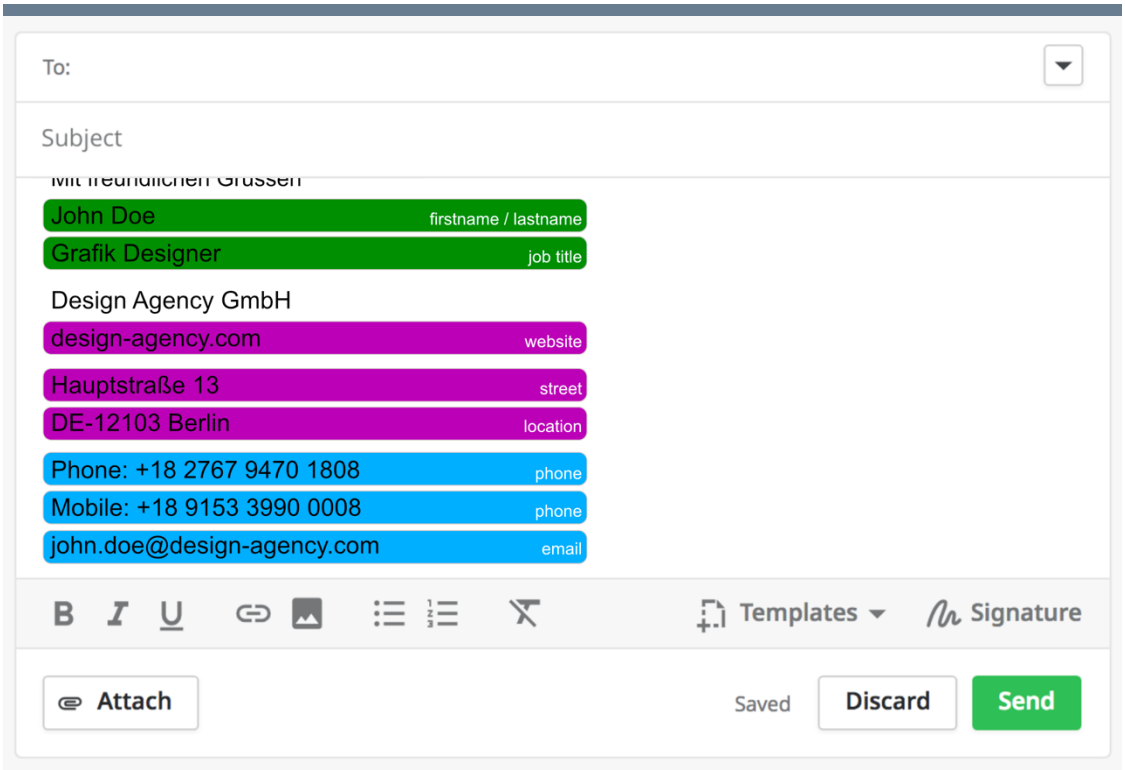

Abschlussbericht

Automatisierte Pflege von Kundenkontaktangaben

Author: Tobias Weissert
Betreuer: Prof. Dr. Jürgen Vogel
Experte: Xavier Monnat

Berner Fachhochschule
Departement Technik und Informatik
BSc in Informatik | Vertiefung : Data Engineering

Erklärung der Diplomandinnen und Diplomanden

Selbständige Arbeit

Ich bestätige mit meiner Unterschrift, dass ich meine vorliegende Bachelor-Thesisselbständig durch geführt habe. Alle Informationsquellen (Fachliteratur, Besprechungen mit Fachleuten, usw.) und anderen Hilfsmittel, die wesentlich zu meiner Arbeit beigetragen haben, sind in meinem Arbeitsbericht im Anhang vollständig aufgeführt. Sämtliche Inhalte, die nicht von mir stammen, sind mit dem genauen Hinweis auf ihre Quelle gekennzeichnet.

Name, Vorname Weissert Tobias.....

Ort, Datum 11.1.2021 .....

Unterschrift 
.....

Inhaltsverzeichnis

Abstract		2
	Ausgangslage	2
	Ergebnisse	2
	Ausblick	2
1	Einleitung	3
2	Ähnliche Projekte	3
	2.1 Apple Mail	3
	2.2 Contacts+	3
	2.3 SigParser	3
3	Material und Methoden	3
	3.1 Material	3
	3.2 Vorarbeit	4
	3.3 Methodik	4
4	Ergebnisse	5
	4.1 Projektverlauf	5
	4.2 Namens Erkennung	6
	4.3 Jobtitel Erkennung	6
	4.3.1 Adresserkennung	6
	4.3.2 Telefonerkennung	7
	4.3.3 Webcrawling	**Error! Bookmark not defined.**
5	Diskussion	8
	5.1 Rechtliches	8
	5.1.1 Herunterladen von personenbezogenen Daten	8
	5.1.2 Speichern von personenbezogenen Daten	8
	5.2 Wirtschaftliches	9
6	Folgerungen	10
Glossar		11
Tabellenverzeichnis		11
Abbildungsverzeichnis		11

Abstract

Ausgangslage

Im heutig schnelllebenden Arbeitsmarkt wechseln Arbeitnehmer innert weniger Jahre ihren Arbeitgeber oft mehrfach. Das CRM (Customer Relation Management) System von Lieferanten oder Partnern hat dadurch schnell falsche oder veraltete Daten. Dies führt zu Mehraufwand beim Kontaktieren der verantwortlichen Person sowie Mehrkosten durch unzustellbaren Postversand. Ausserdem wird in CRM Systemen jeweils die Kontaktperson gepflegt, die im Entscheidungsprozess bei Neuanschaffungen involvierten Personen sind dem Lieferanten oft gar nicht bekannt. Ziel dieser Arbeit ist die Abklärung zur technischen Machbarkeit eines Tools zur Lösung dieses Problems.

Ergebnisse

Das Ziel wurde erreicht und es konnte ein Proofofconcept für ein Tool zur automatischen Stammdatenpflege erstellt werden. Das Tool wurde gegen ein Datenset mit 9860 Kontakten getestet und dabei sind folgende Ergebnisse entstanden:

Analyse E-mail-Signatur

Von 93% der E-Mails konnte die Information in der Signatur richtig erkannt und richtig gelabeld werden. Auch wenn nicht alle Informationen in der Signatur erkannt wurden, konnten die wichtigsten Attribute wie Name, Telefon, E-Mail und Adresse extrahiert werden und können nun in einem weiteren Schritt gegenüber dem aktuellen Stand im CRM verglichen werden. Allerdings bestand das Datenset vor allem aus europäischen Kontakten. Was nicht getestet wurde, sind E-Mails aus anderen Ländern, welche vermutlich ein etwas anderes Format aufweisen: Anderes Postleitzahlsystem oder andere Labels für Telefon oder Adresse.

Analyse der Firmenwebseite

Obwohl nicht jede Webseite gleich aufgebaut ist, konnten bei 86% der Kontakte mittels Webcrawling zusätzliche

Informationen gewonnen werden, falls diese Person auf der Webseite erwähnt wird. Hierzu wird nach Elementen in optischer Nähe zum Namen gesucht. Während der Entwicklung hat sich aber herausgestellt, dass besonders bei grösseren Unternehmen oft nur die Geschäftsleitung auf der Webseite abgebildet ist. So konnte nur bei 36% der Personen im Testset wirklich einen Eintrag auf der Firmenwebseite gefunden werden. Oftmals waren sogar noch weniger Informationen verfügbar als in der E-Mail-Signatur, einzig interessante weitere Information war oft nur ein Foto der Person.

Ausblick

Die entwickelte API zur Analyse der E-Mail-Signatur kann nun in unser firmeninternes CRM System eingebunden werden. So werden automatisch eingehende E-Mails nach neuen oder abweichenden Kontaktangaben durchsucht und entsprechend angepasst. Die Methode, mittels Webcrawling, die Kontakte weiter zu pflegen, benötigt allerdings noch weitere Aufmerksamkeit.

1 Einleitung

Die Idee für diese Arbeit ist kurz vor Weihnachten bei meinem Arbeitgeber entstanden. Die Idee war es damals, nach Rolle personalisierte Weihnachtskarten zu versenden. Heisst, der CEO erhält auf seiner Karte eine andere Botschaft wie z.B. der CFO. Um dies zu ermöglichen wurden circa 10'000 Kontakte manuell klassifiziert und auf deren aktualität geprüft. Dabei flossen mehrere Mannwochen in diese Aufgabe. Trotz sorgfältiger prüfung kamen circa 5% der Weihnachtskarten unzustellbar zurück, einige weitere konnten zwar zugestellt werden, wir wurden aber informiert, dass die entsprechende Person nicht mehr im Unternehmen tätig ist. Dies löste den Gedanken aus, ob es nicht möglich wäre, diese Aufgabe zu automatisieren um so das ganze Jahr über den Stand des CRMs aktuell zu halten und Aus- und Neueintritte frühzeitig zu erkennen. Daraus entstand das Konzept für diese Arbeit. Diese Arbeit isst ein Proof-of-Concept ob es technisch möglich ist, diese Arbeit zu automatisieren und so die manuelle Pflege zu vereinfachen und somit kosten zu sparen.

2 Ähnliche Projekte

2.1 Apple Mail

Apple Mail erlaubt es, die Informationen in der Signatur direkt in das Adressbuch zu importieren. Allerdings beschränkt es sich dabei auf das lokale Adressbuch und dieser Schritt muss vom User jeweils manuel vorgenommen werden.

2.2 Contacts+

Contacts+ bietet einen Assistenten, welcher immer wider nach updates für die gegebenen Kontakte sucht. Woher die Daten allerdings genau stammen ist leider unklar. Dabei sind sicher Twitter, LinkedIn sowie BounceListen von diversen E-Maildiensten.

2.3 SigParser

Dieses Tool erlaubt es E-Mails und deren Signaturen zu parsen und direkt in ein CRM System zu importieren. Dieses Projekt wurde erst kurz vor der fertigstellung der Arbeit entdeckt und macht was das parsen der Signatur angeht etwas sehr ähnliches wie diese Arbeit.

3 Material und Methoden

3.1 Material

Mein Arbeitgeber hat freundlicherweise die Daten zur Evaluation dieser Arbeit zur Verfügung gestellt. Es handelt sich dabei um knapp 10'000 Kontakte. Und zusätzlich über 100'000 Emails, davon wurden 1496 kontrolliert und klassifiziert und schlussendlich zur Kontrolle verwendet. Diesen Daten durften freundlicherweise verwendet werden, unter dem Vorwand, dass diese nicht veröffentlicht werden dürfen und das interne Netzwerk nicht verlassen.

3.2 Vorarbeit

Im Rahmen des Modules Projekt 2 wurden bereits einige Vorarbeiten und technische Abklährungen für diese Arbeit gemacht. Diese werden der Vollständigkeits halber hier trotzdem noch einmals erwähnt, wurden aber ausserhalb der vorgegebenen Zeit bereits erarbeitet.

3.3 Methodik

Für dieses Projekt wird die Projektmethodik Scrum gewählt. Zum einen um die administrative Arbeiten möglichst gering zu halten, ganz nach dem agilen Manifesto: "Individuals and interactions over processes and tools"[1] zum andern, da ich in meinem Arbeitsalltag bereits sehr gute Erfahrung mit dieser Methode sammeln durfte. Für die Arbeit stehen 16 Semesterwochen zur verfügung. Ich habe mich für zweiwöchige Sprints entschieden. Das heisst, nach jeweils zwei Wochen, wird der Fortschritt mit dem Betreuer besprochen und geschaut ob das Ziel so weiterhin erreicht werden kann. Dies erlaubt es auch Änderungen am Vorgehen vorzunehmen und auf allfällige Ressourcenengpässe einzugehen. Somit wird die ganze Arbeit über 8 Sprints verteilt mit einem längeren unterbruch in den Weihnachtsferien über die Festtage. Der letzte Sprint wurde als Reservesprint gewählt, um allfällige Verspätungen abfedern zu können. Somit bleiben 7 effektive Arbeitssprints übrig.

Die Arbeit ist zweigeteilt und beinhaltet einerseits die analyse von E-Mail-Signaturen und andererseits das Webscraping der Firmendomain. Der erste Teil wird besonders in Sprint 2 & 3 behandelt. Der Schwerpunkt Webscaping ligt dann im Sprint 4 & 5. In Sprint 6 werden die Resultate dann mittels eines grösseren Datensets analysiert und die Präzision des Systems ermittelt. Sprint 7 ist vor allem dem schriftlichen Teil der Arbeit gewittmet.

Die Arbeit wird jeweils so aufgeteilt, dass sich der erste Sprint also Sprint 2 & 4 jeweils um die Datenbeschaffung und aufarbetung kümmert und im zweiten Sprint, Sprint 3 & 5 werden die Daten jeweils analysiert. Dabei wird nach dem Vorgehen: "Train, Test, Improve" vorgegangen.

[1] Zitat: Agile Manifesto 2001 https://agilemanifesto.org/

4 Ergebnisse

4.1 Projektverlauf

Auf der folgenden Grafik ist der Verlauf der Arbeit zu sehen. Die rote Linie beschreibt den Idealverlauf, wobei die grüne Linie den tatsächlichen Verlauf darstellt. Zu sehen ist in der Grafik, dass nicht an allen Tagen gearbeitet wurde, dies da die Arbeit berufsbegleitend stattfand. In Sprint 5 wurde leider keine Arbeit geleistet, da hier ein geschäftliches Projekt kurzfristig zu einem Ressourcenengpass führte. Die verlorene Zeit wurde über die Festtage durch einen zusätzlichen Sprint wieder aufgeholt. Wie es für solche Arbeiten wohl üblich ist, gab es am Schluss doch noch einen kurzen Endspurt, die Arbeit konnte aber relativ gut aufgeteilt werden, so dass nur wenige extrastunden notwendig waren.

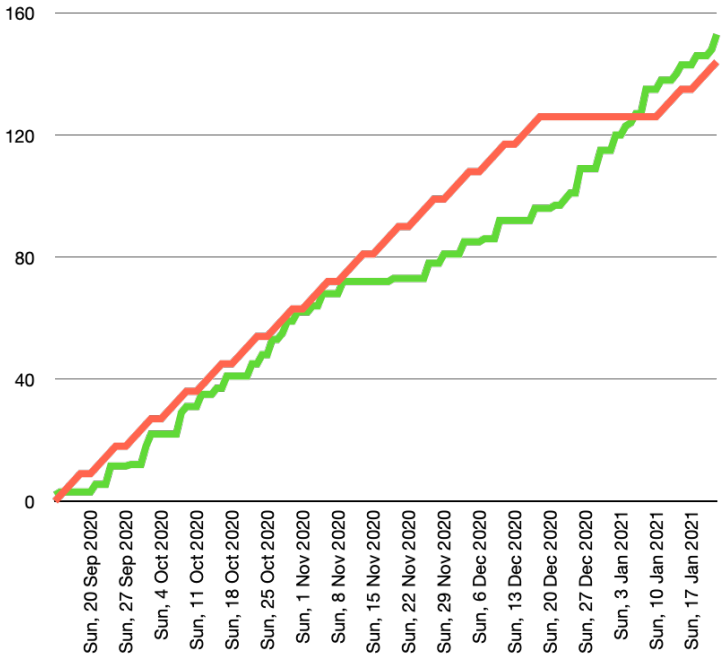

Abbildung 1Burn-up Chart

4.2 Namens Erkennung

Die Namenserkennung wurde bereits im Vorprojekt umgesetzt. Da in einer Signatur nicht aus dem Wortkontext geschlossen werden kann, dass es sich bei einem Wort um einen Namen handelt, wie dies viele andere Wort klassifizierungs Algorithmen tun. Und Namen auch nicht einem bestimmten Muster, folgen, ausser, dass diese im Deutschen wie alle Nomen Grossgeschrieben werden, musste ein Lexikon zuer Erkennung aufgebaut werden, in dem alle Vor- und Nachnamen hinterlegt sind. Hierzu wurde wikidata.org als Datenquelle verwendet. Wikidata ist eine Graphdatenbank, welche Wikipedia unterstützen soll und so das dort gesammelte Wissen maschinenlesbar bereitstellt. Um alle Namen herauszulesen wurden folgende Sparql Skripte verwendet.

Familienname

```
SELECT ?NAME WHERE {
  ?n wdt:P31 wd:Q101352.
  ?n wdt:P1705 ?NAME.
  ?n wdt:P282 wd:Q8229
}
```

Männliche Voramen

```
SELECT ?NAME WHERE {
  ?n wdt:P31 wd:Q12308941;
    wdt:P1705 ?NAME;
    wdt:P282 wd:Q8229.
}
```

Webliche Vornamen

```
SELECT ?NAME WHERE {
  ?n wdt:P31 wd:Q11879590;
    wdt:P1705 ?NAME;
    wdt:P282 wd:Q8229.
}
```

Abbildung 2 Wikidata Export Scripts

Diese Daten wurden danach als CSV exportiert und in einer MYSQL Datenbank abgespeichert. Danach konnte über SQL darauf zugegriffen werden. Dabei konnten zu diesem Zeitpunkt 19964 männliche und 12083 weibliche Vornamen sowie 276544 Nachnamen ins Lexikon aufgenommen werden. Nach dem Abschluss des Projekt 2 wurden ein Name jeweils nicht erkannt, wurde er jeweils manuel dem Testset hinzugefügt.

4.3 Jobtitel Erkennung

Um die Funktion der Person in dem Betrieb zu erkennen musste auch ein Lexikon aufgebaut werden. Hier zu diente jobs.ch als Datenquelle. Da es sich bei jobs.ch um eine single page applikation handelt, hat die Seite eine API, welche die Jobs aus dem Backend holt. Diese API wurde aufgerufen um alle Jobtitel zu speichern. Die API verfügt über ein Paging und erlaubt es maximal 100 Jobs gleichzeitig abzufragen, danach muss ein neuer Request abgesendet werden. Deshalb wurde mit einerm Abstand von jeweils eine Sekunde, um den Server nicht zu stark zu belasen, alle Jobtitel abgefragt. Und Abgespeichert. Dies waren zu diesem Zeitpunkt 65714 verschiedene Jobs. Danach wurde mit einer Regular expression alle Jobs in einer männlichen und in einer weiblichen Variante getrennt und erneut wider in einer MYSQL Datenbank abgespeichert. Nicht zu parsende Jobs wurden ignoriert. Damit konnte ein Lexikon mit 13788 Berufen erstellt werden. Dieses kann nun verwendet werden um die Funktion in der Signatur oder auf der Webseite zu erkennen.

4.3.1 Adresserkennung

Nach Analyse der Daten hat sich herausgestellt, dass die Adressen immer das gleiche Format haben. Diese setzen sich aus seiner 4-5 Stelligen Postleitzahl und einer Ortschaft, einer Strasse und einer Hausnummer oder alternativ dem Begriff Postfach zusammen. Dabei war in den Daten die Hausnummer nie grösser als 3 Ziffern. Auch die Recherche online hat ergeben, dass solche grossen Hausnummern, besonders in Industriegebiete, nicht exisitieren. Ausserhalb der DACH Region existieren weitere Formate, wie Postleitzahlen mit Buchstaben oder das Amerikanische System mit Street and Avenue. Da mein Arbeitgeber ausschliesslich in dieser Region tätig ist, wird ausschlisslich dieses Format unterstützt.

4.3.2 Telefonerkennung

Eine Telefonnummer ist rein optisch und auch technisch relativ einfach zu erkennen. Sie besteht ausschliesslich aus Zahlen oder Zeichen wie ()/+*. Oft sind in einer Signatur allerdings auch mehrere Telefonnummern für direktwahl, mobile oder fax vorhanden. Hier wird aus Praktikabilität jeweils einfach die erste Nummer genommen und die weiteren Nummern als alternative mitgegeben.

4.3.3 Durchsuchen der Webseite nach weiteren Informationen

4.3.3.1 Crawling

In einem ersten Schritt, war der Plan die ganze Domain nach der Person zu durchsuchen. Doch gerade bei grösseren Firmen wie z.B. der BFH war dies aus Performancegründen nicht möglich. Bei folgen jedes Linkes ist die Anzahl Seiten exponentiell angestiegen und selbst nach drei Sprüngen wurde die Person noch nicht immer gefunden. Die zweite Idee war ein Datenset zu übernehmen, dass die gesamte Seite bereits vorgrawled und dieses dann zu durchsuchen. Dabei habe ich mir das Projekt commoncrawl angeschaut, welches ein grosses Datenset an vorgecrwalten Seiten bereits zur verfügung stellt. Allerdings enthielt dieses nur sehr wenige Schweizerseiten und wenn oft auch nicht die komplette Seite. Zudem war die Datenmenge enorm gross um diese auf einem kleinen Webserver, der mir zu Verfügung stand abzuspeichern. Deshalb habe ich mich entschieden das crawling nicht selbst zu machen, sondern von Google übernhemen zu lassen. Mittels der Google API können Suchanfragen an Google gestellt werden und über das keyword site : kann die Domain auf der gesucht werden soll eingeschränkt werden. Somit erhält man alle Seiten, auf der die Person erwähnt wird und muss dann nur noch diese parsen.

4.3.3.2 Parsing

Von dieser Seite wird dann der HTML Sourcecode heruntergeladen und darin nach dem Namen der Person gesucht. Von dort aus wird dann solange das Elternelement dazugenommen, bis sich der Inhalt darin verändert. Somit erhält man die optisch nahen Elemente, diese werden dann mit den gleichen Methoden wie auch bei der E-Mailsignatur nach Inhalten durchsucht.

4.3.3.3 Weitere Quellen

Der Plan war es auch weitere Quellen nach Informationen zu durchsuchen. Dafür hat sich in einem ersten Schritt die Platform LinkedIn angeboten. Hier kann genau gleich wie auch beim crawlen der Domain mittels der Google API und der einschränkung auf eine Domain nach der Person gesucht werden. Allerdings zeigt einem hier LinkedIn jeweils nur ein einziges mal die entsprechende Seite an und beim zweiten mal wird die Registrierungsseite angezeigt. Ich hab mehrere Methoden probiert um dies zu umgehen, löschen der Cookies, aufrufen der Seite mittels Proxy, allerdings war die Abfrage jeweils nur einmal pro IP Adresse möglich. Einzige Methode war mittels VPN die IP Adresse zu verschleiern, allerdings hätten mir hier die Anzahl IP Adressen gefehlt um für das ganze Datenset nach Informationen zu suchen. Bei Recherchen bin ich auf den LinkedIn Sales Navigator gestossen, was wohl die offizielle Lösung für so etwas von LinkedIn ist. Dieser Zugriff kostet aber mehrere tausend Dollar pro Jahr. Weshalb ich hier nicht weitergesucht habe. Mir ist allerdings augefallen, dass in den Vorschautexten von Google jeweils bereits einige Informationen enthalten waren. Google kann also LinkedIn crawlen. Aber auch mit dem User Agent und einer IP Adresse aus der Google Cloud Platform war ich nicht erfolgreich. Leider enthält dieser Text nur wenige Worte und der Jobtitel wird oft abgeschnitten. Aus zeitlichen Gründen, habe ich aber hier nicht mehr weiter geforscht.

5 Diskussion

5.1 Rechtliches

Bei den verarbeiteten Daten handelt es sich um personenpezogene Daten, dass heisst, diese Daten können jeweils einer Person zugeordnet werden und geben Rückschlüsse über diese Person. Obwohl die Schweiz nicht Teil der EU ist, verarbeitet unser Unternehmen auch Daten von europäischen Staatsbrürgern, weshalb in diesem Falle die vor zwei Jahren eingeführte Datenschutzgrundverordnung zum zug kommt.

5.1.1 Verarbeiten von personenbezogenen Daten

Das Herunterladen und Verarbeiten von öffentlichen Daten ist grundsätzlich kein Problem. Wäre das ein Problem, hätte auch Google hier ein Problem. Resp. Jeder Webbrowser, der Daten aus den Internet lädt.

5.1.2 Speichern von personenbezogenen Daten

Grundsätzlich gilt, dass sobald Daten einer Person abgespeichert werden, sei dies manuel oder automatisch muss grundsätzlich das einverständiss der Person eingeholt werden und die Person kann jeder Zeit fordern alle gespeicherten Daten vorzuweisen und auf Wunsch auch löschen lassen.

5.1.3 In der Praxis

Das schriftliche Einverständiss jedes mal zu holen ist selbst im aktuellen manuellen Kontext strikt unmöglich, weshalb dies allgemein in den Datenschutzrichtlinien auf unserer Webseite geregelt ist. Wir handhaben es dann so, dass wir jeweils einmal im jahr die Person darüber informieren, welche Informationen über sie gespeichert sind und sie die möglichkeit hat diese zu korrigieren oder löschen zu lassen.

5.2 Wirtschaftliches

Während meiner Arbeit habe ich auch einige Gespräche geführt mit Unternehmern und Beratern aus dem Bereich CMS. Dabei stoss meine Arbeit auf grosses Interesse. Deshalb habe ich auch einmal nachgefragt, wass denn Ihnen ein solcher Dienst Wert sein würde. Dabei wurden mir folgende Zahlen genannt.

Dienstleistung	Preis
Kontrolle eines Kontaktes	10rp/ Konakt:
Update eines Kontaktes	20rp/ Konakt
Neuanlegen eines Kontaktes	50rp/ Konakt

Tabelle 1 Preisliste

Würde diese Dienstleistung nun in einem mittleren bis grösseren KMUs, wie meines Arbeitgebers verwendet würde dies nach meiner Schätzung nach pro Jahr pro Firma circa folgenden Umsatz ergeben : 8000 x 10Rp + 1200 x 20Rp + 800 x 50Rp = **1440.- CHF**

Würde man die gleiche Rechnung bei einem Grosskonzern machen mit oftmals mehreren Millionen Kontakten, kähme dabei natürlich auch der 10 bis 100 fache Betrag heraus. Würde man einen solchen Grosskonzern oder einige kleinere KMU sals kunden Gewinnen, könnte dieses Produkt durchaus rentabel betrieben werden. Ob der Kunde allerdings wirklich bereit ist, diese Preise pro Kontakt zu bezahlen, kann leider noch nicht gesagt werden, ein Test würde sich aber durchaus lohnen.

6 Folgerungen

6.1 Entwicklung

Aufgrund der Umstellung des aktuellen CMS bei meinem Arbeitgeber hat die Bedeutung für diese Arbeit intern leider etwas an Dringlichkeit verloren. Da ich meinen aktuellen Arbeitgeber in naher Zukunft verlassen werde, wird dieses Projekt vermutlich nicht mehr in Betrieb genommen. Allerdings bin ich auf grosses Interesse bei einem Berater für das neue CMS gestossen, welcher Interesse hat, das Projekt als eigenes Produkt zu vertreiben. Aktuell bin ich darum in Gesprächen wie dies von statten gehen könnte. Vermutlich müsste die Codebasis neu geschrieben werden um dies in ihr aktuelles Produkt zu integrieren und langfristig wartbar zu machen. Die erarbeiteten Lexika könnten aber durchaus weiter verwendet werden.

6.2 Projektverlauf

Durch die Agile vorgehensweise konnte sehr gut auf Änderungen eingegangen werden. Besondere Herausforderung war für mich vorallem die Disziplin im Homeoffice auf Grund der COVID19 Situation so wie eine solche Arbeit berufsbegleitend zu schreiben. Was meiner Meinung nach nicht so gut lief, ist die Kommunikation mit dem Betreuer und Experten, ich hätte hier offener und öffters meinen aktuellen Stand mitteilen sollen. Zudem war auch mein online Repository auf Github nicht immer auf dem aktuellsten Stand, was es für Sie schwer gemacht hat, mehr über den aktuellen Stand zu erfahren.

Glossar

Tabellenverzeichnis

Tabelle 1 Preisliste 9

Abbildungsverzeichnis

Abbildung 1Burn-up Chart 5

Abbildung 2 Wikidata Export Scripts 6