Hochschule
für Technik
Stuttgart

Betreutes Praktisches Studienprojekt

Praktikumsbericht

im Rahmen des Studiengangs
Wirtschaftsinformatik
der Hochschule für Technik Stuttgart

vorgelegt von:

Julian Raubald

(Matrikelnummer 1003812)

Praktikumsbetrieb:

iteratec

Betreut durch:

Prof. Dr. Stefan Knauth

Stuttgart, Tag. Februar. 2024

Inhaltsverzeichnis

1 Einleitung

In diesem Bericht schildere ich die Aufgaben und Projekte, welche ich in meinem Praxissemester im Wintersemester 23/24 bei dem Unternehmen iteratec GmbH durchgeführt und begleitet habe. Dabei gehe ich auf die eingesetzten Technologien ein und teile meine gesammelten Erfahrungen und ordne diese inhaltlich im Rahmen meines Studiums der Wirtschaftsinformatik ein.

1.1 Unternehmen iteratec GmbH

Iteratec ist ein Unternehmen, das sich auf digitale Produktinnovation, Software- und Architekturentwicklung sowie digitale Infrastrukturen spezialisiert hat. Als Partner für die digitale Transformation unterstützt es Unternehmen und öffentliche Organisationen bei der Digitalisierung ihrer Prozesse, Produkte und Dienstleistungen. Dies umfasst die Entwicklung von Innovationen, die technische Umsetzung individueller Softwarelösungen, deren Betrieb und Weiterentwicklung sowie Schulungen in agilen Methoden. Das Unternehmen entwickelt mobile Applikationen, Cloud-Architekturen und bietet Application Performance Management an. Es nutzt auch Technologien wie Blockchain, AI und IoT für die Entwicklung neuer Geschäftsmodelle. Zu den Kunden gehören mittelständische Unternehmen, DAX-Konzerne und Organisationen aus verschiedenen Branchen. 1996 in München gegründet, hat iteratec Standorte in Deutschland, Österreich und Polen und beschäftigt rund 500 Mitarbeiter und 120 Studierende. Seit 2019 sind viele Mitarbeiter über eine Genossenschaft am Unternehmen beteiligt.

1.2 Name und Geschichte

Das Unternehmen wurde 1996 gegründet und existiert bereits seit über 25 Jahren. Während zu dieser Zeit das Wasserfallmodell in der Softwareentwicklung weit verbreitet war hat sich iteratec schon früh auf agile Methoden und Werte spezialisiert. Erkennbar ist das insbesondere an dem Namen, welcher sich aus den Begriffen iterativ und Technologie zusammensetzt. Heute hat iteratec bereits über 1000 Softwareprojekte erfolgreich abgeschlossenen.

1.3 Unternehmensstruktur

Standortübergreifend hat iteratec 4 Hauptgeschäftsführer. An jedem Standort gibt es jeweils einen Geschäftsstellenleiter. Jeder Mitarbeiter hat eine persönliche Führungskraft (pFK), welche für die persönliche Entwicklung vor allem im, aber auch außerhalb des Unternehmens zuständig ist. Jedes Projekt hat einen Product Owner (PO), welcher die Interessen des Kunden im Projekt vertritt, und einen Projektleiter, welcher dafür zuständig ist die Fristen und eingesetzten Ressourcen zu managen. Alle Mitarbeiter begegnen sich Standort- und Positionsübergreifend auf Augenhöhe, was sich unter anderem dadurch zeigt, dass sich alle duzen. Die bisherigen Alleingesellschafter der iteratec GmbH haben im Jahr 2018 entschieden, das Unternehmen weder innerhalb ihrer Familie zu vererben noch dieses an externe zu verkaufen. Daraus entstand die iteratec nurdemtean eG, eine Genossenschaft bei der sich die Mitarbeiter am Unternehmen beteiligen, mit dem Ziel, die Unternehmenskultur zu halten.

1.4 Studierende bei iteratec (SLAB)

Die Abteilung für studierende, das SLAB (Studentenlabor), ist Standort übergreifend und schließt sowohl alle Studierenden als auch deren Betreuer mit ein. Halb-jährlich werden neue SLAB Leiter gewählt, deren Aufgabe es ist, alle anfallenden Aufgaben an studierende zu delegieren und im engen Austausch mit dem SLAB-Verantwortlichen zu stehen. Eine Aufgabe ist das Bereitstellen einer Startbegleitung für neue Studierende.

Andere Aufgaben sind z.B. das organisieren von Workshops und Events.

1.5 Meetings und Veranstaltungen

Bei iteratec gibt es abseits der regulären Projektmeetings viele weitere Veranstaltungen an denen man teilnehmen kann. Im folgenden Kapitel erläutere ich die wichtigsten.

1.5.1 SLAB-Meeting

Das SLAB-Meeting findet zwei mal pro Jahr statt und beschränkt sich auf den Standort. Hier wird die SLAB-Leitung gewählt, welche immer aus zwei Personen besteht. Es wird retrospektiv geschaut und bewertet wie gut die Prozesse rund um die Studentenbetreuung funktionieren. Es werden Verbesserungsvorschläge gesammelt und nach Nutzen und Durchführbarkeit bewertet. Anschließend werden entsprechende Maßnahmen an verantwortliche verteilt die sicherstellen, dass diese durchgeführt werden.

1.5.2 GS-Meeting

Das Geschäftsstellen-Meeting für die Geschäftsstelle Stuttgart, an dem alle Mitarbeiter des Standorts eingeladen sind, findet etwa ein Mal pro Monat statt. Hier werden aktuelle Themen besprochen und Mitarbeiter können Ihre Projekte vorstellen. Die Geschäftsführung verkündet während des Meetings wichtige Neuigkeiten, anstehende Events und präsentiert die aktuellen Geschäftszahlen. Neue Mitarbeiter und studierende werden begrüßt und ausscheidende Mitarbeiter verabschiedet. Zum Ende können Kollegen für die Wall of Fame in einer der drei Leitwerte von iteratec (Permanent Exzellent, Konstruktiv unabhängig und Inspirierend mutig) nominieren und so besondere Leistungen und gute zusammenarbeit würdigen.

1.5.3 BUILD23

Bei der BUILD handelt es sich um eine Messe, welche ausschließlich von iteratec Mitarbeitern besucht werden darf. Sie geht über zwei Tage und fand dieses Jahr bei der Geschäftsstelle in München statt. Im Vorfeld kann jeder Mitarbeiter, Student oder jedes Team sich ein Thema überlegen und als Vorschlag einreichen. Nachdem die Themen freigegeben wurden plant man einen entsprechenden Messestand über drei bis sechs Zeitslots, in denen man sein Thema, Projekt oder Workshop präsentieren kann. Hierzu gibt es an jedem Stand Ausrüstung in Form eines TVs, eines Mikrofons und mehreren Receiver mit Kopfhörern. Ziel der Build ist es das erlangte Wissen Projekt- und Standortübergreifend auszutauschen und sich mit Kollegen zu vernetzen oder Interessensgemeinschaften zu bilden. Im Anschluss an die Build findet die Weihnachtsfeier statt.

2 Tätigkeitsbereich und Aufgabenstellung

In diesem Kapitel werde ich genauer auf meinen Arbeitsbereich und meine Aufgaben eingehen. In den darauf folgenden Kapiteln werde ich ein Projekt (TenderInsights) ausführlich beschreiben, und ein anderes nur grob umreissen (TenderSniffer). An beiden Projekten habe ich ungefähr 3 Monate gearbeitet.

2.1 Arbeitsbereich und Ausstattung

Sämtliche Arbeitsplätze sind sogenannte Flexarbeitsplätze. Das heißt, dass man sich morgens einen freien Schreibtisch sucht und dort seine Arbeit beginnt. Es gibt vereinzelt Bereiche wie das SLAB, wo sich bestimmte Gruppen zum arbeiten treffen aber auch dort hat niemand einen festen Arbeitsplatz. Jeder dieser Arbeitsplätze verfügt über einen höhenverstellbaren Schreibtisch und 2 Monitore inklusive Dockingstation für den Laptop. Jedem Mitarbeiter ist es gestattet seine Arbeit von Zuhause aus durchzuführen, solange man mindestens einen Tag in der Woche vor Ort im Büro arbeitet. Ein Großteil der Kommunikation mit Kollegen erfolgt daher über Videokonferenzen oder Chats mithilfe der Software Microsoft Teams. Für Meetings stehen zahlreiche Meeting räume zur Verfügung, welche man sich jederzeit reservieren kann. Für die Arbeit erhält man als Student einen Dell Laptop mit Windows-Betriebssystem.

2.2 Hauptaufgaben

Zu meinen Hauptaufgaben gehört insbesondere das entwickeln von Software, beziehungsweise das schreiben von Quellcode. Eigenständiges recherchieren und Problemlösung gehören dabei fest zum Arbeitsalltag. Entwickelt wird iterativ nach einer abgewandelten Kanban-Variation. So werden Tickets für einzelne Features oder gefundene Bugs in Jira im Backlog angelegt. Die Entwicklung erfolgt in mehreren Sprints, bei denen Aufgaben aus dem Backlog in Wellen entnommen und in die "NewSpalte geschoben werden. Sobald mit der Arbeit an einer Aufgabe begonnen wurde, schiebt man das Ticket in die In ProgressSpalte, bei erledigten Aufgaben kommen diese in die "ReviewSpalte, bei Problemen oder Unklarheit in die "Waiting-Spalte bis eine Entscheidung gefunden wurde wie man fortfahren will. Bei den ein- oder zwei-wöchentlichen Meetings werden die "NewTickets unter den Entwicklern aufgeteilt und mögliche Probleme oder Unklarheiten besprochen. Aufgaben, welche sich in "Review"befinden kommen in die "DoneSpalte falls alles wie gewünscht funktioniert und der Product Owner zufrieden ist. Da es keine festen WIP-Limits (Work in Progress) gibt kann nicht von reinem Kanban gesprochen werden, allerdings ist es eher unüblich dass ein Entwickler mehr als 2 Aufgaben parallel bearbeitet.

2.3 Verwendete Software und Grundlagen

In diesem Kapitel werden alle Software-Werkzeuge, welche während der gesamten Dauer meines Praktikums Projektübergreifend verwendet werden, erläutert.

2.3.1 Werkzeuge

Visual Studio Code

Visual Studio Code (VSCode) ist ein leistungsfähiger und anpassbarer Quelltext-Editor von Microsoft, der eine breite Palette von Programmiersprachen unterstützt und zahlreiche Erweiterungen für Entwickler bietet.

GitLab

GitLab ist eine auf Git basierende webbasierte DevOps-Plattform, welche für die Versionskontrolle und CI/CD bei der Softwareentwicklung verwendet wird. Für die Testdateien und Dokumentation verwenden wir zusätzlich LFS (Large File Storage) um größere Dateien im Repository abzulegen.

Microsoft Teams

Ein Kommunikationstool von Microsoft, welches genutzt wird um sich mit Kollegen auszutauschen, Dateien zu teilen und Besprechungen zu planen so wie durchzuführen.

Jira

Ein Projektmanagement-Werkzeug von Atlassian, das speziell für Softwareentwicklungsteams konzipiert ist und Funktionen für die Verfolgung von Aufgaben, Bugs und agile Prozesse bietet. Es kam insbesondere beim verwalten des Backlogs und beim bearbeiten von Tickets zum Einsatz.

Miro

Eine digitale, kollaborative Whiteboard-Plattform, die es unserem Team ermöglicht, visuell zu brainstormen, zu planen und Projekte in Echtzeit zu gestalten.

3 TenderInsights

3.1 Kurze Projektbeschreibung

In dem Projekt geht es darum einen PoC - also einen Proof of Concept zu erstellen, der mithilfe von AI bzw. Large Language Models (LLM) große Dokumente nach gesuchten Inhalten durchsucht und aufbereitet. Im Fokus stehen dabei insbesondere öffentliche Ausschreibungen zur Akquise von neuen Projekten für das Unternehmen. Bisher hat ein Mitarbeiter aus PMO die Ausschreibungen händisch grob für passend oder nicht geeignet erklärt und dann an die Akquise weitergeleitet. Dort hat ein Mitarbeiter die Ausschreibung sehr genau analysiert und einen sogenannten One Pager erstellt, der alle Daten welche für eine Entscheidung, ob man sich bewerben möchte oder nicht, relevant sind. Aufgabe des PoC ist es nun diesen One Pager mithilfe von künstlicher Intelligenz und einem LLM aus einer oder mehreren bereitgestellten PDF-Datei zu generieren. Der Anwender soll dies über ein Userinterface im Stile einer Webanwendung bedienen können.

3.2 Verwendete Software und Grundlagen

In diesem Kapitel werden alle Technologien und Begriffe beschrieben, welche im Entwicklungsumfeld des TenderInsights verwendet werden oder für ein Verständnis zuträglich sind.

3.2.1 Technologien

Bei der Entwicklung des TenderInsights werden viele unterschiedliche Technologien eingesetzt. Während meines Praktikums habe ich in meiner Tätigkeit als Entwickler viele davon kennengelernt und verwendet. Nachfolgend erkläre ich die wichtigsten Technologien.

Python

Python ist eine höhere Programmiersprache welche sich durch ihre klare und leicht verständliche Syntax auszeichnet. Sie ist bekannt für ihre zahlreichen Standardbibliotheken und ihrer reichen Auswahl an Modulen und Frameworks. Da OpenAI nur JavaScript und Python unterstützt, und Python aus zuvor genannten Gründen einsteigefreundlicher ist wurde sich für diese Sprache entschieden.

Langchain

Langchain ist ein Open-Source-Framework, das auf die Entwicklung und integration von Sprach-KI-Anwendungen spezialisiert ist. Es erleichtert das einbinden von Sprachmodellen wie GPT-3.5-turbo von openAI in Projekte und wird beim TenderInsights für Textgenerierung verwendet.

OpenAI - GPT

OpenAI ist ein US-Amerikanisches Unternehmen welches sich mit der Erforschung und Entwicklung von künstlicher Intelligenz beschäftigt. Eines der bekanntesten Tools bzw. Frameworks ist GPT - Generative Pretrained Transformer, ein Sprachverarbeitungsmodell zur Textgenerierung, Übersetzung, Zusammenfassung und mehr. GPT wird im TenderInsights verwendet um Information aus Texten zu extrahieren und um die gelieferten Prompts zu bewerten.

Azure

Microsoft Azure ist eine Cloud-Computing-Plattform, welche eine breite Palette von Diensten und Lösungen für Unternehmen und Entwickler bietet. Zu den Entwickler-Tools und -Diensten gehört unter anderem Azure openAI, welches eine openAI API zur Verfügung stellt. Im Gegensatz zu openAI befinden sich die Server in Europa, weshalb im späteren Verlauf des Projekts aus Gründen des Datenschutzes zu Azure gewechselt wurde.

ChromaDB

ChromaDB ist eine Open-Source-Datenbanklösung, welche speziell für die Arbeit mit Vektor-Einbettungen in KI-Anwendungen entwickelt wurde. Da der Kontext für Sprachmodelle stark unter der Größe von durchschnittlichen Dokumenten liegt wird eine Vektor-Datenbank genutzt um die relevanten Stellen in den Dokumenten zu finden und als Kontext mit dem Prompt an das Sprachmodell für die Textgenerierung zu übergeben.

JSON

JavaScript Object Notation, oder kurz JSON, ist ein Daten-Austauschformat welches häufig verendet wird um Daten zwischen einem Server und einer Webanwendung zu übertragen. Da es für Menschen gut lesbar und für Maschinen einfach zu parsen (Entspricht einem normalen Dictionary in Python) ist eignet es sich hervorragend für das Speichern der OnePager Informationen.

3.2.2 Wichtige Begriffe

Tabelle 3.1: Wichtige Begriffe

Name	**Beschreibung**
Prompt	Anleitung oder Frage die dem Sprachmodell vorgibt, was es tun soll. Auch Eingabeaufforderung genannt.
Prompt Engineering	Gestalten und Optimieren von Eingabeaufforderungen (Prompts)
Token	Grundlegender Baustein für die Verarbeitung und das Verständnis von Sprache in KI-Systemen
Embedding	Einbetten von Text in Vektoren
Vektordatenbank	Datenbank welche die aus dem Embedding generierten Vektoren enthält
Similarity Search	Suche von nahestehenden Vektoren innerhalb der Vektordatenbank um gesuchte Dokumenteninhalte zu finden
Large Language Model (LLM)	Fortgeschrittene KI, die umfangreiche Textdaten versteht und generiert, basierend auf umfassendem Training und komplexen Algorithmen

3.3 Projektstand bei Arbeitsbeginn

Zum Zeitpunkt meines Projekteintrittes gibt es bereits ein Frontend welches mithilfe von Streamlit implementiert wurde. PDF-Dateien können über einen File-Uploader im Frontend an die Anwendung übergeben werden. Da die meisten LLM-Modelle eine begrenzte Token-Anzahl für die Abfragen haben ist es nicht möglich das gesamte Dokument an das LLM zu übergeben. Daher wird das Dokument aufgesplittet und mittels einer Vektordatenbank auf Ähnlichkeit mit die gesuchte Information abgeglichen. Die k Ergebnisse mit der größten Ähnlichkeit werden zusammen mit dem Prompt als Abfrage an das LLM übermittelt. Die Abfragen werden mit der Langchain API durchgführt, das verwendete Modell ist "gpt-3.5 turbo". Es gibt 3 Ausschreibungsdokumente in verschiedenen Größen mit denen die Abfragen getestet

werden können. Es gibt 3 Prompts, welche bislang nur bedingt gute Ergebnisse liefern.

3.4 Softwarearchitektur der Anwendung zum Zeitpunkt der BUILD23

Um nachfolgende Kapitel besser verstehen zu können erkläre ich in diesem Kapitel die Grobe Software-Architektur der Anwendung zum Zeitpunkt der BUILD23 anhand der Abbildung 3.1.

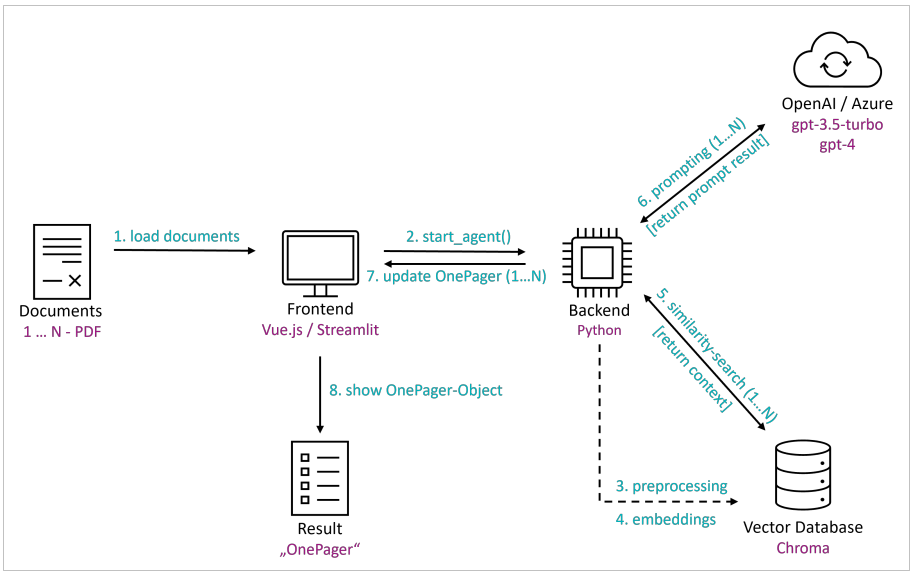

Abbildung 3.1: Schematische Darstellung der Software-Architektur

Zuerst werden die Ausschreibungsdokumente im Frontend hochgeladen (Siehe 1.) und der Agent wird mit den gewünschten Prompts gestartet (Siehe 2.). Der Agent beginnt nun im Backend mit dem PreProcessing, welches die Dokumente in fachlich logische Blöcke aufspaltet und personenbezogene Daten anonymisiert (Siehe 3.). Anschließend wird das Embedding, also eine Repräsentation der Daten als Vektoren, erstellt und als der Vektordatenbank angelegt (Siehe 4.). Nun wird für jeden Prompt eine Similarity-Search durchgeführt (Siehe 5.), bei der dem möglichst passende Stellen aus den Dokumenten aus der Vektordatenbank extrahiert werden, damit diese im Prompt als

Kontext eingebettet für die Textvervollständigung an das LLM übergeben werden (Siehe 6.). Das Ergebnis wird in ein JSON-Objekt übersetzt und der OnePager wird entsprechend aktualisiert (Siehe 7.). Währenddessen baut sich im Frontend Stück für Stück der OnePager mit allen gefundenen Informationen zusammen (Siehe 8.).

3.5 Meine Rolle und Aufgaben im Projekt

Ich bin in der Rolle eines Softwareentwickler in das Projekt gekommen. Meine Aufgabe bestand Anfangs darin, die Prompts für die einzelnen Felder des One Pagers zu formulieren um die gelieferten Ergebnisse zu verbessern. Dies resultierte in den Aufbau einer Test- und Evaluierungsarchitektur in Python mithilfe derer die Prompts vollautomatisch bewertet werden und anhand ihrer Kennzahl ausgewählt und weiter optimiert werden. Durch Komplikationen mit dem Datenschutz wurde der Umstieg von Langchain auf Azure von Microsoft beschlossen, welchen ich durchgeführt habe.

3.5.1 Prompt Engineering

Um gute Ergebnisse durch die Textvervollständigung der Sprachmodelle zu erhalten ist es unumgänglich hochwertige Prompts zu formulieren. Hierfür habe ich mehrere E-Learnings zum Thema Prompt Engineering absolviert, in denen erklärt wird wie man hochwertige Prompts erstellt und ganze Systeme mit einer Chatbotintegration entwickelt. Besondere Methoden welche im Projekt zum Einsatz kommen sind Single-Shot-Prompting, bei dem man ein Beispiel gibt wie eine optimale Antwort auszusehen hat, und die Ausgabe als JSON festzulegen, um die erhaltenen Ergebnisse im Code nutzen und anwenden zu können. Das Abgrenzen von wichtigen Informationen wie dem mitgelieferten Kontext über sogenannte Delimiter hilft dem Modell dabei, Aufgabe und Kontext nicht zu vermischen (wenn auch nicht immer perfekt). Gut eignet sich hierfür '####', da dies als ein Token zählt und selten regulär vorkommt. Um Halluzinationen, also das produzieren von Falschinformationen oder ausgedachten Informationen zu vermeiden wird im Prompt explizit darauf hingewiesen, dass bei nicht vorhandenen Informationen das entsprechende Feld leer gelassen werden soll.

3.5.2 OnePager

Der von PMO erstellte OnePager umfasst alle wesentlichen Information der Ausschreibung um eine fundierte Entscheidung darüber treffen zu können, ob die Ausschreibung als Projekt für iteratec geeignet ist. Informationen sind unter anderem der Projekttitel, eine kurze Projektbeschreibung, Adressdaten über das Ausschreibende Unternehmen oder Amt, wichtige Termine und Deadlines, Bewertungskriterien, Aufgaben und benötigte Unterlagen welche dem Angebot beigefügt werden müssen. Da jede dieser Punkte in unterschiedlichen Regionen der Dokumente zu finden ist und die Qualität der Ergebnisse besser ist wenn man nicht mehrere Fragen in einer Anfrage bündelt wird der OnePager in der Anwendung Frage für Frage zusammengesetzt und ergänzt. Als Format wird JSON verwendet um später den fertigen OnePager gut speichern und weiterverarbeiten zu können (zum Beispiel für die Evaluation) und um dem Frontend die Daten als Variable zur Verfügung zu stellen. Die OnePager-Klasse wurde wurde um viele Hilfsmethoden ausgestattet um das Arbeiten mit den JSON-Objekten zu erleichtert. Zu den Hilfsmethoden gehören unter anderem das initialisieren des OnePager-Template, das laden und abspeichern als Datei, einfaches hinzufügen von Teilinformationen aus den einzelnen Ergebnissen des Sprachmodells in den bestehenden OnePager und Prüffunktionen um nicht JSON-Konforme Antworten zu erkennen.

3.5.3 Verbessern der Kontextauswahl über Similarity Search

Um gezielt die fachlich richtigen Passagen innerhalb des Dokuments zu finden wird eine query gegen die über das Embedding generierte Vektordatenbank gestellt. Als Ergebnis erhält man die n-Textausschnitte, mit der höchsten Trefferquote.

```
1    results = collection.query(
2        query_texts=["This is a query document"],
3        n_results=2)
```

Listing 3.1: Code einer Similarity Search

Bei dem eben beschriebenen Verfahren sprechen wir von einer Similarity Search, also dem finden von ähnlichen Daten innerhalb einer Vektorrepräsentation des

bereitgestellten Dokuments. Da es bei Ausschreibungen viele synonyme Begriffe gibt, welche Informationen zu identischen Fragestellungen geben habe ich diese Begriffe in einem Lexikon erfasst und gesammelt. Die Akquiseabteilung wurde zusätzlich darum gebeten weitere Ergänzungen vorzunehmen um fachlich möglichst viele Begriffe abzudecken. Mit dem so erhaltenen Lexikon können wichtige Keywords für die Similarity Search bereitgestellt werden, vereinzelt wird auch innerhalb der Prompts darauf zurückgegriffen um das Sprachmodell dabei zu unterstützen die richtigen Informationen der richtigen Stelle des OnePagers zuzuordnen. Über ein Prompt.Config Dictionary kann nun für jedes Prompt auf entsprechende Felder wie Suchbegriffe, Promptmethode oder Suchergebnisanzahl individuell zugegriffen werden.

3.5.4 Evaluation

Um festzustellen ob sich Prompts beim Wechsel von Modellen, Promptversionen oder unterschiedlichen PreProcessing-Verfahren verbessern oder verschlechtern habe ich 10 Testdatensätze aus öffentlichen Ausschreibungsportalen zusammmengesucht und entsprechende Musterlösungen erarbeitet, sofern dies in unserem fachlichen Rahmen möglich war. Dies hat uns ermöglicht, ausgewählte Prompts testweise gegen alle 10 Testdatensätze abzufragen. Hierbei wird die neu generierte Antwort mit der Musterlösung verglichen und anschließend anhand eines Bewertungsschemas mit Punkten zwischen 0 und 10 inklusive einer Begründung bewertet, je nachdem wie nah das gelieferte Ergebnis an die Musterlösung herankommt.

Die einzelnen Punkte werden anschließend für jedes Prompt zu einem Average Score zusammengerechnet und mithilfe der Visualisierungsklasse als Plot dargestellt. Dabei können einzelne Prompts über verschiedene Versionen und Zeitpunkte (siehe 3.2) oder mehrere Prompts bei Verwendung unterschiedlicher Sprachmodelle (siehe 3.3) dargestellt und verglichen werden.

3.5.5 Azure Migration

Da die rechtliche Lage bei der Verarbeitung Personenbezogener Daten durch Sprachmodelle wie gpt laut unserer Datenschutzbeauftragten noch sehr unklar ist bis es

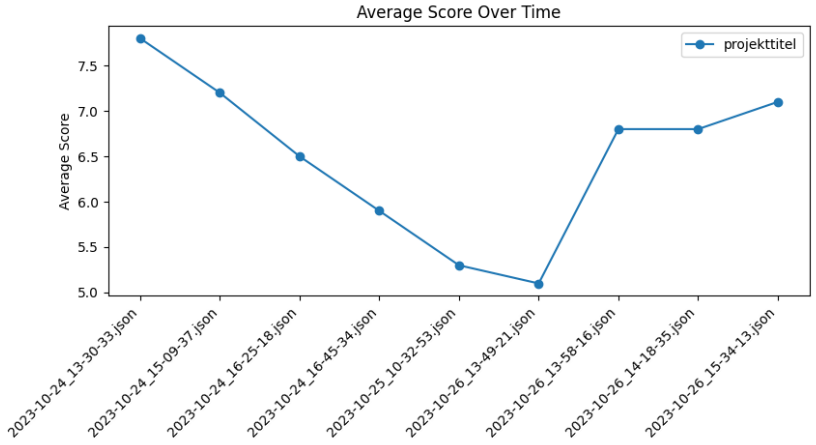

Abbildung 3.2: Visualisierung eines Prompts mit verschiedenen Prompt und Bewertungsmodifikationen über Zeit

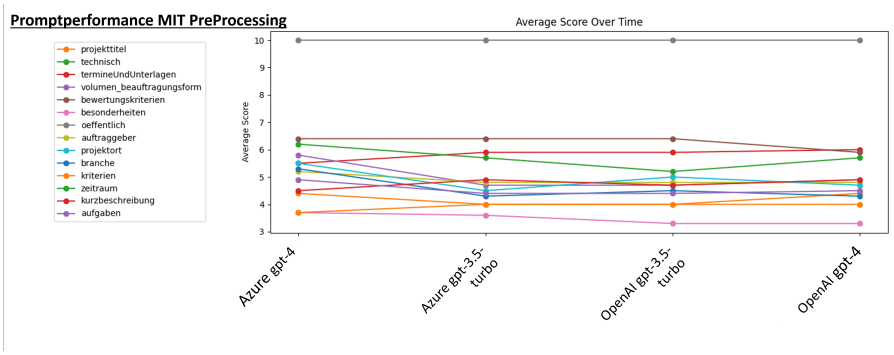

Abbildung 3.3: Visualisierung der Prompts mithilfe unterschiedlicher Modelle. Pre-Processing ist aktiviert

erste Urteile geben wird, wurde sich darauf geeinigt auf azure openAI zu migrieren da dort europäische Server verwendet werden statt amerikanischer. Ein großer Nachteil von Azure ist aber, dass man für das erstellen der Embeddings für die Vektordatenbank auf 16 Inputs beschränkt ist, während openAI hier keine Einschränkungen vorgibt. Größere Dokumente können leicht auf 1 Input pro Seite kommen, was ein Einbetten von Dokumenten ab 20 Seiten verhindert. Um das Problem zu lösen habe ich statt die vorgegebene Embedding-Methode zu verwenden eine eigene Methode erstellt, welche die Dokumente bzw. die Inputs auf mehrere Listen mit einer Größe kleiner gleich 16 aufteilt. Anschließend werden die einzelnen Listen über eine REST API zu embeddings umgewandelt und wieder zusammengesetzt. Am ende wird eine

Collection mit den Embeddings, den Dokumenten und entsprechenden Metadaten erstellt, welche zentral in einem VectorDataManager zur Verfügung gestellt wird. Gegen diese Collection können anschließend Abfragen wie Similarity-Searches (siehe 3.5.3) durchgeführt werden.

3.5.6 Dokumentation

Sämtlicher von mir geschriebener Quellcode wurde mit Ein- und Mehrzeiligen Kommentaren ergänzt, um das Lesen und Verstehen des Codes zu vereinfachen. Ich habe die gesammelten Erfahrungen in einem Bericht festgehalten und diesen im Repository abgelegt. Zudem wurden Architecture Decision Records, kurz ADR, erstellt und ebenfalls dem Projekt angehängt. Diese ADRs sind dazu da, später Einsicht darüber zu geben warum sich für bzw. gegen ein Framework oder eine Technologie entschieden wurde.

3.5.7 Messestand auf der BUILD23

Zum Abschluss des Projektes wurden wir angefragt, ob wir aufgrund des hohen Interesses an dem Projekt eine Präsentation während des GS-Meetings und der BUILD23 halten möchten. Das erstellen der Präsentation und Vortragen dieser gehörte auch zu meinem Aufgabenbereich.

3.6 Herausforderungen und Lösungsansätze

3.6.1 Optimierung der Prompts

Um die Prompts zu verbessern ist es notwendig, diese an unterschiedlichen Ausschreibungen zu testen und das gelieferte Ergebnis mit der richtigen Antwort zu vergleichen. Da das größte Dokument über 100 Seiten hat und wir keine Musterlösungen haben war es schwierig Aussagen über die Qualität der Prompts zu treffen. Die Lösung war das Einführen einer Metrik welche die Qualität der Prompts misst. Es wurden 10

Dokumente ausgearbeitet und sämtliche wichtige Informationen in OnePager-JSON Dateien gespeichert. Anschließend wurde eine Testarchitektur geschaffen, in welcher man die gewünschten Prompts vollautomatisiert gegen die 10 Testdokumente abfragt und anschließend die Resultate zusammen mit der Musterlösung von ChatGPT auf inhaltliche Übereinstimmung überprüfen und bewerten lässt.

[SIEHE ANHANG XXX - Evaluationsprompt]

Aus den so generierten Bewertungen lassen sich nun Aussagekräftige Kennzahlen generieren. Zur Darstellung wurde eine Visualisierungsklasse geschrieben, welche die Punktzahl der Prompts graphisch darstellt (siehe 3.4) und so schnell erkennbar ist, ob der aktuelle Prompt zu einer Verbesserung oder Verschlechterung der Ergebnisse führt.

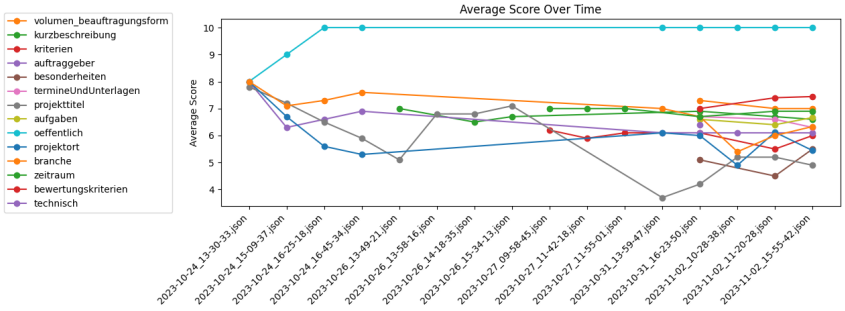

Abbildung 3.4: Visualisierung der durchschnittlichen Promptqualität im Laufe der Zeit

3.6.2 Datenschutz

Das übermitteln und verarbeiten von Personenbezogenen Daten ist in Deutschland nur mit Ausdrücklicher Genehmigung der entsprechenden Person zulässig. Laut unserer Datenschutzbeauftragten ist unklar inwiefern das Übermitteln der Daten aus den Dokumenten über die OpenAI API hierunter einzustufen ist. Offiziell heißt es, dass die Daten, welche über die API geteilt werden, nicht zu Trainingszwecken genutzt werden. Da die Server von OpenAI allerdings in den USA liegen gestalten sich Probleme mit europäischen Datenschutzrecht. Als Lösung wurde die Umstellung auf Azure in Betracht gezogen, da die Server auf europäischen Boden stehen und

damit zumindest nach europäischen Recht Datenschutzkonform sind. Zusätzlich anonymisieren wir so gut wie alle Personenbezogenen Daten bevor wir diese an das Sprachmodell übermitteln, indem wir den Kontext vor dem Abschicken gegen eine Liste an Namen prüfen, und so Namen und Mailadressen herausfiltern.

3.7 Ergebnisse und Erfahrungen

3.7.1 Proof of Concept

Aus den gesammelten Ergebnissen und Erfahrungen lässt sich herleiten, dass der PoC definitiv zeigt, dass die Anwendung so möglich ist. Auch wenn nicht immer alle OnePager korrekt generiert wurden, und teils immer noch Halluzinationen vorkommen und wichtige Informationen fehlen so funktioniert die Anwendung im Kern so wie sie es soll. Wichtig für die Zukunft ist es Klarheit beim Umgang mit Personenbezogenen Daten im Zusammenhang mit Sprachmodellen zu schaffen und gegebenenfalls eine geeignete Filterung dieser Daten zu implementieren. Durch neue Modelle, wie dem gegen Ende des Projekts angekündigten gpt-4-turbo, welches eine Kontextlänge von 128K Tokens besitzt, wird man nicht länger gezwungen sein sich auf kleine Textausschnitte zu begrenzen. Man kann ganze Kapitel übergeben und Limitierungen haben mehr wirtschaftliche Gründe als Technische. Auch die Fehleranfälligkeit wird durch einen eingebauten JSON-Parser weiter sinken. Unsere Tests zeigen keine Nennenswerte Unterschiede zwischen den Modellen gpt-3.5-turbo und gpt-4 was die Qualität der Prompts angeht, daher ist es unwahrscheinlich, dass das neue Modell hier tatsächlich einen Unterschied macht.

3.7.2 Persönlich

Persönlich konnte ich in dem Projekt lernen, wie man mithilfe eines Jira-Boards als Gruppe iterativ Software entwickelt. Ich habe mit Python eine neue Programmiersprache gelernt, welche vielseitig einsetzbar ist und vor allem im Bereich der Künstlichen Intelligenz und Machine Learning viele Möglichkeiten bietet. Zudem

konnte ich ein tieferes Verständnis für Prompts und deren Einsatzmöglichkeiten entwickeln. Das entwickeln einer ganze Applikation, welche Sprachmodelle oder andere LLMs integriert ist für mich nun verständlich und durchführbar. Ich konnte auch Erfahrungen im Dokumentieren von Entwicklungsartefakten und Entscheidungen sammeln, welche sich in zukünftigen Projekten weiter vertiefen lassen.

4 TenderSniffer

Dieses Kapitel nur kurz.

4.1 Kurze Projektbeschreibung

TenderSniffer ist ein Projekt, welches aus mehreren Applikationen besteht und zum Ziel hat verschiedene Ausschreibungsportale nach neuen Ausschreibungen zu durchsuchen und die gefundenen Ausschreibungen visuell strukturiert darzustellen. Teilanwendungen hierbei sind der TenderCrawler, welcher die Ausschreibungen von den einzelnen Plattformen zieht und in die Datenbank speichert. Der TenderWeb ist ein Graphisches User Interface (GUI), welches die Ausschreibungen in der Datenbank darstellt und Benutzereingaben abspeichert. Mithilfe von verschiedenen Schaltflächen kann die Ausschreibung entweder an die Akquiseabteilung weitergeleitet werden oder für unpassend deklarieren.

4.2 Verwendete Software und Grundlagen

In diesem Kapitel werden alle SW-Werkzeuge und Technologien beschrieben, welche im Entwicklungsumfeld des TenderSniffers verwendet werden oder für ein Verständnis zuträglich sind.

4.2.1 Werkzeuge

IntelliJ IDEA

Integrierte Entwicklungsumgebung des Softwareunternehmens JetBrains. Wird für die Entwicklung mit Java und Kotlin eingesetzt.

PostgreSQL

Freies Datenbankmanagementsystem, welches seit 1997 von einer Open-Source-Community weiterentwickelt wird. Es wird oft mit Postgres abgekürzt.

SonarQube

Plattform für statische Analyse und Bewertung von Quelltext. Die Ergebnisse der Analyse werden über eine Website dargestellt.

Lint

Werkzeug, welches den Quellcode auf Fehler, unübliche Muster, nicht eingehaltene Stilrichtlinien und potenzielle Bugs überprüfen, um eine bessere Codequalität und -konsistenz zu gewährleisten.

4.2.2 Technologien

Java

Java ist eine objektorientierte höhere Programmiersprache welche 2010 von Oracle übernommen wurde. Sie findet neben Computerapplikationen auch Einsatz bei Apps für Smartphones, Tablets und Spielekonsolen.

JavaScript

JavaScript ist eine vielseitige und weit verbreitete Programmiersprache, die hauptsächlich für die Entwicklung von interaktiven Webseiten und Webanwendungen eingesetzt wird.

Spring Boot

Spring Boot ist ein Java-Framework zur Vereinfachung der Entwicklung und Bereitstellung von Spring-basierten Anwendungen, das auf Konvention statt Konfiguration setzt und zahlreiche Funktionen für einen schnellen Projektstart bietet.

Vue.js

Vue.js ist ein JavaScript-Framework, das für den Aufbau interaktiver Web-Oberflächen und Single-Page-Applikationen verwendet wird.

Jest

Jest ist ein JavaScript-Test-Framework, das für seine einfache Konfiguration und effiziente Leistung bei der Entwicklung von Webanwendungen bekannt ist. Jest ist ein JavaScript-Test-Framework, das häufig für das Testen von React- und Vue.js-Anwendungen eingesetzt wird, wobei es Funktionen wie einfache Konfiguration und integrierte Mocking-Unterstützung bietet.

4.3 Meine Rolle und Aufgaben im Projekt

Ich bin in der Rolle eines Softwareentwickler zum Projekt hinzugestoßen, es wird agil mithilfe von Jiraboards und Sprints entwickelt. Ein Sprint geht über vier Wochen, da ein großteil der Projektbeteiligten Werkstudenten sind und entsprechend nur ein bis zwei Tage pro Woche arbeiten. Die Endnutzer (PMO- und Akquiseabteilungen)

erstellen eigenständig neue tickets und beschreiben darin aus Benutzersicht welche Funktionen, Änderungen und Fehler gewünscht sind, bzw. beseitigt werden sollen. Neben den täglichen Austauschterminen, den sogenannten "Dailies"gibt es jeden Freitag einen größeren Besprechungstermin, die "Weeklies". Am ende eines jeden Sprints findet ein "Review", bevor der nächste Sprint begonnen wird.

4.3.1 Kommentare ausblenden

Meine Einstiegsaufgabe war es in der Kommentarsektion einen Button im Frontend des TenderWeb zu programmieren, mit dem der Kommentarbereich eingeklappt werden kann sobald dieser mehr als 3 Kommentare enthält. Hierfür habe ich die vorhandene commentBox.vue um den besagten button erweitert und mithilfe von javascript die Logik hinter dem button implementiert. Nach mehreren Schleifen mit dem Team und den Nutzern habe ich für die finale Version tests mithilfe von jest geschrieben. Im Anschluss habe ich den Merge Request erstellt und nachdem die DevOps Pipeline erfolgreich durchgelaufen ist und die Funktion auf dem Testbuild einwandfrei funktioniert hat wurde das Feature auf der Produktionsumgebung ausgeliefert und das Ticket abgeschlossen.

4.3.2 Auftraggeberdetails erfassen

Die zweite und umfangreichste Aufgabe war es, die Daten der Auftraggeber in einer eigenen Box anzuzeigen und editierbar zu gestalten, so dass Änderungen konsistent sind. Hierfür habe ich das Datenbankschema um eine client-tabelle erweitert und im Backend entsprechende domain-, repository-, service- und Controllerklassen geschrieben, mit denen über HTTP Requests die API angesprochen werden kann und Auftaggeber sowohl abgerufen als auch gespeichert werden können. Im Frontend habe ich den store, welcher als zentraler Ort zur Speicherung und Verwaltung des Anwendungszustands, was eine konsistente Datenverwaltung und -verteilung über die gesamte Anwendung hinweg ermöglicht, um entsprechende Methoden erweitert. Anschließend habe ich eine clientBox mit vue erstellt, welche diese Daten aus dem store anzeigt und bei Änderungen mithilfe von store Mutationen und Aktionen

die neuen Daten über eine POST Request an das Backend übermittelt und somit konsistent in der Datenbank abgespeichert werden.

5 Fazit und Ausblick

5.1 Zusammenfassung der wichtigsten Erkenntnisse

5.2 Einordnung der erlernten Inhalte und Fähigkeiten im Kontext deines Studiums

Allg. Programmieren 1 und 2 -> Java

PERSPEKTIVWECHSEL?

SW-Entw. nach Srum -> Tickets als UserStorie oder Epics geschrieben -> PO -> Sprints -> Spalten wie in Kanban? CI/CD Pipeline

Git

5.3 Weiterer Werdegang

Wie wird sich das Praxissemester auf deinen weiteren Werdegang auswirken?

6 Anhang

Eventuell Codebeispiele, Diagramme, Screenshots oder sonstige unterstützende Materialien, die im Hauptteil zu umfangreich wären

6.1 Liste der verwendeten Tools und Technologien

6.2 Literaturverzeichnis

https://www.deeplearning.ai/short-courses/building-systems-with-chatgpt/

7 Eidesstattliche Erklärung

Bestätigung, dass du den Bericht selbstständig verfasst hast (sofern von deiner Hochschule gefordert).

8 Zusammenfassung und Ausblick

Im letzten Kapitel sollte die Arbeit zusammengefasst und ein Fazit gezogen werden. Außerdem sollte beschrieben werden, wie es mit dem Projekt weitergehen kann und welche Punkte vielleicht interessant wären aber im Rahmen der Arbeit nicht bearbeitet werden konnten.

A Anhang

Der Anhang kann Teile der Arbeit enthalten, die im Hauptteil zu weit führen würden, aber trotzdem für manche Leser interessant sein könnten. Das können z. B. die Ergebnisse weiterer Messungen sein, die im Hauptteil nicht betrachtet werden aber trotzdem durchgeführt wurden. Es ist ebenfalls möglich längere Codeabschnitte anzuhängen. Jedoch sollte der Anhang kein Ersatz für ein Repository sein und nicht einfach den gesamten Code enthalten.

B Abbildungsverzeichnis

C Tabellenverzeichnis

D Listings

Erklärung

Ich versichere an Eides statt, die vorliegende Arbeit selbstständig verfasst und nur die angegebenen Quellen benutzt zu haben.

Unterschrift

Lübeck, Tagesdatum