder aus einer Kombination von einer *expression* Regel, von einer *operator* Regel und von einer weiteren *expression* Regel ableiten lässt.

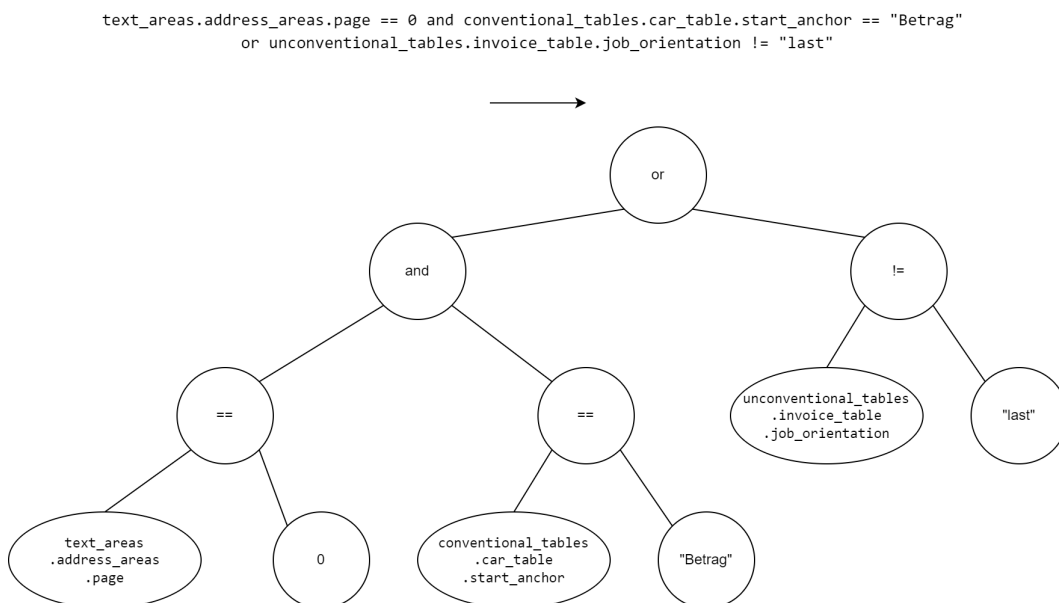


Abbildung 3.2: Generierter Syntaxbaum vom LALR-Parser

Bei der *value* Regel wird die Mehrdeutigkeit der Terminale definiert, für die alle primitiven Datentypen abgedeckt werden. Die Terminale *PROPERTY\_NAME* wird hier hervorgehoben, da diese den regulären Ausdruck einer Eigenschaft



eines Extraktionsobjekts abbilden soll. Dabei ist dies als eine Verkettung durch weitere Objektnamen durch die Punktnotation definiert. Für die *operator* Regel bestehen die Mehrdeutigkeiten aus der *comp\_operator* Regel, der *ident\_operator* Regel und der *memb\_operator* Regel. In der von mir aufgezählten Reihenfolge bilden die Vergleichsoperatoren, die Identitätsoperatoren und die Memberoperatoren ab. Die in Python existenten arithmetischen Operatoren, Zuweisungsoperatoren und bitweisen Operatoren werden für die aus der Grammatik erzeugten Sprache hingegen nicht definiert. Des Weiteren wird das Klammern für die Ausdrücke nicht unterstützt, weswegen die Operatorrangfolge durch die Produktionsregeln der Grammatik fest vorgegeben ist. Die restlichen Zeilen beinhalten die Importe der zusätzlichen Module für die Definition der Grammatik. Abbildung 3.2 zeigt ein Beispiel wie aus einem Regelausdruck ein Syntaxbaum durch den Parser kreiert werden kann. Mittels der Transformer Klasse der Lark-Bibliothek wird als Zwischenschritt die Syntax der einzelnen Blätter und Knoten des Baums so angepasst, dass diese für die weitere Verarbeitung und Evaluation der Regelmaschine bereitstehen.

### 3.2.4 Regelmaschine

Nachdem unterdessen die Syntax und Grammatik der Regeln gegeben ist, wird die Regelmaschine als Modul entwickelt. Dieses Modul besteht aus den Klassen für den Regelinterpreter und den Regelspeicher. Während die Regelinterpreter-Klasse die konkrete Logik der Vorwärtsverkettung beinhaltet, kümmert sich die Regelspeicher-Klasse um das Bereitstellen und um die Speicherung der Regeln und Fakten. Softwaretechnisch wird für diese Speicher-Klasse das Singleton Entwurfsmuster benutzt, mit diesem die Regelmaschine nur eine Regelspeicher Instanz referenzieren kann. Dadurch ist ein fester Zustand der Regel- und Faktenmenge gegeben.

Die Speicherung der Regeln und Fakten kann über den persistenten Ansatz der Datenbanken ermöglicht werden. Jedoch erlaubt der Ansatz der Speicherung der Regeln und Fakten via einer Datei im Dateisystem den schnellen Zugriff und die schnelle Bearbeitung jener. Des Weiteren stellt sich hier die Frage, in welchem Dateiformat die Regeln und Fakten festgehalten werden: Ein gängiges Serialisierungsformat wie JSON eignet sich durch ein festgelegtes Schema zur Speicherung, da die Daten strukturiert aufbewahrt werden. Demgegenüber existieren weitere Dateiformate, die eine geeignete Alternative darstellen. Eines dieser Alternativen ist das Dateiformat YAML, welches auch als eine vereinfachte Auszeichnungssprache gilt. Jene Sprache bildet die Obermenge zu JSON ab und ermöglicht neben den auch in JSON definierten generischen Datentypen

einen übersichtlicheren, aber auch komplexeren Ansatz für ein Serialisierungsformat. Die Sprache YAML besitzt im Vergleich zu JSON eine bessere Lesbarkeit der Syntax, ist kompakter gestaltet und bietet eine hohe Kompatibilität. [6]. Die Syntax für mögliche Kommentare ist obendrein auch verfügbar, wodurch sich diese Auszeichnungssprache für Konfigurationsdateien eignet, das wiederum für die Definitionen der Regeln und Fakten abgebildet werden kann.

Die Serialisierungsformate unterscheiden sich letztendlich in der Zielsetzung: JSON wird für den leichten Datenaustausch für Webanwendungen aufgrund der leichten Serialisierung und Übersetzung mit dem Parser benutzt. YAML dagegen hat eine erhöhte Komplexität für die Serialisierung und Übersetzung, aber ist für Dateien konzipiert, die von Menschen bearbeitet und genutzt werden.

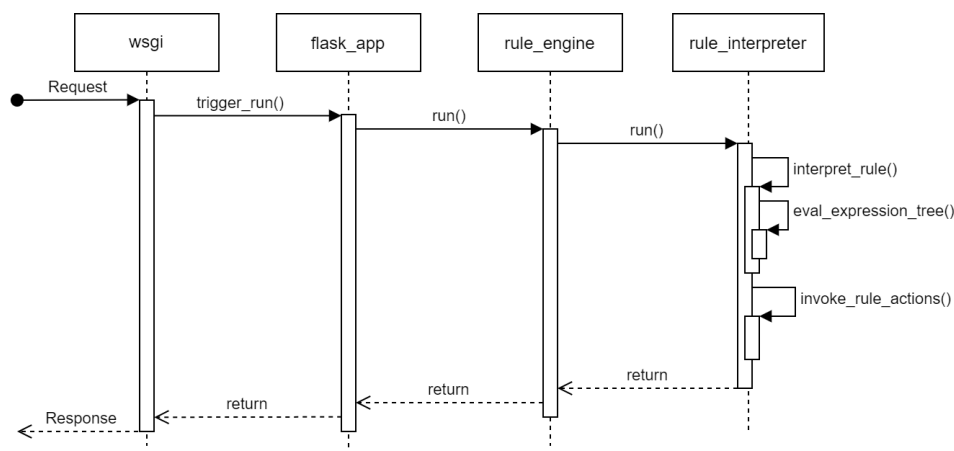


Abbildung 3.3: Durchlauf der Regelinterpretationen durch ein Request

Mit Hilfe einer Bibliothek für die Serialisierung und Übersetzung von YAML-Dateien werden ebendaher die Regeln und Fakten während der Laufzeit geladen, bearbeitet und gespeichert. Dabei wird sich für den Ansatz entschieden, die Regeln und Fakten für jede einzelne Dokumentenart zu speichern und zu verwenden. Während die booleschen Ausdrücke der Regeln als String dem Parser für die Grammatik übergeben werden, muss aus den Fakten nach dem Laden während der Laufzeit eine für die Evaluation angepasste Version generiert werden. Hierfür werden durch eine Funktionalität die Fakten als eine Menge beziehungsweise als ein Dictionary erzeugt, die den Pfad zur Eigenschaft in der Punktnotation als Schlüssel besitzen, angegeben. Ebenjene Konvertierung der Fakten ist für die Auswertung mit den Regeln im späteren Prozess geeignet.

Für den Ablauf wird, wie in Kapitel 3.2.2 beschrieben, zunächst die Regelmaschine beim Starten des Systems instanziiert. Diese wiederum instanziiert den Regelspeicher, welcher die Regeln und Fakten bereit hält, und den Regelinterpre-

ter, welcher den Parser für die Grammatik Instanz erstellt. Die Funktionalitäten des Interpreters werden aufgerufen, sobald die HTTP Anfrage bezüglich des Durchlaufens der Regelevaluation im WSGI ankommt.

Die Abbildung 3.3 zeigt den Verlauf bei dieser Anfrage und den damit entstehenden Aufrufstapel. So löst zunächst die Route-Anfrage die gekoppelte Funktion *trigger\_run()* der Flask Instanz aus, welche wiederum die *run()* Methode der Regelmaschine betätigt. Diese ruft dann die *run()* Methode des Regelinterpreters mit der Angabe, für welche Dokumentart der Vorgang stattfinden soll, auf. Dort werden iterativ die aktuell gespeicherten Regeln mit Hilfe der Funktion *interpret\_rule()* zunächst evaluiert. Diese Funktion übergibt den Ausdruck der einzelnen Regel dem Parser, mit welchem die Funktion *eval\_expression\_tree()* aufgerufen wird, um den Syntaxbaum zu erstellen.

Innerhalb dieser Methode wird rekursiv über den Syntaxbaum navigiert und die einzelnen booleschen Ausdrücke von den Blättern bis zur Wurzel aufgelöst. Dieser boolescher Wert wird danach bis zum Methodenaufruf in der *run()* Funktion zurückgeben, sodass daraufhin die Regel bei einem positiven Ergebnis zu der Liste der erfüllten Regeln hinzugefügt wird. Ist die Schleife der Regelevaluation abgeschlossen, werden die Aktionen der erfüllten Regeln mit ihren Parametern in der *invoke\_rule\_action()* Funktion ausgeführt. Hierbei werden die Aktionen abhängig des ausgewählten Dokumententyps aufgerufen. Die Ergebnisse der aufgerufenen Funktionen werden dann mit den Regelnamen indiziert und bis zur WSGI Klasse, in welcher der Wert als HTTP Antwort zurückgeschickt wird, gesammelt zurückzugeben. Der Vorgang deckt somit den Hauptprozess des regelbasierten Ansatzes im Teilsystem.

### 3.3 Entwicklungsumgebung

Damit die Funktionalitäten der Teilsysteme, die die unterschiedlichen Ansätze der künstlichen Intelligenz umsetzen, genutzt werden, bedarf es unter anderem die Steuerung jener über menschliche Eingaben. So müssen sowohl die Möglichkeiten der Konfiguration der Regeln und Fakten für die Benutzer:innen als auch das Starten des maschinellen Lernens über eine Benutzungsschnittstelle gegeben sein. Wie in Kapitel 3.2.4 bereits erwähnt, muss eine Umgebung zur Verfügung gestellt werden, die das Bearbeiten der Regel- und Faktendefinitionen erlaubt.

Die Hauptfunktionalitäten wie das Starten, das Überprüfen der Änderungen und das Pausieren sollen direkt erreichbar sein, weswegen sich die Positionierung dieser in der Navigationsleiste eignet. Des Weiteren soll der Dokumen-

tentyp für den Extraktionsprozess auswählbar sein und die Dokumente für den ausgewählten Dokumententyp in dieser Umgebung als eine Seitenansicht angezeigt werden. Auch sind die Ergebnisse der Extraktionsprozesse für die Benutzer:innen für den Abschluss und für die Entwicklung relevant, weswegen eine Anzeige dafür erstellt wird. Das Layout der Benutzungsoberfläche orientiert sich hierbei an bekannten Entwicklungsumgebungen wie Visual Studio [39] oder IntelliJ [34]. Durch dieses Layout wird eine intuitivere Interaktion geboten.

Da die Schnittstellen über das Netzwerk der Client-Server-Architektur gegeben sind und die Priorität auf der Zugänglichkeit und Integration in das System liegt, habe ich mich für die Umsetzung der Entwicklungsumgebung als Webanwendung entschieden. Für die graphische Schnittstelle der Benutzer:innen des Systems wird daher die UI Bibliothek Vue benutzt, die für eine kleine prototypische Umsetzung mit anderen Bibliotheken wie React bei dem Kompilieren und Laden zeitlich mithalten kann [13]. Bei Vue habe ich mich für die dritte Hauptversion mit der Composition Application Interface (API), die eine reaktive Programmierung für das ermöglicht, und für die Nutzung von dem Dateiformat Single File Components (SFC) entschieden [58]. Des Weiteren wird das Build Tool Vite benutzt, die mit Hilfe eines Optimierungsansatzes die Kompilierung und die Übersetzung des Codes auf das Frontend umlagert [57].

Für Vite wird daher eine Konfigurationsdatei angelegt, die Einträge über den Entwicklungsserver und die verschiedenen Worker für den im späteren Verlauf vorgestellten Editor besitzt. Dazu nutzt man konventionell der Node Package Manager (npm), um weitere Abhängigkeiten für das Frontend zu installieren [40]. Weitere Eigenschaften des Frontends sind die Nutzung von der Programmiersprache TypeScript, mit welcher die Typsicherheit für die dynamische Skriptsprache gewährleistet wird [2], und dem Entwicklungswerkzeug PostCSS. Mit PostCSS wird eine detaillierte und übersichtliche Definition für Style Selektoren ermöglicht [45]. Als Einstiegspunkt wird klassischerweise eine Hyper Text Markup Language (HTML) Datei verwendet, die das Skript zum Instanzieren der Vue Anwendung auslöst. An jener Stelle werden wiederum die weiteren über npm bereits installierten Bibliotheken als Plugin in die Vue Anwendung installiert.

Für die Vue Anwendung wird die Struktur modular gehalten, weswegen verschiedene Vue Komponente definiert werden. Diese werden in die Hauptkomponenten und weiteren Unterkomponenten hierarchisiert. So instanziiert die Hauptkomponente die Komponenten für die Bereiche der Hauptansicht. Darunter fallen die Bereiche für die Navigationsleiste, für den Editor, für die PDF-Ansicht und für den Output für diese die jeweiligen Komponenten als Modul definiert sind. In Abbildung 3.4 die im Rahmen meiner Thesis entwickelten prototypische Version

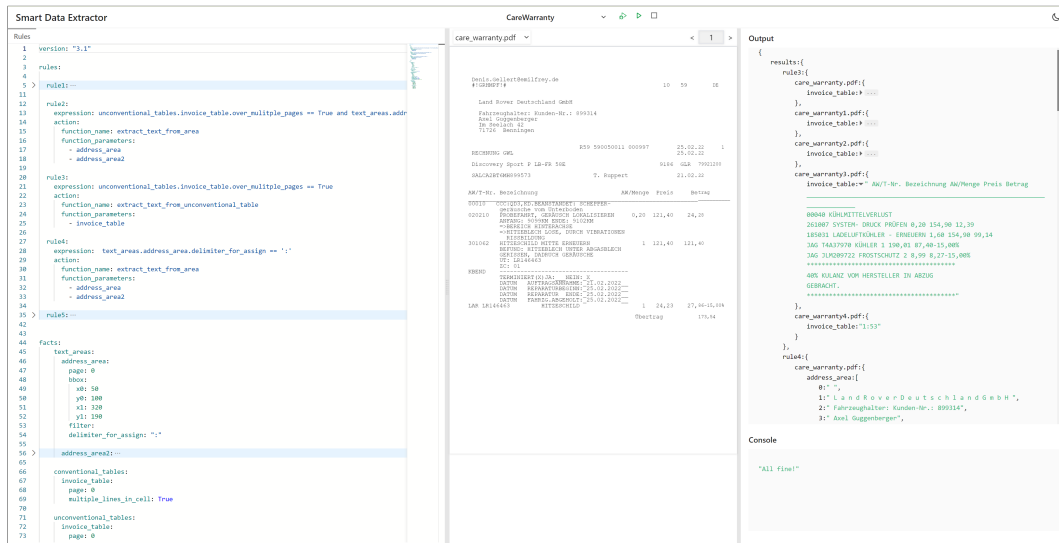


Abbildung 3.4: Hauptansicht der Entwicklungsumgebung

der Entwicklungsumgebung angezeigt. Hier sind die beschriebenen Bereiche zu sehen, die in eigene Module unterteilt wurden. Zwischen den Hauptbereichen befinden sich Trennbalken, die durch eine weitere Bibliothek implementiert wurde und mit deren Hilfe die Größen der verschiedenen Bereiche verändert werden können. Eine wichtige Eigenschaft ist die Nutzung eines State-Management-Systems unter Zuhilfenahme der Bibliothek *Pinia* [44]. Dies ermöglicht das Zwischenspeichern der einzelnen Zustände der Komponenten, die dem Server mitgeteilt werden kann, oder das Erhalten des letzten Zustands für das Neu-laden der Webanwendung. So können beim initialen Aufruf die letzten Werte geladen werden, sodass diese Anwendung offline-fähig dargestellt werden kann.

### 3.3.1 Editor

Die Konfiguration der Regeln und Fakten bedarf eine UI Schnittstelle für das Erstellen und Bearbeiten dieser. Dies wird gängigerweise in Form eines Texteditors ermöglicht. Im Webbereich stehen für die Integration eines Texteditors mindestens mehrere Bibliotheken als Kandidaten zu Verfügung. Im Rahmen des Projekts wurde sich für die Nutzung des Monaco-Editor entschieden, welches die Kernkomponente des Editors Visual Studio Code [38] darstellt und unabhängig von ihr als Open-Source Projekt bereitgestellt wird.

Die folgenden Eigenschaften führten zu dieser Auswahl: Der Monaco-Editor bietet viele Funktionalitäten für eine angenehme Bearbeitung, darunter die Autovervollständigung durch IntelliSense, Textsuche und Code Linting an. Die

Bibliothek wird durch eine weite Community gewartet, besitzt allerdings einen hohen Umfang des Software-Bundlings und erlaubt auch keine Lazy Loading der Funktionalitäten als Entwurfsmuster. Des Weiteren müssen die Worker für die einzeln genutzten Sprachen konfiguriert werden. Diese Worker werden in der zuständigen Vue Komponente definiert.

Danach wird nach der Reihenfolge das JSON Schema vom Server angefragt und in den Editor geladen, welches zusätzlich für IntelliSense und das Code Linting definiert wurde [35]. Im weiteren Code Verlauf werden weitere Einstellungen, wie die gewählte Sprache, für den Monaco-Editor angegeben, sodass dieser instanziiert werden kann. So wird die Sprache für das Highlighting auf YAML gestellt, welches als Dateiformat für die Regeln und Fakten gewählt wurde. Hier wird zunächst ein Request zum Laden der Einträge des ausgewählten Dokumenttyps ausgeführt.

Als Callback wird daraufhin festgelegt, dass nach jeder Textänderung wenige Sekunden später ein Request für die Speicherung dieser Änderungen zu dem jeweiligen Dokumententyp geschickt wird. Somit werden die Zustände der Konfiguration auf der Server Seite festgehalten. Auf der Benutzungsoberfläche wird dann der Texteditor angezeigt, der auch die Navigation über Tabs unterstützt. So können die Regeln und Fakten in einem Tab oder in zwei getrennte Tabs gehalten werden, was wiederum eine übersichtlichere Aufteilung verschafft.

### **3.3.2 Dokumentenanzeige**

Damit die Entwicklung und das Anpassen der Regeln und Fakten angenehmer verläuft, bedarf es einer Anzeige der Dokumente als Orientierungspunkt. Diese umgesetzte Anzeige erlaubt es, eine einzelne Seitenansicht der Dokumente des ausgewählten Dokumenttyps darzustellen. Hierfür steht ein Selektor bereit, welcher eine Auswahl aus allen Dokumenten des Dokumenttyps gewährleistet und einen schnellen Wechsel anbietet. Daneben ist das Vor- und Zurückblättern mit zwei Knöpfen beziehungsweise das Umblättern mit Hilfe einer Seitenanzahl gegeben.

Damit die Positionierung des zu extrahierenden Bereichs konfiguriert werden kann, helfen die Koordinaten, die abhängig des Mauszeigers als Referenzpunkt angezeigt werden. Dies wurde dynamisch mit der Hover-Funktion umgesetzt, mit dieser die Benutzer:innen direkt die X- und Y-Koordinaten erhalten können. Außerdem soll die Anzeige im Zusammenspiel mit dem Editor es ermöglichen, die Bounding Box eines Bereichs durch das Ziehen mit der Maus auf der ausgewählten Seite zu generieren und diese in eine ausgewählte Stelle im Editor einzusetzen.

### 3.3.3 Output

Nachdem die Benutzer:innen ihre Konfiguration festgelegt und den Durchlauf gestartet haben, bedarf es für die Interaktion eine Rückgabe des Systems. Hierfür wurde sich für ein Ausgabebereich entschieden, welches zwei Fenster besitzt: Das erste Fenster stellt die Ergebnisse in einer Anzeige für JSON Dateien dar und ermöglicht damit eine genauere Betrachtung für die Benutzer:innen. Hierbei werden sowohl die konkreten Ergebnisse des Durchlaufs der Regeln als auch die Ergebnisse der Auswirkungen der Änderungen angezeigt.

Beim zweiten Fenster werden Informationen, die abseits von Ergebnissen, wie Zustände des Systems oder Fehlermeldungen angezeigt. Dieses Fenster soll es ermöglichen, einen aktuellen Zustand und eine Rückmeldung zu präsentieren, sodass im Zuge dessen die Mensch-Computer-Interaktion verbessert und die Nutzungsbarriere zum System niedriger wird. Unter den Zuständen zählen die erfolgreiche Inbetriebnahme, Hinweise zur Nutzung und Fehlermeldungen, die im System auftreten potentiell können.

## 3.4 Machine Learning

Während die Teilsysteme der symbolischen künstlichen Intelligenz und der Entwicklungsumgebung implementiert sind, bedarf es nun die Konzeption und Umsetzung des Machine-Learning-Ansatz. Für die gleiche Problematik, die bereits von dem Teilsystem der symbolischen KI thematisiert und verarbeitet wurde, soll mit dem Ansatz ein weiterer Lösungsweg zur Bewältigung dieser gewählt werden.

Neben dem maschinellen Lernen existieren unter anderem die Arten des überwachten Lernens (supervised Learning) und des nicht überwachten Lernens (unsupervised Learning). Beide haben ihre Einsatzgebiete auch innerhalb der Dokumentenextraktion. Hierbei zählen zu dem Bereich der Extraktion verschiedene Dokumentarten wie Text- oder HTML Dokumenten hinzu, bei welchem die letztere genannte Art präzise Ergebnisse bei der Verwendung der Methoden des überwachten Lernens verschafft [8]. Nicht überwachtetes Lernen hat im Gegensatz zu Dokumenten, die kein Labeling besitzen, einen höheren Wirksamkeit.

Über beide Arten des maschinellen Lernens können Daten strukturiert und modelliert werden. Doch existieren neben den weit verbreiteten zwei Kategorien auch weitere Bereiche im Machine Learning. Das semi-überwachte Lernen verhilft das Labeling durch weitere Annahmen der Struktur unter Zuhilfenahme von semi-überwachter Klassifikation und durch eingeschränktes Clustering zum

Durchbruch [19]. Das schwache überwachte Lernen verwendet als Technik die Datenmodelle über neu generierte Daten. Dies ist für den Fall vorgesehen, wenn man mit neuen Daten aus Dokumenten arbeiten möchte. In solch einem Fall eröffnet die Bibliothek Fonduer mit der Erstellung einer Pipeline-Struktur eine neue Lösungsmöglichkeit.

### 3.4.1 Pipeline

Fonduer setzt den Machine-Learning-Ansatz mit Hilfe der Kategorie eines abgeschwächten überwachten Lernen (weak supervised learning) um. Genauer gesagt wird hier die Idee des Knowledge Base Construction (KBC), also des Aufbaus einer Wissensbasis über stark formatierte Daten verfolgt. Diese Wissensbasis wird über eine Datenbank abgebildet, die die Daten strukturiert bereithält. Daher besteht für die Nutzung dieser Bibliothek eine Abhängigkeit von dem Datenbanksystem PostgreSQL [4]. Des Weiteren wird die Bibliothek spaCy genutzt, weswegen die Sprachmodelle für die deutsche Sprache in dem Dockerfile des Teilsystems mitgeladen und verlinkt werden. Über die Kompositionsdatei von Docker werden dann die Container für die Machine-Learning Komponente und die dazugehörige Datenbank virtuell aufgesetzt und mit dieser entwickelt.

Damit dies auch für vielschichtige Daten funktioniert, wird hier ein spezielles Datenmodell benutzt, welches die Dokumente und ihre Besonderheiten darstellt. So fallen verschieden Formate von Dokumenten wie XML, HTML und auch PDF unter dem Begriff der reichlich formatierten Daten. Hier sollen die textlichen, visuellen und strukturellen Eigenschaften der Dokumente mit ihrer semantischen Informationen verbunden werden, um sie für die weitere Verarbeitung zu nutzen. Diese Eigenschaft wird als Multimodalität bezeichnet und existiert in der Prozesskette an verschiedenen Stellen.

Die Eingabe für das Teilsystem findet in Form der Dokumente statt. Auch für dieses Teilsystem wird sich auf die Dokumente bezogen, die im Rahmen der Thesis betrachtet werden. Hierfür werden, wie auch im anderen Extraktionssystem, die PDF-Dokumente in ihre drei Dokumenttypen unterteilt. Die Menge der Dokumente für die jeweilige Art unterscheiden sich in ihrer Dimension: Für die Dokumente der Care Garantieaufträge ist die Anzahl der Elemente dreistellig, für die Formel1 Rechnung vierstellig und für die KFZ300 Rechnung fünfstellig. Dies soll eine Relation der Verarbeitung ermöglichen und damit einen zeitlichen Zeitverlauf darstellen.

Der Durchlauf von den Eingabedaten zu der Erstellung des Machine-Learning Modells wird als Pipeline bezeichnet. In Abbildung 3.5 wird die komplette Pi-



pipeline für die Verarbeitung des maschinellen Lernens dargestellt. So sind die einzelnen Phasen — aufgeteilt in zwei Kategorien — zu sehen. Die erste Hälfte der Phasen ist statisch vorgegeben und kann strukturell nicht verändert werden. Bei der zweiten Hälfte sind Eingriffe der Benutzer:innen beziehungsweise bei den Entwickler:innen möglich um die Datenmodelle zu präzisieren. Insgesamt sind die fünf Phasen relevant, die in der Hauptfunktionalität im Teilsystem umgesetzt wurden: In der ersten Phase wird zunächst die KBC über ein spezielles Schema und die daraus folgenden iterative Erstellung eines Eingabekorpus initialisiert. Dafür wird die Hostadresse der benötigten Datenbank angegeben und dergleichen einem Meta-Objekt, welches die Brücke zur Persistenz darstellt, übergeben, um die Datenbank parallel zu starten.

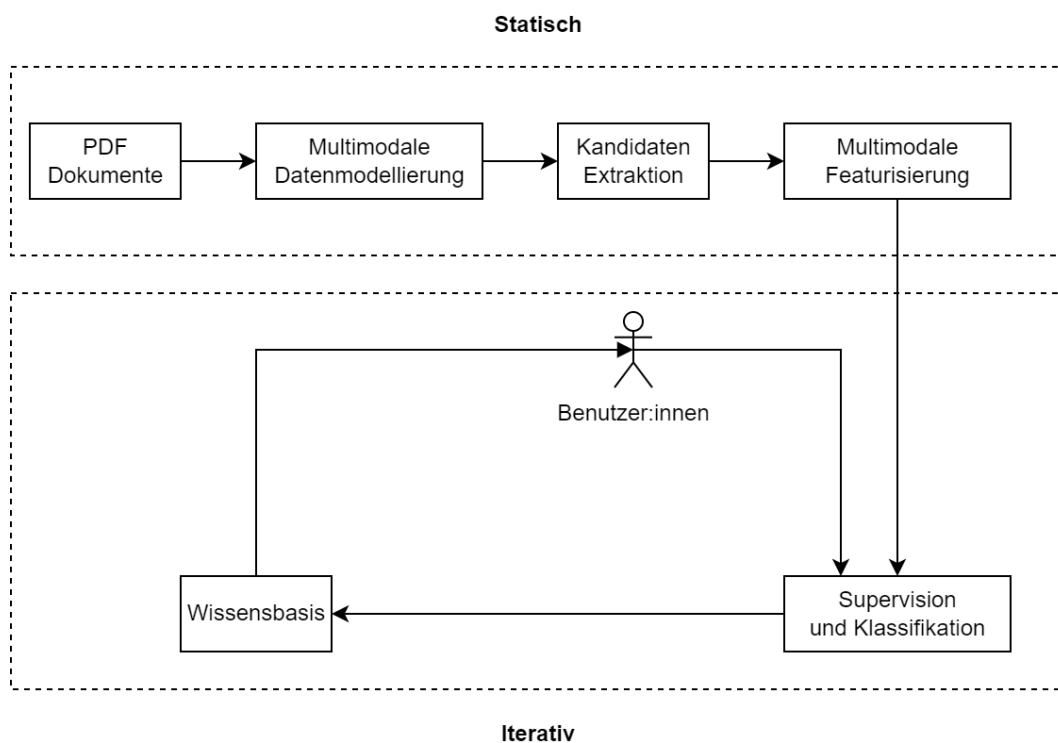


Abbildung 3.5: Fonduer Pipeline für das maschinelle Lernen

Die multimodale Datenmodellierung wird in dieser Phase durch das parallele Parsen der Dokumente ermöglicht, welches inhaltliche, strukturelle, tabellarische und visuelle Informationen bereit hält. Für das Parsing findet hierbei die Nutzung der HOCDocPreprocessor-Klasse statt, welche den Pfad der zu übersetzenden Dateien bei Instanziierung mitgegeben wird. Als Zwischenschritt werden die PDF-Dokumente als HOCD Dateien, einem Optical Character Recognition (OCR) Datenformat, welches auf HTML aufbaut, mittels einer internen

Klasse übersetzt. Ein weiterer Weg wäre es, dies über den HTMLPreprocessor umzusetzen, was aber bei digital verarbeiteten Dateien nicht von der Dokumentation empfohlen wird.

Als nächstes wird der Korpus-Parser benutzt, der mit Hilfe der Bibliothek spaCy die Dokumente in einzelne Sätze und Tokens aufteilt und über weitere Parametrisierung die Inklusion der strukturellen und sprachlichen Informationen angibt. In der zweiten Phase werden Mentions, welche die potentiellen Kandidaten für die Datenextraktion darstellen, definiert. In dem Anwendungsfall des Projekts wird für jedes Feld, das sich in den Aufträgen und den Rechnungen befindet und extrahiert werden soll, eine Mentions festgelegt. Beispiele hierfür sind, wie aus dem Teilsystem der symbolischen KI bereits bekannt, die Jobs, die Reparaturzeiten, die Fahrzeugmodelle und die Jobbeschreibungen.

Damit diese Mentions auch vollwertige Kandidaten des Extraktionsziels werden, bedarf es die Definitionen von Matchers, die über eine Filterangabe eine präzise Erkennung geben sollen. Die Matchers gibt es in verschiedenen Variationen. In diesem Anwendungsfall wird sich für die Filterung über reguläre Ausdrücke entschieden, da diese bei der symbolischen KI ihre Anwendungen finden. So wird zu jedem Mention-Objekt ein Matcher-Objekt entwickelt. Dazu kommen noch jeweils Definitionen für die N-Gramm Darstellung der Mention-Objekte, um so die Größe der Fragmente festzulegen.

Damit die Extraction der Mentions mit all diesen zusätzlichen Einschränkungen stattfinden kann, wird ein MentionExtractor-Objekt definiert, welches die Mentions und deren Matchers und N-Gramms mitgegeben wird, um mit dieser Kombination an Informationen die erwünschten Daten zu extrahieren. Nachdem diese Informationen auf das MentionExtractor-Objekt angewendet werden, wird das Schema der Kandidaten für die Extraktion angegeben. Hierfür braucht es eine Definition für jede Dokumentenart und ihre eigene Unterklasse der Candidate-Klassen. Folglich wird ein Warranty-Objekt für die Care-Aufträge, ein Invoice-Objekt für die Formel 1 Rechnungen und ein Invoice-Objekt für die KFZ3000 Rechnungen definiert. Diesen wird wie bei der Mentions-Klasse dieselben Matcher überreicht.

Dazu werden Throttlers, das sind Funktionen, die als weitere starke Filter fungieren, definiert, um die Genauigkeit der validen Extraktionskandidaten zu ermöglichen. In diesem Anwendungsfall definieren wir daher Throttler-Objekte, die sich mit den Ausrichtungen der verschiedenen Einträgen beschäftigen. So wird zum Beispiel überprüft, ob die Kandidaten für die Reparaturzeiten, der Beschreibung und der Menge horizontal zueinander ausgerichtet sind. Nachdem dies festgelegt ist, werden diese Paradigmen in einem CandidateExtractor-

Objekt kombiniert und angewendet. Für die Anwendung werden die Dokumente in Trainings- und Testsätze unterteilt.

In der dritten Phase werden den Kandidaten der Trainings- und Testsätzen durch multimodale Featurisierung Merkmale annotiert. Hierfür wird ein Featurizer-Objekt definiert, welchem die Verbindung zur Datenbank und den Kandidaten der jeweiligen Dokumentenart übergeben wird. In der vierten Phase findet das schwache überwachte Lernen und die Klassifikation über Labeling-Funktionen statt. Dafür werden diese Funktionen, welche weitere Bedingungen für die Kandidaten stellen und die es zu erfüllen gilt, definiert, um als valider Kandidat weiterverarbeitet zu werden.

Die Labeling Funktionalitäten werden hierbei durch die Bibliothek Snorkel zur Verfügung gestellt, welche das Labeling der Trainingssätze durch das schwache maschinelle Lernen schneller und günstiger der Ressourcenkosten gestaltet [14]. In dem Listing 3.2 werden zwei Beispiele der für das Teilsystem implementierte Label-Funktionen dargestellt. Hierfür wird neben den zwei gängigen Wahrheitswerten ein Wahrheitswert mit dem Schlüsselwort ABSTAIN definiert, welches gesetzt wird, wenn keine Entscheidung für eines der zwei Wahrheitswerte TRUE oder FALSE zu dem ausgeführten Zeitpunkt für das Labeling möglich ist.

```
1      ABSTAIN = -1
2      FALSE = 0
3      TRUE = 1
4
5      @labeling_function()
6      def LF_repair_times_row(c):
7          if 'repair_times' in get_row_ngrams(c.description):
8              return TRUE
9          else:
10             ABSTAIN
11
12     @labeling_function()
13     def LF_stamp_time_row(c):
14         if 'stamp_time' in get_row_ngrams(c.description):
15             return TRUE
16         else:
17             ABSTAIN
```

Listing 3.2: Labeling-Funktionen für die Supervision

Daher reichen zunächst einfache Funktionsmuster, die die tabellarischen Eigenschaften der Felder, wie zum Beispiel der Fahrzeugmodellnamen, des Kfz-Kennzeichens und des zuletzt abgelesenen Kilometerzählerstands, überprüfen.

Stehen die Labeling Funktionen, können diese auf das Labeler-Objekt angewendet werden. Hierbei können weitere Statistiken des Vorgangs als Feedback mitgegeben werden. Ist das Training abgeschlossen, werden als nächstes in der Pipeline die Labeling-Funktionen modelliert, sodass Konflikte betrachtet werden können. Mit Wahrscheinlichkeitsberechnungen werden dann weitere Labels auf die Trainingssätze angewendet, sodass danach ein generatives Modell angelernt ist.

Dies ermöglicht dem System und der Entwicklung iterativ die Labeling Funktionen für eine höhere Genauigkeit zu verbessern. Hierfür müssen die Konflikte und die Überlappungen programmatisch weiter überprüft werden. Für das weitere Training der Datensätze wird das discriminative Modell durch Regression antrainiert. Dafür benutzt Fonduer das Machine-Learning Framework Emmental, das die Funktionalitäten für die Modellierung bereitstellt [16]. In diesem Prozess werden weitere Tokens und Rauschproben generiert, damit das Modelltraining unterstützt wird. Zum Schluss können die Testsätze mit ihrem Labeling genutzt werden, um die Vorhersage durch das Training der KI zu evaluieren.

### **3.4.2 Webserver**

Damit die Steuerung der künstliche Intelligenz des maschinellen Lernens und ihre Funktionalitäten gegeben ist, wird die Kommunikation der Teilsysteme benötigt. Diese Kommunikation findet über das interne Netzwerk statt und ermöglicht damit auch die Interaktion der Docker Container Umgebung. Dass diese nötigen Eigenschaften für das Teilsystem erfüllt werden können, wird derselbe Ansatz des als erstes vorgestelltes Teilsystems umgesetzt. Das bedeutet eine Nutzung der WSGI Schnittstelle, die auch die Verwendung einer Flask Anwendung als Webserver zur Folge hat.

So wird auch eine Factory-Methode für verschiedene Anwendungsfälle, wie unter anderem die Continuous Integration, bereitgestellt. Die Entwicklungsumgebung erlaubt hierbei das Starten der Verarbeitung der Trainingsdaten, weswegen sich hier die die Funktionalität nur auf den Durchlauf mittels des Machine-Learning-Ansatzes reduziert.

## Kapitel 4

# Ergebnisse

Nachdem ich nun im bisherigen Verlauf meine Vorgehensweise und Durchführung geschildert habe, zeige ich nun meine Ergebnisse:

Mein Projekt hat insgesamt verschiedene Merkmale meiner eingangs geschilderten Problematik hervorgehoben. So wurde im Zuge meiner Analyse zuerst gezeigt, dass eine weitere Verarbeitung der aus den Autohäusern entstandenen Dokumente ein sinnvoller Anwendungsbereich für das von den Mitarbeiter:innen genutzte ERP-System ist. Des Weiteren wurde deutlich, dass die Dokumente, auf die ich mich im Rahmen der Thesis beschränkt habe, viele verschiedene Formen besitzen und teilweise beliebig gewählt werden können. Die Unterschiede zwischen den einzelnen Dokumenten liegen mehrheitlich in den Details der einzelnen Extraktionsbereiche. Das Data-Mining dafür erwies sich mit der passenden Extraktionsbibliothek als hilfreich, da die Daten für das System und deren Logik greifbarer und leichter zu verarbeiten waren als zum Beispiel über gängige OCR Erkennung.

Eine Anpassung an die Definitionen des Dokumentenmusters muss in der symbolischen künstlichen Intelligenz durch eine Kategorisierung in Dokumentarten vorgenommen werden. Im Zuge dieser Anpassung fiel auf, dass sowohl die Definitionen von Tabellen in elektronisch verarbeiteten Dokumenten, als auch die Erkennung dieser Tabellen sich als recht umständlich erweist, da die nötige Standardisierung einer Tabelle fehlte. Verschiedene Semantiken für die einzelnen vorgestellten Felder benötigen eine weitere Differenzierung der Extraktion. So bedarf es zunächst Analysen der Struktur und des Kontexts, um eine angemessene Verarbeitung zu ermöglichen. Texte und Tabellen können über Regel- und Faktendefinitionen extrahiert werden und übersichtlich in einer Webanwendung dargestellt werden.

Das System, das ich im Rahmen dieser Thesis entwickelt habe, erwies sich als

Prototyp späterer Systeme mit derselben Aufgabe. Die Struktur für die Teilsysteme wurde durch eine container-getriebene Entwicklung modular gehalten und sorgte durch die Architektur für eine gemeinsame Kommunikationsebene für diese Komponenten. So konnte ein Projekt aufgesetzt werden, welches die neuesten Technologien in den Bereichen, in denen diese vorkommen und agieren, nutzte, um die erwünschten Extraktionsziele zu erhalten oder mindestens eine Annäherung zu erreichen.

Der Webansatz für die Entwicklungsumgebung zeigte bis auf die zusätzliche Build-Zeit während des initialen Ladens eine konsistente Verwendung und kein fehlerhaftes Verhalten auf. Zudem war die Entwicklung einer Grammatik für den Anwendungsfall wirksam, da sich durch die Entwicklung eine Sprache für die Ausdrücke entwickeln ließ, mit welcher die Benutzer:innen und die Entwickler:innen schnelle und simple Regeln konfigurieren konnten. Die Verarbeitung der Dokumente wird durch die daraus generierten Webanfragen ermöglicht, sodass der Durchstich der Funktionalitäten stattfindet.

Des Weiteren können die Unterschiede bei Veränderungen der Regeln und Fakten angezeigt werden, was die iterative Entwicklung dieser Konfigurationen anbietet. Für die Umsetzung des Machine-Learning Ansatzes erwiesen sich der Ansatz des schwachen maschinellen Lernens als eine vielversprechende Alternative. So war die Umsetzung einer Pipeline für die PDF Dokumente möglich. Dafür wurden für jede Dokumentenart Trainings- und Testsätze der Dokumente genutzt, um die Supervision und Klassifikation zu nutzen.

Die Datenbank wurde mit Informationen durch die KI befüllt, wodurch sowohl Tabellen für die Erwähnungen und Kandidaten, als auch für die einzelnen Felder, die für die Extraktionsfunktionalität definiert wurden, erstellt. Hierbei habe ich beobachtet, dass über einen längeren Zeitverlauf die Tabellen der Sätze und Paragraphen zuerst genutzt wurden, um die multimodale Modellierung der Dokumente abzubilden. Für den weiteren Verlauf zeigte sich aber auch, dass das genutzte Machine-Learning Framework sehr viele Ressourcen benötigt. Trotz Virtualisierung traten bei einem längeren Trainingsdurchlauf Abstürze auf, welche die Einträge in der Datenbank trotz Wiederherstellung über Mounting beeinträchtigten.

## 4.1 Testing

Für die prototypische und zukünftige Entwicklung des Projekts ist das Testing für die weitere Benutzung relevant. Hauptsächlich hierfür verantwortlich soll die Continuous Integration Pipeline sein. Zu Grunde liegen hier die Images der

Teilsysteme, die als Dockerfile beschrieben sind und in der Container Registry vom unternehmensinternen Repository gespeichert werden. Über verschiedene Stages sollen dann die Hauptaspekte der Teilsysteme in das System integriert werden. Hier werden diese Komponente auf verschiedenen Eigenschaften überprüft.

Für das Teilsystem der symbolischen KI wurden das Unit Testing und das nicht funktionale Testing über die Continuous Integration benutzt. Für die Entwicklungsumgebung wurde der Ansatz der UI Tests gewählt, um nicht nur die Benutzung sondern auch die mögliche Automatisierung für weitere Eingaben zu überprüfen. Bei dem für das maschinelle Lernen zuständige Teilsystem wurde das Testing über die Testsätze, die durch das beschriebene ERP-System bereitgestellt wurden, ermöglicht.

## Kapitel 5

# Konklusion und Ausblick

In dieser Arbeit ging es darum, Informationen aus digital verarbeiteten PDF-Dokumenten zu extrahieren und aus diesen neue Daten zu generieren. Die Dokumente, um die es in dieser Arbeit ging, stammten von Kund:innen von Ilexius GmbH. Jene Kund:innen sind unter anderem Beschäftigte von Autohäusern und deren Werkstätten. Mit meiner Systementwicklung, die ich in dieser Thesis demonstriert habe, werden Informationen aus dem Endprodukt aller Arbeitsschritte der Kund:innen für die Entwickelnden besser zugänglich gemacht, indem die Daten leichter extrahiert werden, sodass Ilexius diese Daten für weitere Arbeitsschritte im ERP-System bereitgestellt werden kann. Für meine Entwicklung habe ich das Data-Mining im Bereich der Dokumentenextraktion genutzt, das dabei hilft, laufende Prozesse detaillierter zu steuern und zu überwachen.

Neben dem Data-Mining nutzte ich zudem den Ansatz einer symbolischen künstlichen Intelligenz. Diese KI ermöglichte mir während meines Arbeitsprozesses in jedem Schritt eine Feinschleifung meiner bisherigen Arbeit der manuellen Extraktion. Durch diese KI kann sich jede/r Benutzer/in meines hier entwickelten Systems grobe bis zu detaillierten und angepassten Definitionen mit Hilfe einer eigenständigen Umgebung erstellen. Jene Definitionen erlauben dem/der Benutzer:in mittels Feedback des Systems eine Annäherung. Im späteren Verlauf meiner Arbeit wurde der Ansatz diese Annäherung im Teilsystem des Machine Learnings betrachtet. Bevor meine symbolische KI als auch das Machine Learning starten konnte, habe ich zu Beginn die Funktionalitäten der Bibliotheken genutzt, um durch Extraktionsergebnisse der Datenansätze die Erfolgsquote zu überprüfen. Danach definierte ich das Format für die Regeln und Fakten des Extraktionsvorgang unter Zuhilfenahme einer eigens entworfenen Grammatik. Anschließend stellte ich die Benutzungsoberfläche fertig, sodass ich diese mit der



symbolischen KI verbinden konnte. Zeitgleich konzeptionierte und entwickelte ich die zu dem Machine Learning gehörende Subsysteme. Dies tat ich in einem iterativen Ansatz, sodass eine Pipeline-Struktur erbaut wurde, mit der das Training des Machine Learnings baldmöglichst beginnen konnte. Im Anschluss daran fing ich an, mich in dem gegebenen Rahmen mit dem Software-Testing zu beschäftigen. Dadurch konnte die Ausführung aller meiner noch folgenden Prozesse gewährleistet werden. Als letzte Vorbereitung passte ich die grafische Benutzungsoberfläche weiter an, dass sich dieses mit den Systemen verbindet, wodurch die Ausführbarkeit der Funktionalität überprüft werden konnte.

Durch meine Arbeit habe ich es geschafft, ein System zu schreiben, das die Extraktion von PDF-Dokumenten ermöglicht. Dafür habe ich insgesamt zwei verschiedene Ansätze verwendet.

Während der Entwicklung meines Systems stieß ich allerdings auch auf einige Komplikationen, die es in Zukunft noch zu bewältigen und zu verbessern gilt. So ist unter anderem der Extraktionsvorgang äußerst abhängig von der Semantik des Extraktionsziels des jeweiligen Dokumentationstyps. In zukünftigen Arbeiten sollte das System deshalb so entwickelt werden, dass weitere Dokumentenarten ihre angepasste Extraktionslogik erhalten. Durch diese Weiterentwicklung könnten Zeit, als auch Ressourcen gespart werden. Darüber hinaus müssen, damit die Sprache für die Regelausdrücke erweitert werden kann, in Zukunft weitere Operatoren als auch die Klammerung in der Grammatik definiert werden.

Andernfalls bleibt die Sprache mit ihren jeweiligen Ausdrücken beschränkt, sodass mein System auch nur begrenzt nutzbar wäre. Zuletzt bedarf es für das maschinelle Lernen noch viele weitere Trainingseinheiten für andere Dokumentarten und eine weitere Anpassung der Label-Funktionen. Aufgrund des begrenzten Zeitraums, in dem mein System entwickelt worden ist, ist das Training des Machine Learnings nicht ganz so fortgeschritten, wie ich es zu Beginn meiner Thesis erhofft hatte. Dementsprechend gilt es, das Machine Learnings in den nächsten Monaten fortzusetzen, sodass weitere Anpassungsleistungen von Seiten des Machine Learnings erbracht werden.

Letztendlich habe ich ein System konzipiert und umgesetzt, das Entwickler:innen die Extraktion von Dokumenten für Mitarbeitende in Autohäusern ermöglicht. In der immer mehr digitalisierten Zukunft kann das von mir konzipierte System jedoch nicht nur auf Dokumente aus Autohäusern angewendet werden, sondern auch auf Dokumente jenseits der Automobilindustrie.

## Literaturverzeichnis

- [1] Milam W. Aiken and Olivia R. Liu Sheng. Nexpert object. *Expert Systems*, 7(1):54–57, 1990.
- [2] Gavin Bierman, Martín Abadi, and Mads Torgersen. Understanding type-script. In *European Conference on Object-Oriented Programming*, pages 257–281. Springer, 2014.
- [3] clips. CLIPS: A Tool for Building Expert Systems. <https://www.clipsrules.net/>.
- [4] Ramez Elmasri and Sham Navathe. *Fundamentals of Database Systems*. Pearson/Addison Wesley, Boston, 5th ed edition, 2007.
- [5] David Embrey. The Skill, Rule and Knowledge Based Classification. *Human Error*, page 10.
- [6] Malin Eriksson and Victor Hallberg. Comparison between json and yaml for data serialization. *The School of Computer Science and Engineering Royal Institute of Technology*, pages 1–25, 2011.
- [7] Borko Furht, Esad Akar, and Whitney Angelica Andrews. *Creating User Interface*, pages 7–11. Springer International Publishing, Cham, 2018.
- [8] Mysore Gopinath and Abhijith Athreya. *Automatic Detection of Section Title and Prose Text in HTML Documents Using Unsupervised and Supervised Learning*. PhD thesis, University of Cincinnati, 2018.
- [9] Miguel Grinberg. *Flask web development: developing web applications with python*. O'Reilly Media, Inc.", 2018.
- [10] Kevin C Klement. Propositional logic. 2004.
- [11] Di Liu, Tao Gu, and Jiang-Ping Xue. Rule Engine based on improvement Rete algorithm. In *The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding*, pages 346–349, 2010.
- [12] Chris A. Mattmann and Jukka L. Zitting. *Tika in Action*. Manning Publications, Shelter Island, NY, 2012. Includes index.
- [13] Markus Pikkanen. React and vue performance comparison. 2021.
- [14] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel. *Proceedings of the VLDB Endowment*, 11(3):269–282, nov 2017.

- [15] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.
- [16] Sen Wu. Emmental: A framework for building multimodal multi-task learning systems, 2019.
- [17] Sen Wu, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas, Philip Levis, and Christopher Ré. Fonduer: Knowledge base construction from richly formatted data. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1301–1316. ACM, 2018.
- [18] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, August 2017.
- [19] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

## Online-Quellen

- [20] Bedeutung AW (Arbeitswert). <https://wissen.autoixpert.de/hc/de/articles/360008973819-Bedeutung-AW-Arbeitswert->. [letzter Zugriff: 3. Mai 2022].
- [21] abbyy. PDF Software: Open, Read & Edit PDFs. <https://pdf.abbyy.com/>. [letzter Zugriff: 11. März 2022].
- [22] amazon. Intelligente Extraktion von Text und Daten mit OCR – Amazon Textract – Amazon Web Services. <https://aws.amazon.com/de/textract/>. [letzter Zugriff: 11. März 2022].
- [23] apache. Apache Tika – Apache Tika. <https://tika.apache.org/>. [letzter Zugriff: 12. März 2022].
- [24] atlanhq. Comparison with other PDF Table Extraction libraries and tools · atlanhq/camelot Wiki. <https://github.com/atlanhq/camelot>. [letzter Zugriff: 12. März 2022].
- [25] camelot. Camelot: PDF Table Extraction for Humans. Atlan Technologies Pvt Ltd, July 2022. [letzter Zugriff: 12. März 2022].
- [26] Adobe I/O-Adobe Developers. Extract Text from PDF | Extract Data from PDF | Visualizer - Adobe Developers. <https://developer.adobe.com/document-services/apis/pdf-extract/>. [letzter Zugriff: 12. März 2022].

- [27] Ben Dickson. What is symbolic artificial intelligence? <https://bdtechtalks.com/2019/11/18/what-is-symbolic-artificial-intelligence/>, November 2019. [letzter Zugriff: 20. März 2022].
- [28] Docsumo. PDF Scraper - Scrape data from pdf | PDF data extraction. <https://docsumo.com/blog/extract-data-from-pdf>, June 2022. [letzter Zugriff: 20. März 2022].
- [29] explosion. spaCy · Industrial-strength Natural Language Processing in Python. <https://spacy.io/>. [letzter Zugriff: 4. Juni 2022].
- [30] flask. API — Flask Documentation (2.1.x). <https://flask.palletsprojects.com/en/2.1.x/api/#url-route-registrations>. [letzter Zugriff: 4. März 2022].
- [31] Bruce Frederiksen. Welcome to Pyke. <http://pyke.sourceforge.net/index.html>. [letzter Zugriff: 3. Mai 2022].
- [32] Google. Common Expression Language. <https://chromium.googlesource.com/external/github.com/google/cel-go/+refs/tags/v0.9.0/README.md>. [letzter Zugriff: 7. Mai 2022].
- [33] Paul Haley. Haley / ART syntax lives on in open-source Java rules – Commercial Intelligence, 2008. [letzter Zugriff: 4. Juni 2022].
- [34] jetbrains. IntelliJ IDEA: The Capable & Ergonomic Java IDE by JetBrains. <https://www.jetbrains.com/idea/>. [letzter Zugriff: 7. Juni 2022].
- [35] json. JSON Schema. <https://json-schema.org/>. [letzter Zugriff: 4. Juni 2022].
- [36] lark-parser. Lark-parser/lark: Lark is a parsing toolkit for Python, built with a focus on ergonomics, performance and modularity. <https://github.com/lark-parser/lark>. [letzter Zugriff: 7. Mai 2022].
- [37] measures. Measuresforjustice/texticator: Texticator is a tool to extract text from documents and generate structured data. <https://github.com/measuresforjustice/texticator>. [letzter Zugriff: 20. März 2022].
- [38] microsoft. Visual Studio Code - Code Editing. Redefined. <https://code.visualstudio.com/>. [letzter Zugriff: 7. Mai 2022].
- [39] microsoft. Visual Studio: IDE und Code-Editor für Softwareentwickler und -teams. <https://visualstudio.microsoft.com/de/>. [letzter Zugriff: 7. Juni 2022].
- [40] npm. Npm. <https://www.npmjs.com/>. [letzter Zugriff: 7. Mai 2022].

- [41] omar. Dear ImGui, July 2022. [letzter Zugriff: 14. Mai 2022].
- [42] pdfminer. Welcome to pdfminer.six's documentation! — pdfminer.six \_\_VERSION\_\_ documentation. <https://pdfminersix.readthedocs.io/en/latest/>. [letzter Zugriff: 2. Mai 2022].
- [43] Roberto Abdelkader Martínez Pérez. Nilp0inter/experta, July 2022. [letzter Zugriff: 4. Mai 2022].
- [44] pinia. Pinia. <https://pinia.vuejs.org>. [letzter Zugriff: 4. Juni 2022].
- [45] postcss. PostCSS - a tool for transforming CSS with JavaScript. <http://postcss.org/>. [letzter Zugriff: 4. März 2022].
- [46] pyTorch. PyTorch. <https://www.pytorch.org>. [letzter Zugriff: 4. Juni 2022].
- [47] qt. Qt | Cross-platform software development for embedded & desktop. <https://www.qt.io>. [letzter Zugriff: 14. Mai 2022].
- [48] Jesus Ruiz. Durable\_rules, March 2022. [letzter Zugriff: 13. April 2022].
- [49] Jeremy Singer-Vine. Pdfplumber, July 2022. [letzter Zugriff: 2. Mai 2022].
- [50] stichdata. What is Data Extraction? [ Tools & Techniques ]. <https://www.stichdata.com/what-is-data-extraction/>. [letzter Zugriff: 13. April 2022].
- [51] svelte. Svelte • Cybernetically enhanced web apps. <https://svelte.dev/>. [letzter Zugriff: 14. Mai 2022].
- [52] tabulapdf. Tabulapdf/tabula: Tabula is a tool for liberating data tables trapped inside PDF files. <https://github.com/tabulapdf/tabula>. [letzter Zugriff: 13. April 2022].
- [53] tensorflow. TensorFlow. <https://www.tensorflow.org/?hl=de>. [letzter Zugriff: 14. Mai 2022].
- [54] Laura Theuerer. Was macht das After Sales Management im Autohaus? | Karriere-Ratgeber - kfz-betrieb Jobs. <https://jobs.kfz-betrieb.de/karriere-ratgeber/was-macht-das-after-sales-management-im-autohaus-51.html>. [letzter Zugriff: 14. April 2022].
- [55] unixuser. PDFMiner. <https://www.unixuser.org/~euske/python/pdfminer/>. [letzter Zugriff: 2. Mai 2022].
- [56] Guido van Rossum, Barry Warsaw, and Nick Coghlan. Style guide for Python code. PEP 8, 2001. [letzter Zugriff: 4. Juni 2022].

- [57] vite. Vite. <https://vitejs.dev>. [letzter Zugriff: 7. Mai 2022].
- [58] vue. Composition API FAQ | Vue.js. <https://vuejs.org/guide/introduction.html#api-styles>. [letzter Zugriff: 8. Juni 2022].