Information Retrieval aus Simap Meldungen

als

Masterarbeit

an der

Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Universität Bern

eingereicht bei

PD Dr. Matthias Stürmer

Institut für Wirtschaftsinformatik

Forschungsstelle Digitale Nachhaltigkeit

von

Schweizer Dominic Jonas

von Neckertal SG

im 11. Semester

Matrikelnummer: 10-808-988

Studienadresse

Dorfberg 555

3550 Langnau i.E

079 819 91 10

dominic.schweizer@dorfberg.ch

Bern, 26.04.21

Zusammenfassung

In dieser Arbeit wurde ein Algorithmus entwickelt, der basierend auf einer Liste von Abfragetexten ähnliche Passagen in einer Sammlung von Texten erkennt und identifiziert. Das Resultat enthält neben dem Text auch die identifizierten Passagen.

Der Algorithmus wurde auf die Eignungskriterien und die Zuschlagskriterien der Simap-Ausschreibungen angewendet. Simap.ch ist das schweizerische Informationssystem über das öffentliche Beschaffungswesen.

In einer Stichprobe von 2000 Meldungstexten konnten 2942 Passagen mit Eignungskriterien und 4976 Passagen mit Zuschlagskriterien identifiziert werden. In Kombination mit der Suchmaschine Elasticsearch kann mit den iterativ festgelegten Abfragen das relevanteste Dokument für ein Projekt gefunden werden.

Summary

The subject of this work describes the development of an algorithm which tags sequences in a collection of texts based on a set of queries. The algorithm returns the text with the tagged sequences as result. This algorithm was applied to the procurement notices from simap.ch, the Swiss platform of public procurement. From a sample of 2000 notices 2942 sequences containing suitability criteria and 4976 containing award criteria were identified. The set queries can then be used to search with Elasticsearch for the most relevant files within the procurement projects.

Inhaltsverzeichnis

ZUSAMMENFASSUNG I
SUMMARY I
INHALTSVERZEICHNIS II
1 EINLEITUNG 4
1.1 Ausgangslage 5
1.2 Problemstellung 8
1.3 Zielsetzung 10
1.4 Aufbau der Arbeit, Methodik des Vorgehens 11
2 UNTERSUCHUNG DATENSATZ 13
2.1 Grundlegende Begriffe 13
2.2 Entwicklung und Verteilung der Meldungen 14
2.3 Beschreibung Datensatz 17
3 METHODEN TEXT MINING 20
3.1 Preprocessing und Vektorisierung Text Daten 22
3.2 Information Retrieval 26
4 ENTWICKLUNG ALGORITHMUS 28
4.1 Grundlegende Begriffe 29
4.2 Dritte und letzte Iteration 32
4.3 Zweite Iteration 35
4.4 Erste Iteration 36
5 RESULTATE 39
5.1 Auswertung der Iterationen 39
5.2 Ermittlung Meta-Daten der Resultate 41
5.3 Untersuchung Eignungskriterien 42
5.3.1 Kategorisierung Eignungskriterien 43
5.3.2 Auswertung Eignungskriterien 45
5.4 Untersuchung Zuschlagskriterien 49
5.4.1 Kategorisierung Zuschlagskriterien 49
5.4.2 Auswertung Zuschlagskriterien 51
5.5 Untersuchung Ausschreibungsunterlagen 54

6 DISKUSSION 56
6.1 Zusammenfassung und Diskussion 56
6.2 Ausblick 57
ANHANG 60
Dank 60
Code 60
ABBILDUNGSVERZEICHNIS 60
TABELLENVERZEICHNIS 62
LITERATURVERZEICHNIS 63
SELBSTÄNDIGKEITSERKLÄRUNG 68

1 Einleitung

Die Plattform simap.ch wird seit 01.03.2009 vom Bund und 23 der 26 Kantone genutzt. Bei der Vorstellung anlässlich einer Pressekonferenz am 29.06.2009 wurde die Bedeutung des öffentlichen Beschaffungswesens von BR Doris Leuthard unterstrichen: 8% des BIP wird für die öffentliche Beschaffung verwendet. Simap.ch bietet nichts anderes als dass es *den elektronischen Zugang zum öffentlichen Beschaffungsvolumen von ca. 40 Milliarden Franken pro Jahr ermöglicht* (Verein Simap, 2009).

Seit 2011 wird von allen Kantonen sowie dem Bund und einigen grösseren Städten das System www.simap.ch als Plattform für ihre Publikationen im öffentlichen Beschaffungswesen eingesetzt (Verein Simap, 2011).

Die Forschungsstelle Digitale Nachhaltigkeit (FDN) wertet die von simap.ch publizierten Meldung aus und informiert mit der Plattform Beschaffungsstatisik.ch (*Beschaffungsstatistik.ch*, 2021) und Intelliprocure.ch (*IntelliProcure - Intelligence im öffentlichen Beschaffungswesen*, 2018) über das öffentliche Beschaffungswesen. Die Auswertung erfolgte dabei zuerst durch eine Analyse der publizierten Meldungstexte, welche mit einem Webcrawler heruntergeladen und mit Regulären Ausdrücken ausgewertet wurden (Gunawan et al., 2019). Die Auswertung beinhaltet unter anderem die Zuweisung der Ausschreibungen zu einer Auftraggeber Art oder die Zuweisung zu einem bestimmten Sektor. Abbildung 1 zeigt die Startseite der Plattform «beschaffungsstatistik.ch» mit Informationen über den Trend der Anzahl Ausschreibungen pro Monat.

Abbildung 1 Beschaffungsstatistik.ch: Trend-Ausschreibungen

Seit 2019 wird die Simap.ch SOAP-Schnittstelle verwendet, um die Meldungen im XML-Format herunterzuladen (KPS Solutions, 2020a). Die Analyse der XML-Meldungen kann anhand der XML-Elemente durchgeführt werden und erhöht die Zuverlässigkeit der Auswertung. Ausserdem ist es so möglich

weitere Informationen wie zum Beispiel die Eignungskriterien (EK) und Zuschlagskriterien (ZK) mit einer hohen Genauigkeit aus den Meldungen zu extrahieren.

2018 wurden in der Schweiz 5.56 Mia CHF für Beschaffungen durch die zentrale Bundesverwaltung ausgegeben (BBL, 2019). Neben der zentralen Bundesverwaltung gibt es auch noch weitere Auftraggeber Arten. Die Dokumentation der Simap.ch XML-Files zeigt bei den Auftraggebern neben der zentralen Bundesverwaltung noch weitere: *Dezentrale Bundesverwaltung*, *Kanton*, *dezentral Kanton* sowie Gemeinde und dezentral Gemeinde. Zusätzlich sind auch Andere und Ausland als Auftraggeber Art aufgeführt.

Von den 40 Mia. im Jahr, welche 2009 erwähnt wurden (Verein Simap, 2009), kann die Auswertung 2020 allerdings nur ~37.5% zuordnen. Trotz dem geringeren Volumen bietet der öffentliche Beschaffungsmarkt vielen Anbietern eine gute Chance, im Corona Jahr 2020 wurden insgesamt 8'725 Zuschläge mit einem Volumen von 15'201'968'424 CHF vergeben (*IntelliProcure - Intelligence im öffentlichen Beschaffungswesen*, 2018).

1.1 Ausgangslage

Die Datenbank der Forschungsstelle Digitale Nachhaltigkeit (FDN) umfasst am 26. März 2021 183'236 Meldungen zu 96'405 Projekten. Die Projekte enthalten eine oder mehrere Meldungen zu einem Beschaffungsobjekt. Zum Beispiel enthält Projekt 208253 vier Meldungen: jeweils einen Zuschlag und eine Ausschreibung in deutscher und französischer Sprache. Zusätzlich sind für 24'181 Projekte die Ausschreibungsunterlagen vorhanden. Von den 183'236 Meldungen sind 88’418 Ausschreibungen, welche Informationen über Eignungskriterien (EK), Zuschlagskriterien (ZK) und geforderte Nachweise (GN) beinhalten können. Eignungskriterien müssen vom Anbieter erfüllt werden, damit er an einem Ausschreibungsverfahren teilnehmen kann (Endtner & Stürmer, 2019). Ein Nichterfüllen eines Kriteriums führt zum Ausschluss aus dem Ausschreibungsverfahren (Tiefbaumamt Kanton Bern, 2010). Zuschlagskriterien dienen dazu das wirtschaftlich günstigste Angebot zu ermitteln; dieses enthält den Zuschlag (Tiefbaumamt Kanton Bern, 2010). Die geforderten Nachweise können die Eignungskriterien präzisieren.

Diese Informationen sind für verschiedene Parteien interessant: In der Forschung können die Kriterien beispielsweise verwendet werden um festzustellen ob, und welche Nachhaltigkeitskriterien in der öffentlichen Beschaffung relevant sind (Welz & Stuermer, 2020). Für die Auftraggeber können die Informationen anderer Ausschreibungen Anregungen und Ideen liefern, und schlussendlich profitieren die Anbieter von den Informationen, indem sie schnell abschätzen können, ob sie für eine Ausschreibung geeignet sind oder nicht.

Neben den Meldungen umfasst die Datenbank ebenfalls eine grosse Anzahl von Ausschreibungsunterlagen. Insgesamt sind am 26. März 2021 die Unterlagen zu 24’181 Projekten verfügbar. Der Datensatz zu den Ausschreibungsunterlagen umfasst 2,1 TB mit insgesamt 617’252 Dateien (*IntelliProcure - Intelligence im öffentlichen Beschaffungswesen*, 2018). Um die EK und ZK aus den Unterlagen zu extrahieren, wurden bereits zwei Ansätze mit Machine Learning untersucht. 2019 untersuchte Endtner die Ausschreibungsunterlagen auf für ZK relevanten Passagen (Endtner, 2019). Dazu verwendete er sämtliche Dokumente, in denen die Worte: *{ Eignungskriteri*, Zuschlagskriteri*, Zulassungskriteri* }* vorkommen und teilte diese anhand ihrer Struktur in Segmente auf. Stürmer und Endtner wendeten 2020 ein ähnliches Verfahren an um Eignungskriterien aus den Unterlagen zu extrahieren, allerdings konnten auch dort nicht für alle Projekte Eignungskriterien ermittelt werden (Endtner & Stürmer, 2019).

Die Verwendung der Ausschreibungsunterlagen bedeutet, dass sehr grosse Datensätze aus unstrukturierten Informationen verarbeitet werden müssen. Die Informationen sind unstrukturiert, weil sie aus reinem Text ohne Strukturierungselemente wie zum Beispiel XML-Tags bestehen. Die Meldungen sind durch das XML-Format bereits semi-strukturiert und Informationen über die EK, ZK und GN können in den XML-Elementen OB01.COND.TECHNICAL; AWARD.CRITERIA und OB01.COND.PROOF abgerufen werden(KPS Solutions, 2020b). Abbildung 2 zeigt einen Ausschnitt aus dem XML der Meldung 980029. Das Element, welches die Informationen über das EK enthält, wurde markiert.

```
<OB01.COND.CONTRACTOR.NOTICE>Sind zugelassen, müssen jedoch deklariert werden.</OB01.COND.CONTRACTOR.NOTICE>
<OB01.COND.TECHNICAL VALUE="NOTICE">Alle wirtschaftlich und technisch leistungsfähigen Firmen, die zudem die
  nachfolgenden Eignungsnachweise erbringen, sind aufgerufen, ein Angebot in CHF zu unterbreiten.
  E1: Referenz des Unternehmers
  E2: Qualitätssicherung mittels unternehmerbezogenem Qualitätsmanagementsystem
  E3: Ausreichende personelle Ressourcen zur termingerechten Realisierung des Bauvorhabens
  E4: Wirtschaftliche / finanzielle Leistungsfähigkeit</OB01.COND.TECHNICAL>
<OB01.COND.PROOF VALUE="NOTICE">Die nachfolgenden Eignungsnachweise / Bestätigungen müssen zusammen mit den
  Angebotsunterlagen eingereicht werden, da ansonsten nicht auf das Angebot eingegangen werden kann:
```

Abbildung 2 XML der Meldung 980029 mit markierten EK

Die Angabe der Informationen ist für die Ersteller der Meldung freiwillig. Falls keine Informationen in den Elementen eingetragen werden, bekommt das Element das Attribut DOCUMENTS. Wenn das Attribut DOCUMENTS gesetzt ist, enthält die auf simap.ch publizierte Meldung einen Standarttext mit einem Verweis auf die Dokumente. So steht zum Beispiel bei den Eignungskriterien der Text: *Siehe Ausschreibungsunterlagen.* Die Abbildung 3, zeigt wie viele der 88'418 Ausschreibungen Informationen über EK, ZK und GN haben.

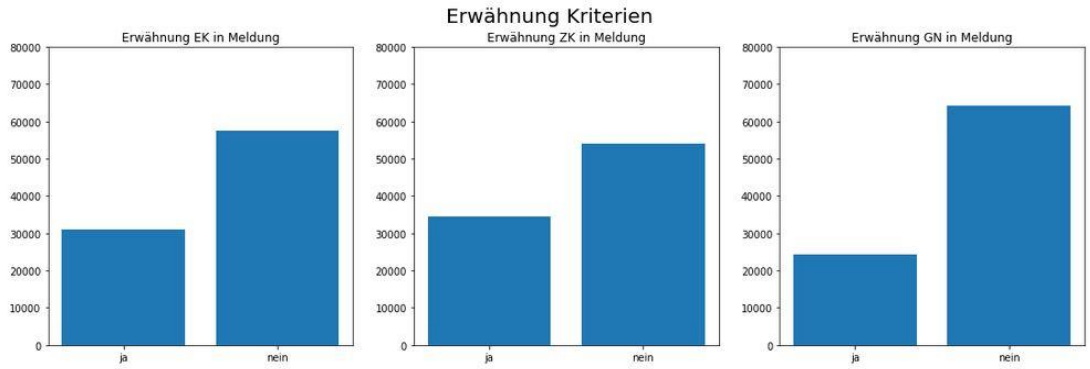


Abbildung 3 Anteil Elemente mit Text zu EK, ZK und GN in Meldungen

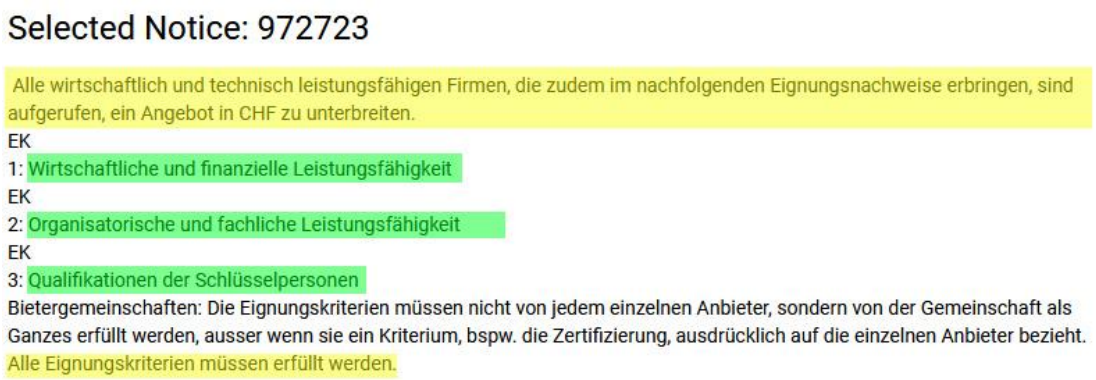
Abbildung 3 zeigt, dass nicht für alle Meldungen Informationen zu den EK, ZK und GN verfügbar sind. Allerdings ist Anzahl der verfügbaren Meldungen grösser als die Anzahl der verfügbaren Ausschreibungsunterlagen, so gibt es 88'815 Ausschreibungen und davon haben 31'069 Informationen zu den Eignungskriterien, 24’172 Informationen zu den geforderten Nachweisen und 34’349 zu den Zuschlagskriterien.

Die Zuschlagskriterien enthalten noch weitere mögliche Attribute: CRITERIA ermöglicht die Zuschlagskriterien als gewichtete Liste zu erfassen und in der Meldung darzustellen. PRICE setzt Preis als einziges Zuschlagskriterium. Bei EK, ZK und GN gibt es ebenfalls NO ATTRIBUTE. Das bedeutet kein Attribut vorhanden was vor allem bei älteren Meldungen vorkommt.

Neben dem Klassifizierungsansatz (Endtner, 2019), (Endtner & Stürmer, 2019) gibt es im Forschungsgebiet Text Mining weitere Ansätze um die Informationen zu den EK, ZK und GN aus den Meldungen und Unterlagen zu extrahieren (Allahyari et al., 2017). Information Retrieval zum Beispiel beschäftigt sich mit dem Ermitteln von Informationen aus einer grossen Dokumentensammlung (Mitra & Chaudhuri, 2000). Die ersten Web-Suchmaschinen verwendeten Information Retrieval Algorithmen (Singhal, 2001). Mit den in Elasticsearch (*Elasticsearch*, o. J.) indexierten und in Intelliprocure.ch verfügbaren Ausschreibungsunterlagen (*IntelliProcure - Intelligence im öffentlichen Beschaffungswesen*, 2018) und den Informationen über die EK, ZK und GN in den Meldungstexten bietet sich somit eine alternative Möglichkeit um die relevanten Dateien zu ermitteln: Anhand der gefundenen Informationen über die EK, ZK und GN können mit der Suchmaschine Elasticsearch ähnliche Dokumente gefunden werden. Allerdings stellt sich hiermit die Frage: Wie kann man die Informationen über EK, ZK und GN in möglichst hoher Qualität aus den Meldungen generieren?

1.2 Problemstellung

Mit der Extraktion der Texte aus den Elementen OB01.COND.TECHNICAL; AWARD.CRITERIA und OB01.COND.PROOF sind drei Sammlungen von Texten vorhanden. In jeder Sammlung befinden sich relevante Passagen, welche Informationen über EK, ZK und GN enthalten; allerdings gibt es auch Passagen, welche auf die Unterlagen verweisen oder gar keine relevanten Informationen enthalten. Stürmer und Endtner unterteilten die Eignungskriterien in ihrer Arbeit in relevant und unrelevant. So sind Passagen, welche den Charakter von Eignungskriterien wie zum Beispiel *Alle Eignungskriterien müssen erfüllt werden* aufweisen, nicht relevant (Endtner & Stürmer, 2019). Die Texte in den Sammlungen können sowohl relevante als auch nicht relevante Informationen enthalten. Die Meldung 972723 in Abbildung 4 enthält zum Beispiel eine Einleitung und einen Abschluss aus nicht relevanten Informationen, die Aufzählung im Mittelteil enthält relevante Informationen über die Eignungskriterien.


Selected Notice: 972723

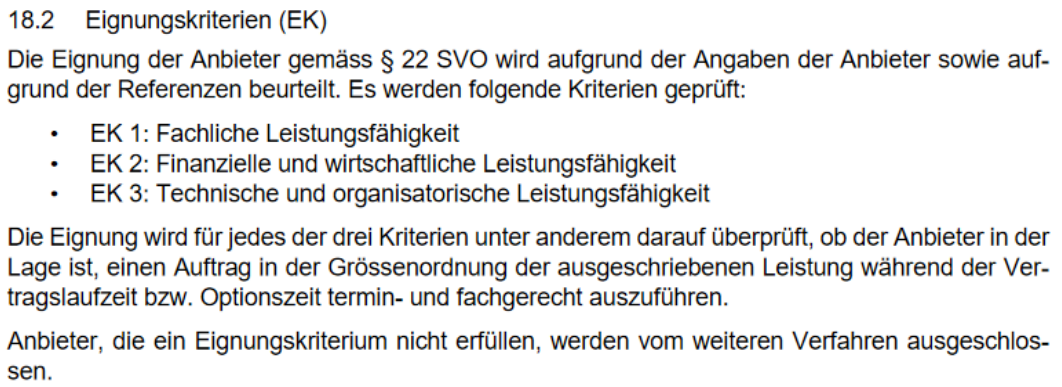
Alle wirtschaftlich und technisch leistungsfähigen Firmen, die zudem im nachfolgenden Eignungsnachweise erbringen, sind aufgerufen, ein Angebot in CHF zu unterbreiten.
EK
1: Wirtschaftliche und finanzielle Leistungsfähigkeit
EK
2: Organisatorische und fachliche Leistungsfähigkeit
EK
3: Qualifikationen der Schlüsselpersonen
Bietergemeinschaften: Die Eignungskriterien müssen nicht von jedem einzelnen Anbieter, sondern von der Gemeinschaft als Ganzes erfüllt werden, ausser wenn sie ein Kriterium, bspw. die Zertifizierung, ausdrücklich auf die einzelnen Anbieter bezieht.
Alle Eignungskriterien müssen erfüllt werden.

Abbildung 4 Text des EK Elements der Meldung 972723 mit identifizierten Passagen.

Um eine hochwertige Liste an Informationen für die Elasticsearch Suchabfragen zu generieren, müssen die relevanten und die unrelevanten Informationen getrennt werden. Zusätzlich ist es von Vorteil, wenn die relevanten Informationen in möglichst kleine, unabhängige Segmente unterteilt werden, um bessere Suchresultate zu erzielen. Zum Beispiel kann die Meldung 972723 in der Abbildung 4 in drei relevante Abfragen unterteilt werden:

- Wirtschaftliche und finanzielle Leistungsfähigkeit
- Organisatorische und fachliche Leistungsfähigkeit
- Qualifikationen der Schlüsselpersonen

Mit diesen drei Abfragen können in Elasticsearch bereits 9'674 Dateien in insgesamt 7489 Projekten gefunden werden. Die von Elasticsearch bestbewertete Datei heisst */185479/teil_a_ausschreibungsbestimmungen.pdf*


18.2 Eignungskriterien (EK)

Die Eignung der Anbieter gemäss § 22 SVO wird aufgrund der Angaben der Anbieter sowie aufgrund der Referenzen beurteilt. Es werden folgende Kriterien geprüft:

- EK 1: Fachliche Leistungsfähigkeit
- EK 2: Finanzielle und wirtschaftliche Leistungsfähigkeit
- EK 3: Technische und organisatorische Leistungsfähigkeit

Die Eignung wird für jedes der drei Kriterien unter anderem darauf überprüft, ob der Anbieter in der Lage ist, einen Auftrag in der Grössenordnung der ausgeschriebenen Leistung während der Vertragslaufzeit bzw. Optionszeit termin- und fachgerecht auszuführen.

Anbieter, die ein Eignungskriterium nicht erfüllen, werden vom weiteren Verfahren ausgeschlossen.

Abbildung 5 Ausschnitt aus Ausschreibungsbestimmung.pdf zeigt die Eignungskriterien

Der in Abbildung 5 gezeigte Ausschnitt zeigt, dass die Suche nach den drei Segmenten gute Resultate liefert. Die gefundene Datei enthält Eignungskriterien und weitere Informationen.

1.3 Zielsetzung

Das Ziel dieser Arbeit ist es einen Algorithmus zu entwickeln, um ähnliche Passagen aus einer Sammlung von Meldungstexten anhand von vorher definierten Abfragen zu EK, ZK und GN identifizieren zu können. Dabei kann jeder Meldungstext mehrere verschiedene Passagen enthalten. Zum Beispiel enthält die Meldung 972723 aus Abbildung 4 insgesamt fünf identifizierte Passagen, wovon drei als relevant eingestuft wurden. Abbildung 6 zeigt eine schematische Darstellung des Aufbaus. Der Algorithmus nimmt zwei Inputs entgegen, nämlich die Sammlung der Texte und die Sammlung der Abfragen und erstellt daraus einen Output: Die Sammlung der Texte mit den identifizierten Abfragen.

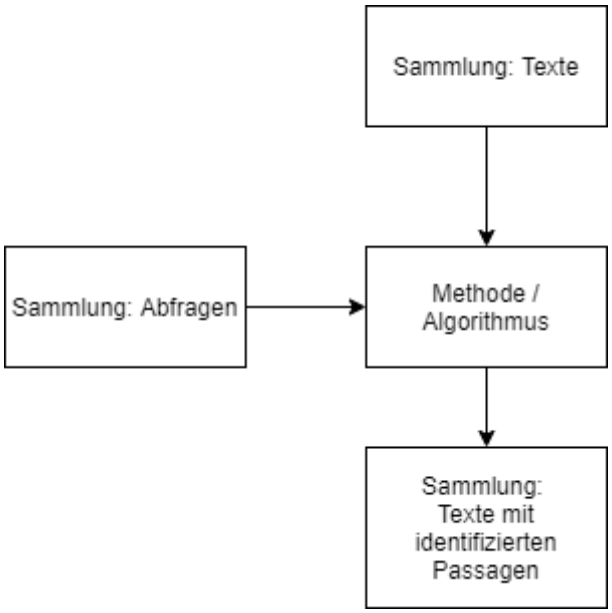


Abbildung 6 Schematische Darstellung Algorithmus

Der entwickelte Algorithmus kann anschliessend gegen die Meldungstexte validiert werden. Mit einem iterativen Vorgehen wird die Sammlung der Abfragen ausgebaut. Dabei wird der Algorithmus ausgeführt und die Resultate werden von Hand untersucht. Wenn es über mehrere Meldungen ähnliche Segmente gibt, welche vom Algorithmus noch nicht identifiziert wurden, werden diese der Sammlung der Abfragen hinzugefügt. Nachdem die Abfragen ergänzt wurden, kann der Algorithmus neu gestartet werden und die nächste Iteration beginnt.

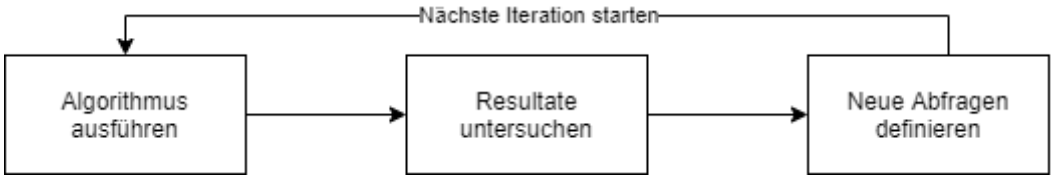


Abbildung 7 Ablauf einer Iteration, um neue Passagen zu identifizieren.

Sobald eine genügend hohe Abdeckung von identifizierten Passagen in den Meldungstexten vorhanden oder eine bestimmte Anzahl von Abfragen erreicht ist, kann die Sammlung der Abfragen ausgewertet werden. Die Einträge in der Sammlung können anschliessend in den in Elasticsearch indexierten Ausschreibungsunterlagen gesucht werden, um zu prüfen ob die gefundenen Passagen auch effektiv in den Ausschreibungsunterlagen vorkommen.

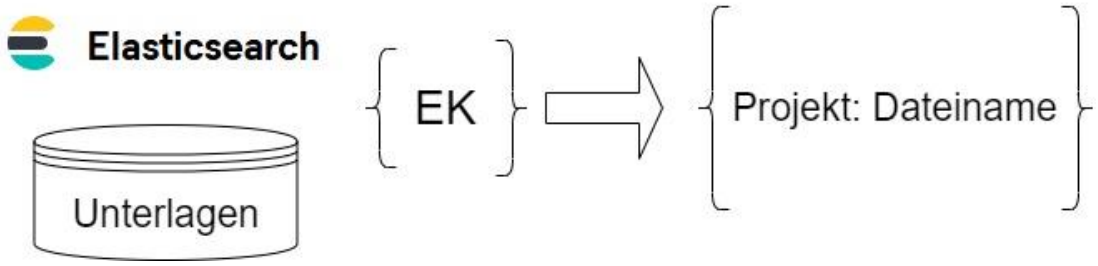


Abbildung 8 Schematische Darstellung Abfrage in Elasticsearch

In Abbildung 8 wird die Sammlung der Abfragen als EK bezeichnet und liefert als Resultat eine Liste von Projektnummern und Dateinamen. Damit können Nutzer die für sie relevanten Unterlagen einfacher finden.

Diese Arbeit bietet damit folgende Mehrwerte in den Bereichen Text Mining und Öffentliche Beschaffungen:

- Algorithmus, um ähnliche Passagen anhand von vordefinierten Abfragen in einer Sammlung von Texten zu finden.
- Liste von Eignungskriterien
- Liste von Zuschlagskriterien
- Liste mit den für die Eignungskriterien relevanten Dokumenten.

1.4 Aufbau der Arbeit, Methodik des Vorgehens

Zuerst werden die Meldungstexte explorativ untersucht. Dabei wird versucht herauszufinden, ob es wiederkehrende Muster in den Texten gibt und ob die Texte einfach mit dem Computer verarbeitbar sind. Anschliessend werden verschiedene Ansätze aus den Forschungsgebiet Text Mining diskutiert, um

herauszufinden, ob und wie sie auf das bestehende Problem angewendet werden können. Anschliessend wird aus den geeigneten Ansätzen der Algorithmus in mehrere Iterationen mit Throw-Away-Prototyping erstellt. Die Verwendung von Throw-Away-Prototypen hat den Vorteil, dass jede Iteration des Algorithmus von Grund auf aufgebaut werden kann und so Fehler aus vorherigen Iterationen nicht die zukünftige Entwicklung beeinträchtigen. Throw-Away-Prototyping kann unter anderem auch für das Sammeln von Anforderungen verwendet werden (Ryan & Doubleday, 2007). Nach der letzten Iteration wird die Sammlung der Abfragen erweitert, indem die in Abbildung 7 beschriebene Methode durchgeführt wird. Die Abfragen werden anschliessend wie in Abbildung 8 auf die in Elasticsearch indexierten Dateien angewendet, um relevante Dateien zu finden.

Der Aufbau folgt dem methodischen Vorgehen: Kapitel 2 beschreibt die Untersuchung der Meldungstexte. In Kapitel 3 wird das Gebiet Text Mining und das Preprocessing von Text erläutert. In Kapitel 4 wird der Algorithmus sowie die einzelnen Iterationen und deren Ergebnisse beschrieben. Kapitel 5 zeigt die Resultate aus den Abfragen der EK und ZK sowie die Ergebnisse der Elasticsearch Suchabfragen. Kapitel 6 schliesst die Arbeit mit der Diskussion der Resultate sowie einigen Vorschlägen für das weitere Vorgehen ab.

2 Untersuchung Datensatz

Im zweiten Kapitel werden zuerst einige grundlegende Begriffe in 2.1 erklärt, welche für das Verständnis dieser Arbeit notwendig sind. Dazu gehören unter anderem die Unterscheidung zwischen strukturierten, semi-strukturierten und unstrukturierten Daten und das Dateiformat XML. Anschliessend wird in 2.2 die Entwicklung und die Verteilung der Simap-Meldungen beschrieben und in 2.3 wird der Datensatz explorativ untersucht. Dabei soll vor allem der Nutzen des Datensatzes diskutiert werden. Abschliessend diskutiere ich die verschiedenen Gebiete im Forschungsbereich Text Mining, um mögliche Lösungen für die gegebene Problemstellung zu finden.

2.1 Grundlegende Begriffe

Bei Datenanalysen wird grundsätzlich zwischen drei verschiedenen Datenstrukturierungen unterschieden: Strukturierte Daten, semi-Strukturierte Daten und unstrukturierte Daten (Sint et al., 2009). Strukturierte Daten sind in vektorisierter Form verfügbar, dabei kann jeder Wert eindeutig durch Koordinaten identifiziert werden. Zum Beispiel repräsentieren Tabellen zweidimensional strukturierte Daten, wo jeder Wert über die Reihe und die Spalte identifiziert werden kann. Abbildung 9 zeigt einen Ausschnitt aus der Tabelle der Datenauswertung aus 2.3 als Beispiel für einen strukturierten Datensatz.

	pub_date	notice_number	type	lang_notice
0	2009-03-02	349793	OB01	DE
1	2009-03-02	351277	OB01	DE
2	2009-03-02	351301	OB01	FR
3	2009-03-02	352385	OB02	DE
4	2009-03-02	354577	OB02	DE

Abbildung 9 Strukturierter Datensatz

Semistrukturierte Daten sind nicht notwendigerweise in tabellarischer Form vorhanden. Die Werte können jedoch trotzdem über diverse Methoden abgerufen werden. Daten in XML-Dateien aus Abbildung 2 sind semistrukturiert. Die Informationen sind in einer Baumstruktur angeordnet, und um sie abzurufen muss das entsprechende Element gefunden werden. Ein weiteres Beispiel sind Key:Value Datenstrukturen wie zum Beispiel das JSON Format. Im JSON

sind die Informationen über Schlüssel verfügbar. Die Meldungsnummer (*notice_number)* in Abbildung 10 kann zum Beispiel als *data.notice_number* aufgerufen werden.


```
{
  "notice_number": "984833",
  "coverage": 0.640625,
  "sections": [
    {
      "start": 0,
      "start_token_index": 0,
      "end_token_index": 4,
      "similarity": 0,
      "dev_len": 999,
      "tag": "untagged",
      "number_of_tokens": 4,
      "text": "EK1Allgemeine Anforderungen:\n",
      "end": 29
    }
  ]
}
```


Abbildung 10 Resultat des Algorithmus aus 4.2 als JSON

Informationen in unstrukturierten Daten können nicht direkt abgerufen werden. Diese Arbeit zum Beispiel besteht aus unstrukturiertem Text, der für den Leser gut verständlich ist. Allerdings ist es schwierig ihn maschinell auszuwerten, um so Erkenntnisse zu gewinnen. Damit unstrukturierte Daten maschinell ausgewertet werden können, müssen diese zuerst aufbereitet werden. Das Forschungsgebiet Text Mining beschäftigt sich mit der Auswertung von Text und versucht mit verschiedenen Methoden Erkenntnisse aus Text zu generieren (Allahyari et al., 2017). In 3 werden verschiedene Methoden diskutiert.

2.2 Entwicklung und Verteilung der Meldungen

Im Zeitraum vom 01.03.2009 bis zum 26.03.2021 wurden insgesamt 183'236 Meldungen an 3'793 Tagen publiziert. Der Tag mit wenigsten Meldungen ist gleichzeitig der erste Tag der Nutzung. Am 02.03.2009 wurde nur eine Meldung veröffentlicht. Der letzte Tag im Zeitraum hat mit 236 Meldungen am meisten publizierte Meldungen pro Tag. Durchschnittlich wurden 40.32 Meldungen pro Tag publiziert, wobei der Median bei 37 Meldungen pro Tag liegt.

Es gibt also einige Tage, an denen sehr viele Meldungen publiziert wurden. Die untenstehenden Boxplots in Abbildung 11 bestätigen diese Annahme.

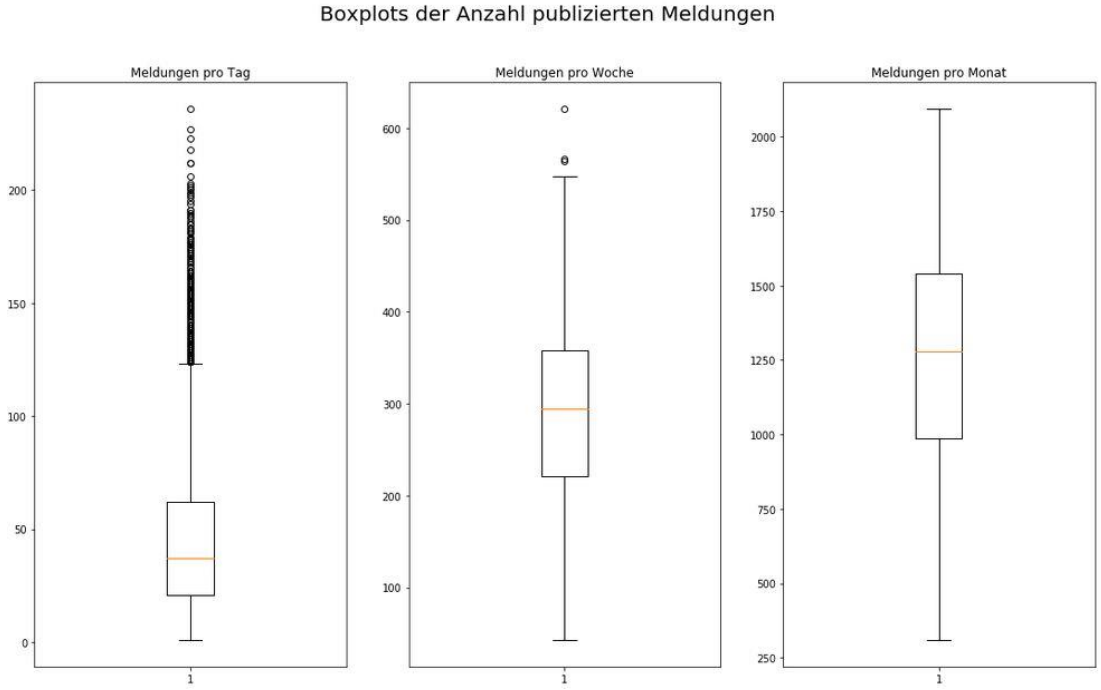


Abbildung 11 Boxplot Anzahl publizierter Meldungen pro Tag, Woche und Monat

Die Anzahl der Ausreisser reduziert sich, wenn die Meldungen pro Woche oder pro Monat gezählt werden, ausserdem erhöht sich die IQR von 41 auf 137 pro Woche und 554 pro Monat. Die hohe Anzahl der Ausreisser könnte auch daran liegen, dass die Verteilung der publizierten Meldungen pro Wochentag nicht gleichmässig ist. Das Bar Chart in Abbildung 12 zeigt, dass die meisten Meldungen jeweils am Freitag publiziert werden. An den anderen Wochentagen werden weniger Meldungen publiziert, wobei die Varianz zwischen Montag und Donnerstag gering ist. An den Wochenenden werden am wenigsten Meldungen publiziert.

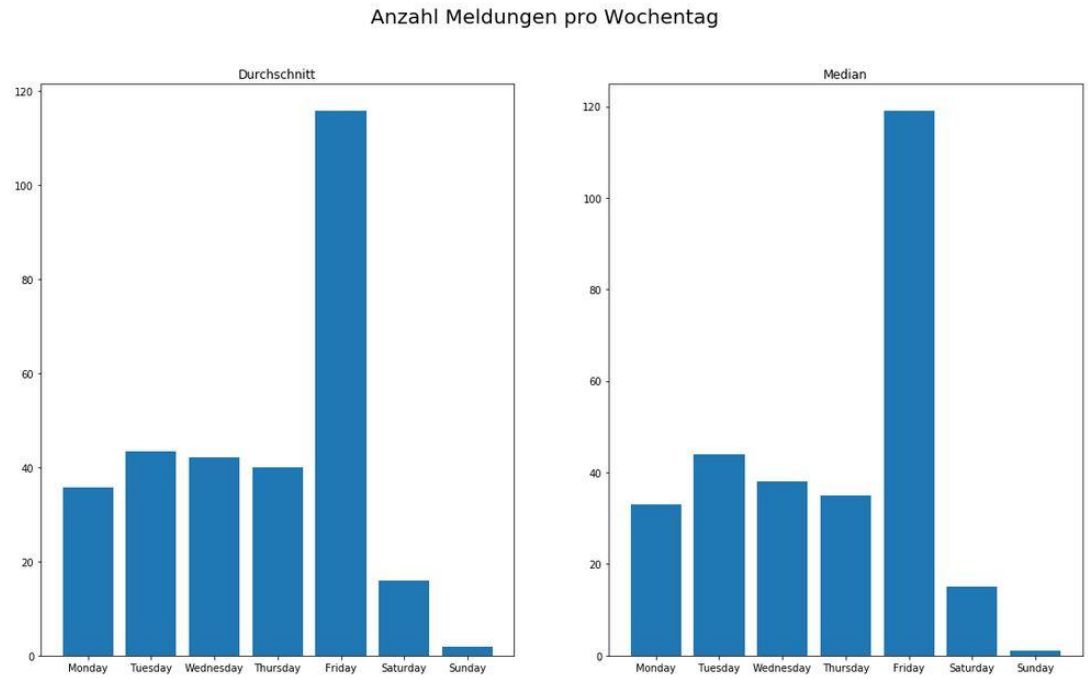


Abbildung 12 Anzahl Meldungen pro Wochentag als Median und Durchschnitt

Die Entwicklung der Anzahl Meldungen pro Woche und pro Monat zeigt einen positiven Trend. Bei beiden Charts in Abbildung 13 ist sichtbar, dass gerade am Anfang von 2009 bis 2011 eine starke Zunahme stattgefunden hat. Dies könnte durch das Onboarding von weiteren Benutzern verursacht worden sein (Verein Simap, 2011). Eine weitere interessante Erkenntnis ist, dass die Anzahl Meldungen pro Woche periodische, starke Ausreisser haben. Dies könnte mit der Altjahreswoche zusammenhängen, bei der viele Behörden ihren Betrieb reduzieren. Grundsätzlich lässt sich anhand der Entwicklung feststellen, dass das Interesse an Simap stetig wächst und immer mehr Meldungen auf simap.ch publiziert werden.

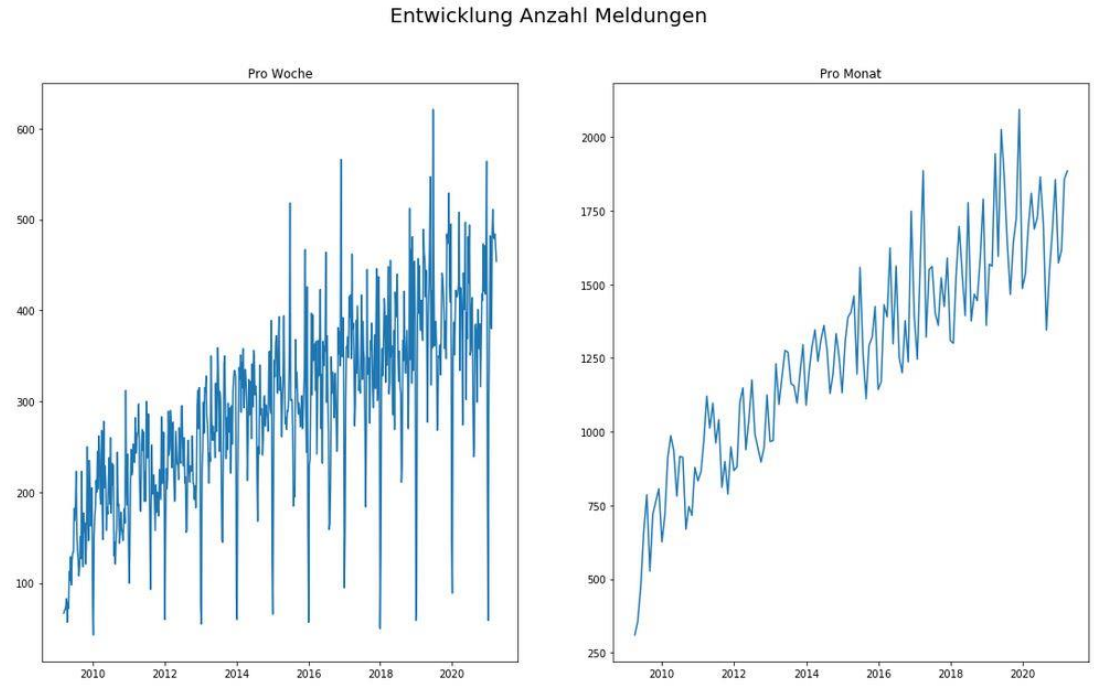


Abbildung 13 Entwicklung der Anzahl der Meldungen pro Woche und pro Monat

2.3 Beschreibung Datensatz

Für die gesamte Arbeit wird der Simap-ver4 Datensatz der Forschungsstelle Digitale Nachhaltigkeit verwendet (*Beschaffungsstatistik.ch*, 2021). Dieser beinhaltet die Meldung im sowohl im HTML als auch im XML-Format. Dabei kann das HTML direkt im Browser dargestellt werden, während das XML weitere Informationen enthält und zuerst aufbereitet werden muss. Für die Untersuchung werden alle vorhandenen Meldungen vom Zeitraum 01.03.2009, dem offiziellen Start von Simap.ch (Verein Simap, 2009) bis zum 26.03.2020 verwendet.

Insgesamt gibt es zehn verschiedene Meldungstypen auf Simap.ch. Die Meldungstypen werden dabei mit OB00-OB09 klassifiziert und die SOAP-Dokumentation gibt jedem Typ eine Beschreibung, welche in Tabelle 1 erläutert sind. Die zehn Meldungstypen werden auf simap.ch in drei Kategorien *Ausschreibungen*, *Zuschläge* und *andere Veröffentlichungen* zusammengefasst.

Tabelle 1: Meldungstypen mit Beschreibung und Kategorie

Meldungstyp	Beschreibung	Kategorie
OB00	Vorankündigung	Ausschreibungen
OB01	Ausschreibung	Ausschreibungen

OB02	Zuschlag	Zuschläge
OB03	Teilnehmerauswahl	Andere Veröffentlichungen
OB04	Abbruch	Andere Veröffentlichungen
OB05	Ausschreibung (Zusammenfas-sung)	Ausschreibungen
OB06	Berichtigung	Andere Veröffentlichungen
OB07	Wettbewerb	Ausschreibungen
OB08	Vergebene Wettbewerbe	Zuschläge
OB09	Widerruf	Andere Veröffentlichungen

Das Balkendiagramm in Abbildung 14 zeigt, dass die meisten der 183'236 Meldungen entweder dem Typ OB01-Ausschreibungen oder dem Meldungstyp Zuschläge OB02-Zuschläge angehören. Interessant ist dabei, dass die Anzahl der Ausschreibungen höher ist als die Anzahl der Zuschläge, da jeder Ausschreibung eigentlich ein Zuschlag oder ein Abbruch folgen müsste und es zusätzlich Zuschläge gibt, welche den Abschluss von freihändigen Verfahren publizieren. Das zweite Balkendiagramm in Abbildung 14 zeigt, wie viele Meldungen in welcher Sprache publiziert wurden. Die Meldungen können in deutscher, französischer, englischer und italienischer Sprache veröffentlicht werden. Dabei gibt es am meisten Meldungen in deutscher Sprache und französischer Sprache. Englische und italienische Meldungen sind eher selten.

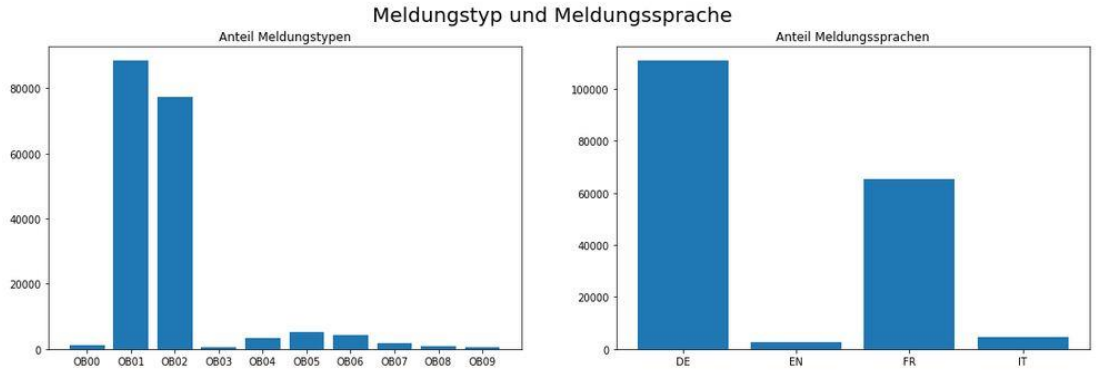


Abbildung 14 Verteilung der Meldungen nach Typ und Sprache

Insgesamt gibt es 57'188 Ausschreibung in deutscher Sprache.

Wie bereits in Abbildung 3 enthalten jeweils ungefähr ein Drittel der Ausschreibungen Informationen über EK und ZK, und ein Viertel der Meldungen enthält Informationen über die geforderten Nachweise. Insgesamt sind 31'069 Texte zu den EK, 34'349 Texte zu den ZK und 24'172 Texte zu den GN verfügbar. Die Untersuchung der Länge der Textblöcke zeigt, dass der längste Text mit 66'966 Zeichen in den geforderten Nachweisen der Ausschreibung 1166559 steht. In diesem Text werden die Eignungskriterien sowie die geforderten Nachweise für alle 8 Lose der Ausschreibung definiert.

3 Methoden Text Mining

Text Mining beschreibt den Prozess, um hochwertige Informationen aus strukturierten, semi-strukturierten und unstrukturierten Texten zu generieren (Allahyari et al., 2017). In einer Erhebung über das Gebiet Text Mining (Allahyari et al., 2017) werden insgesamt zehn Teilgebiete für Text Mining aufgelistet und untersucht. In dem 2012 erschienenen Buch zu Text Mining werden sieben Teilgebiete identifiziert („The Seven Practice Areas of Text Analytics“, 2012). Das VENN Diagramm in Abbildung 15 zeigt die Teilgebiete und vor allem deren Überschneidung mit anderen Forschungsgebieten.

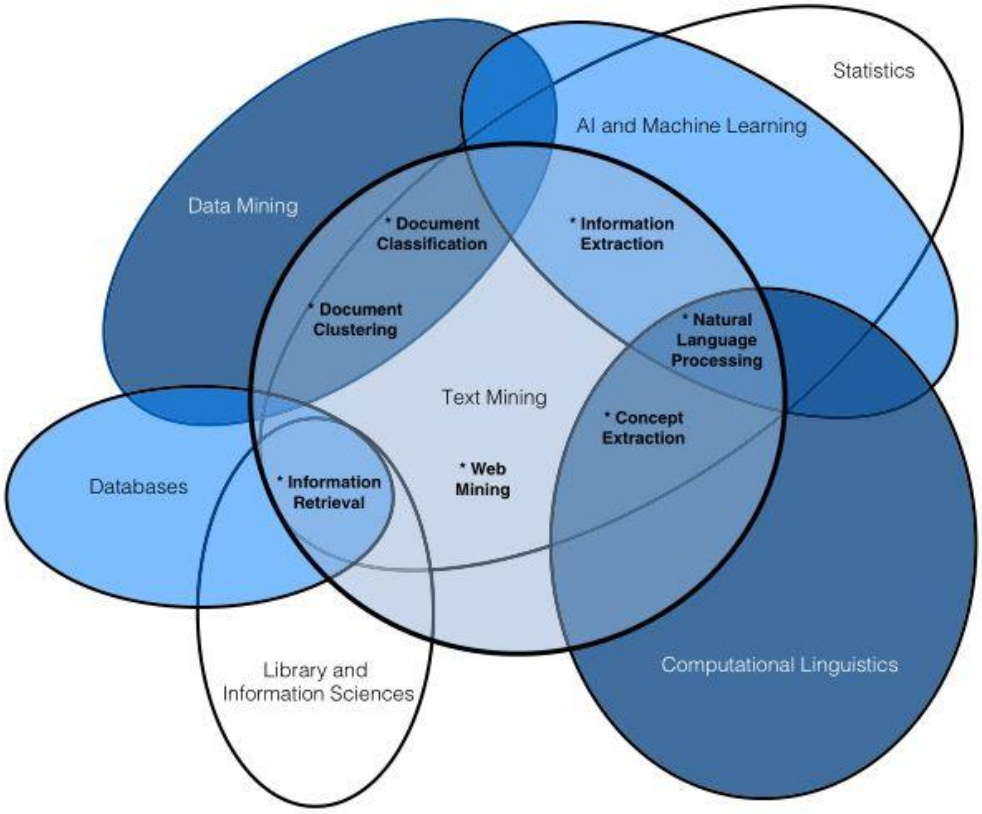


Abbildung 15 Teilgebiete und Überschneidungen von Text Mining („The Seven Practice Areas of Text Analytics“, 2012)

Die bisherigen Arbeiten mit den Ausschreibungsunterlagen (Endtner, 2019) und (Endtner & Stürmer, 2019) lassen sich in das Forschungsgebiet *Document Classification* einordnen.

Tabelle 2 beschreibt einige der Teilgebiete gemäss der Auflistung von (Allahyari et al., 2017) und Abbildung 15. Für diese Arbeit ist vor allem das Teilgebiet Information Retrieval und Natural Language Processing relevant.

Tabelle 2 Teilgebiete von Text Mining nach(Allahyari et al., 2017)

Information Retrieval (IR)	Beschreibt den Prozess, um relevante Dokumente in einer Sammlung unstrukturierter Daten zu finden. Das Ziel von IR ist nicht verborgene Muster in den Daten zu finden, sondern den Zugang zu Informationen zu vereinfachen. Eine Beispielanwendung von IR sind Suchmaschinen
Natural Language Processing (NLP)	NLP ist ein Forschungsgebiet zwischen Informatik, Künstlicher Intelligenz und Linguistik mit dem Ziel mit Computern natürliche Sprache zu verstehen. Viele Text Mining Methoden verwenden NLP-Techniken im Preprocessing.
Information Extraction (IE)	Information Extraction beschreibt die automatische Extraction von Informationen aus unstrukturierten oder semi-strukturierten Daten. Named Entity Recognition und Relation Extraction sind zwei grundlegende Aufgaben von (IE).
Unüberwachtes Lernen	Mit unüberwachtem Lernen können verborgene Strukturen in den Daten gefunden werden. Dabei werden vor allem Clustering und Topic Modelling häufig verwendet. Beim Clustering werden die Dokumente in Cluster aufgeteilt, wobei die Ähnlichkeit innerhalb der Cluster höher ist als die Ähnlichkeit zwischen den Clustern. Topic Modelling wird auch als weiches Clustering bezeichnet. Anstelle einer fixen Zuteilung wird für jedes Dokument die Zugehörigkeit zum Topic ermittelt.
Überwachtes Lernen	Sammlung von Methoden, um Dokumente anhand eines vortrainierten Musters zu klassifizieren. Die Methoden müssen zuerst anhand von Trainingsdaten trainiert werden.

Für die meisten Text Mining Anwendungen ist es notwendig zuerst den Text meist in mehreren Schritten vorzubereiten, in der Literatur werden die verschiedenen Schritte unter dem Begriff Preprocessing zusammengefasst.

In 3.1 wird die Vorbereitung der Texte und die Transformation in ein für den Computer verständliches Format beschrieben. 3.2 diskutiert anschliessend das Forschungsgebiet Information Retrieval.

3.1 Preprocessing und Vektorisierung Text Daten

Das Ziel vom Preproccessing und der Vektorisierung ist es, die Textdaten in eine für den Computer verwendbare Form zu transformieren. In der Literatur wird das Preprocessing und die Vektorisierungen nacheinander ausgeführt (Allahyari et al., 2017).

Ein Framework für Überwachtes Lernen aus Text kann aus den Schritten Preprocessing, Feature Extraktion oder Vektorisierung, Feature Selection und Klassifizierung bestehen. Im Schritt Preprocessing wird der Text vorbereitet, die Vektorisierung transformiert den Text in eine Matrix aus Zahlen. Mit der Feature Selection können überflüssige Dimensionen des Vektors entfernt werden. Die Klassifizierung erfolgt schlussendlich mit einem Klassifizierungsalgorithmus wie zum Beispiel einem Entscheidungsbaum (Uysal & Gunal, 2014).

In dieser Arbeit werden die Schritte Preprocessing, Vektorisierung und Feature Selection in der Abbildung 16 anhand einiger Passagen aus den Eignungskriterien erklärt. Die dabei verwendeten Begriffe werden nacheinander definiert: Eine Sammlung von Texten wird als Corpus und der einzelne Text als Dokument bezeichnet. Jedes Dokument besteht aus einem oder mehreren Tokens, und das Token ist die kleinstmögliche Einheit. ‘‘*Technische Leistungsfähigkeit*‘‘ besteht aus zwei Tokens: *‘Technisch‘* und ‘‘*Leistungsfähigkeit’’*.

Das Preprocessing transformiert das Dokument in eine Liste von Tokens. Dabei ist die Tokenisierung nicht trivial. Eine Trennung bei jedem Leerzeichen funktioniert in den meisten Fällen, allerdings würden dann die Satztrennzeichen an den Worten hängen. Die Satztrennzeichen können ebenfalls als Tokens verwendet werden. Allerdings hängt dies vom verwendeten Werkzeug ab. Der Tokenizer von Spacy betrachtet die Satztrennzeichen als eigene Tokens, während der Tokenizer in Sci-Kit-Learn die Satztrennzeichen vor der Tokenisierung entfernt (explosion.ai, 2021), (*sklearn.feature_extraction.text.CountVectorizer*, o. J.). Die Tokenisierung wird auch in (Dridan & Oepen, 2012; Manning et al., 2009; Yogish et al., 2019) diskutiert. Ein weiterer

Schritt im Preprocessing ist das Entfernen von Stoppwörtern. Für die Entfernung kann entweder ein Werkzeug wie Spacy oder eine Stoppwortliste wie zum Beispiel (*Solariz/German_stopwords*, o. J.)verwendet werden. In Abbildung 16 gehören die Tokens *'und', 'des'* und *'der'* zu den Stoppwörtern und werden entfernt. Die in Abbildung 16 verwendete Lemmatisierung oder die hier nicht verwendete Stammformreduktion (Stemming) reduzieren die Tokens auf die Wortgrundform oder den Wortstamm. Das Ziel ist die Anzahl der unterschiedlichen Ausprägungen eines Wortes zu minimieren. So wird *'finanziellen'* und '*finanzielle*' auf die Grundform '*finanziell*' sowie '*Anbieters*' und '*Anbietenden'* auf '*Anbieter'* reduziert.

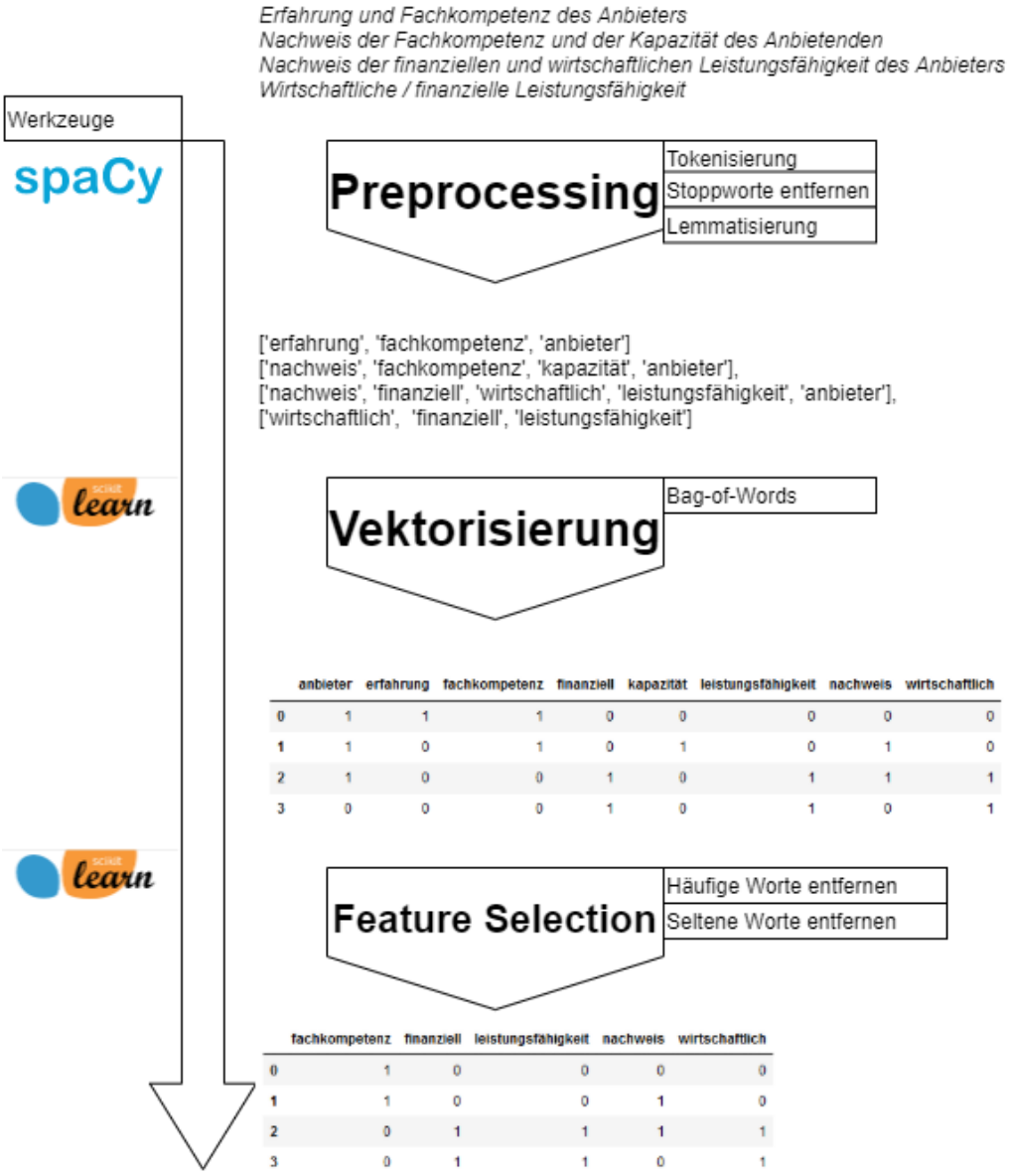


Abbildung 16 Darstellung Preprocessing; Vektorisierung und Feature Selection

Nach dem Preprocessing sind die Texte als Liste von Tokens verfügbar. Um diese Listen maschinell auswerten zu können müssen sie vektorisiert werden. Zur Vektorisierung gibt es verschiedene Ansätze, einige verwenden ein vortrainiertes Model für die Vektorisierung, während andere kein Model benötigen. Die Bag-of-Words (BoW) ist am gebräuchlichsten und berücksichtigt, wie häufig ein Wort im Text vorkommt. Allerdings wird die Reihenfolge der Worte in der BoW Darstellung nicht berücksichtigt (Allahyari et al., 2017). Abbildung 16 zeigt, dass aus den vier Listen von Tokens vier Vektoren mit acht Dimensionen gebildet wurden. Jeder Dimension kann ein Wort zugeordnet werden, der Vektor 0 ist (1,1,1,0,0,0,0,0), das bedeutet, er hat jeweils ein Token mit den Inhalten (‘*anbieter*’, *fachkompetenz’*, ‘*erfahrung’*).

Neben der Term-Count(TC) aus Abbildung 16 und Abbildung 17 gibt es weitere BoW-Ausprägungen, eine sehr häufig benutzte Variante ist die Term-Frequency-Inverse-Dokument-Frequency (TF-IDF), welche nicht nur die Häufigkeit des Wortes im Text und die Textlänge, sondern auch die Häufigkeit des Wortes im Corpus berücksichtigt(Chen, 2020). Abbildung 17 zeigt die vier Vektoren als die Term Count– Darstellung. Für die Vektorisierung kann ein Werkzeug wie (*sklearn.feature_extraction.text.CountVectorizer*, o. J.) verwendet werden.

	anbieter	erfahrung	fachkompetenz	finanziell	kapazität	leistungsfähigkeit	nachweis	wirtschaftlich
0	1	1	1	0	0	0	0	0
1	1	0	1	0	1	0	1	0
2	1	0	0	1	0	1	1	1
3	0	0	0	1	0	1	0	1

Abbildung 17 BoW: Term-Count Darstellung

Damit die Vergleichbarkeit zwischen den Dokumenten bei unterschiedlicher Länge vergleichbar bleibt, können die Vektoren normalisiert werden. In Sklearn und in Text Mining wird die L2-Normalisierung benutzt, welche dazu

führt, dass die Summe der Quadrate eines Vektors 1 ist. Abbildung 18 zeigt die normalisierten Vektoren aus Abbildung 17.

	anbieter	erfahrung	fachkompetenz	finanziell	kapazität	leistungsfähigkeit	nachweis	wirtschaftlich
0	0.577350	0.57735	0.57735	0.000000	0.0	0.000000	0.000000	0.000000
1	0.500000	0.00000	0.50000	0.000000	0.5	0.000000	0.500000	0.000000
2	0.447214	0.00000	0.00000	0.447214	0.0	0.447214	0.447214	0.447214
3	0.000000	0.00000	0.00000	0.577350	0.0	0.577350	0.000000	0.577350

Abbildung 18 Normalisierte Vektoren: Term Frequency Darstellung(L2)

Um eine TF-IDF Darstellung zu erreichen, wird die Term Frequency mit der Inverse Document Frequency multipliziert. Abbildung 19 zeigt die TF-IDF Matrix.

	anbieter	erfahrung	fachkompetenz	finanziell	kapazität	leistungsfähigkeit	nachweis	wirtschaftlich
0	0.448100	0.702035	0.553492	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.392053	0.000000	0.484263	0.000000	0.614226	0.000000	0.484263	0.000000
2	0.375218	0.000000	0.000000	0.463468	0.000000	0.463468	0.463468	0.463468
3	0.000000	0.000000	0.000000	0.577350	0.000000	0.577350	0.000000	0.577350

Abbildung 19 TF-IDF: Term Frequency- Inverse Document Frequency

Der letzte Schritt ist die Feature Selection. Dieser Schritt wird nicht für alle Text Mining Techniken ausgeführt. Doch die Klassifizierung profitiert davon, da das Datenset auf die relevanten Dimensionen reduziert wird. Bei der Feature Selection werden Worte, die für die Klassifizierung nicht relevant sind, entfernt (K. Dalal & A. Zaveri, 2011). Eine einfache Feature Selection Methode ist zum Beispiel ein Schwellenwert. Dieser kann absolut oder relativ sein und als Minimum oder Maximum eingesetzt werden. In Abbildung 16 wird ein absoluter minimaler und ein relativ maximaler Schwellenwert eingesetzt: Alle Worte, die nur einmal im Corpus oder in mehr als 70% der Texte vorkommen, werden entfernt. Nach dem Schritt Feature Selection haben die Vektoren nur noch 6 Dimensionen da die Dimensionen *'erfahrung'*, *'kapazität'* wegen der Minimalschwelle und ‘*anbieter'* wegen der Maximalschwelle entfernt wurden.
Die in Abbildung 16 gezeigten Schritte können je nach Anwendungsfall anders konfiguriert werden.

3.2 Information Retrieval

Mit der in 3.1 beschriebenen Schritten können nun verschiedene Text-und Data-Mining Methoden angewendet werden. Einen guten Überblick über die Methoden bietet (Allahyari et al., 2017; Hotho et al., 2005). Das Gebiet Information Retrieval und dessen Anwendungen wird ausserdem in diesen Arbeiten genauer erläutert: (Bassil, 2012; Henrich, 2008; Manning et al., 2009; Mitra & Chaudhuri, 2000; Singhal, 2001)

Gegenstand des Information Retrieval ist die Suche nach Dokumenten.(Henrich, 2008)

In ihrem Buch über Information Retrieval definieren Manning et al. Information Retrieval als *das Finden von unstrukturierten Daten, meistens Textdokumenten, welche das Informationsbedürfnis befriedigen, aus grossen Sammlungen*.(Manning et al., 2009)

Eine Möglichkeit, um relevante Dokumente zu finden ist, die Ähnlichkeit des Dokuments mit einer eingegebenen Abfrage zu vergleichen.

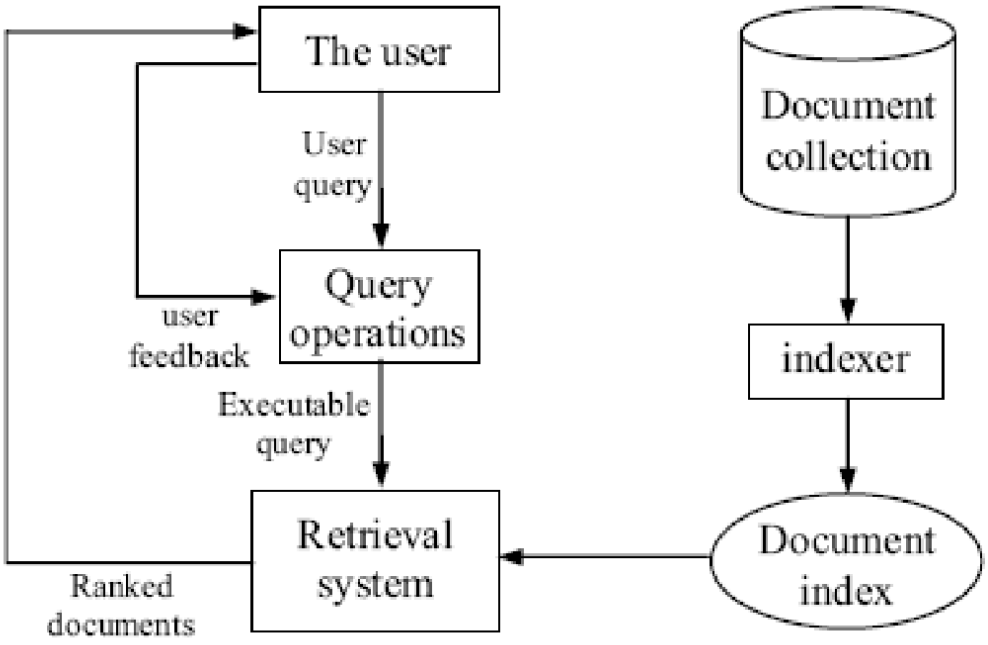


Abbildung 20 IR System (Bassil, 2012)

Die Ähnlichkeit zwischen zwei Dokumenten kann mit dem Winkel ihrer Vektoren bestimmt werden. Durch die L2-Normalisierung kann der Cosinus des Winkels aus dem Skalarprodukt der beiden Vektoren bestimmt werden. Eine

Ähnlichkeit von 1 bedeutet, dass die beiden Dokumente identisch sind, eine Ähnlichkeit von 0, dass die beiden Dokumente keine gemeinsamen Worte haben.

$$S(d_1, d_2) = \sum_{k=1}^{m} w(d_1, t_k) \cdot w(d_2, t_k)$$

Abbildung 21 Berechnung der Ähnlichkeit(Similarity) mit dem Skalarprodukt der beiden Vektoren

Abbildung 21 zeigt die Formeln zur Berechnung der Ähnlichkeit. Mit der Anwendung auf die in Abbildung 19 enthaltenen Vektoren kann eine Distanzmatrix erstellt werden:

	0	1	2	3
0	1.000000	0.443714	0.168135	0.000000
1	0.443714	1.000000	0.371546	0.000000
2	0.168135	0.371546	1.000000	0.802751
3	0.000000	0.000000	0.802751	1.000000

Abbildung 22 Ähnlichkeits-Matrix der in Abbildung 16 definierten Texte

Die Ähnlichkeiten in Abbildung 22 wurde mit Sklearn als Cosinus Ähnlichkeit (sklearn, o. J.)berechnet. Es zeigt sich, dass die Texte: *'Nachweis der Fachkompetenz und der Kapazität des Anbietenden'* und *'Nachweis der finanziellen und wirtschaftlichen Leistungsfähigkeit des Anbieters'* mit 80% die höchste Ähnlichkeit haben. Eine Abfrage mit *'Leistungsfähigkeit dem Auftrag entsprechend.'*

Das von der FDN verwendete Elasticsearch basiert auf der Lucene Syntax, die Lucene Syntax verwendet unter anderem die Cosinus Ähnlichkeit, um die Score eines Suchergebnisses zu berechnen. (Elastic, 2021; *TFIDFSimilarity (Lucene 7.6.0 API)*, o. J.)

4 Entwicklung Algorithmus

Wie in 1.3 formuliert ist das Ziel dieser Arbeit ein Verfahren zu entwickeln, mit dem ähnliche Textpassagen in einer Sammlung von Texten identifiziert werden können. Dabei können in jedem Text der Sammlung unterschiedliche Textpassagen identifiziert werden. Die identifizierten Textpassagen werden mit einem Tag versehen, um die Auswertung zu vereinfachen. In der weiteren Beschreibung wird dieses Vorgehen als Tagging der Texte aus der Textsammlung beschrieben. In Abbildung 23 und Abbildung 24 sehen wir die Eignungskriterien der Meldung 980029 vor und nach dem Tagging.

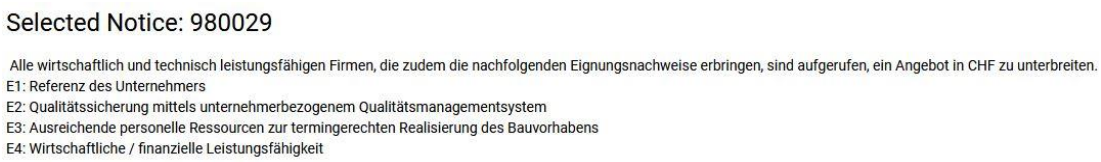


Abbildung 23 Text des Elements EK der Meldung 980029

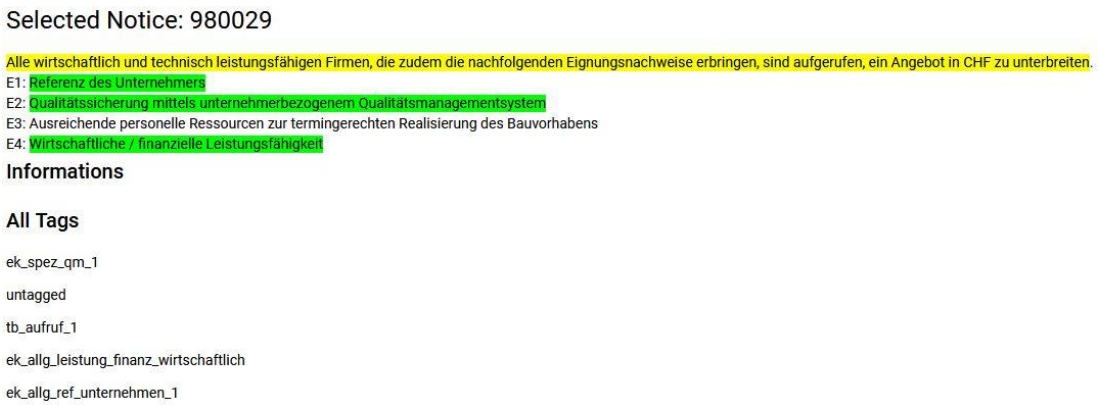


Abbildung 24 Text des Elements EK der Meldung 980029 mit identifizierten Passagen

Als Qualitätskriterium des Taggings wird das Verhältnis der Anzahl getaggten Wörter zu der Anzahl aller Wörter genommen. Es wird im weiteren Text als Coverage oder Abdeckung bezeichnet. Es soll zeigen, wieviel % des Textes mit Tags abgedeckt werden konnte. Der in Abbildung 24 dargestellte Text erreicht eine Abdeckung von 78.94%

Für das Tagging der Meldungen wurde in mehreren Iterationen ein Algorithmus entwickelt. Dabei existierte nach jeder Iteration ein Prototyp, der gegen die Eignungskriterien evaluiert wurde. Für die Evaluation wurden jeweils 2000 Meldungstexte getaggt, und die getaggten Texte wurden anschliessend stichprobenweise untersucht. Bei der Untersuchung wurden folgende Qualitätskriterien bestimmt:

1. Abschluss und Laufzeit des Algorithmus
2. Durchschnittliche Abdeckung der getaggten Texte
3. Qualität der getaggten Passagen:
    a. Enthält die Passage die gleiche Information wie die Abfrage
    b. Enthält die Passage nicht die gleiche Information

Mit den Qualitätskriterien kann bestimmt werden, ob der Algorithmus die Anforderungen erfüllt. Wenn die durchschnittliche Abdeckung tief ist, hat der Algorithmus wenige Passagen identifiziert und getaggt. Anschliessend werden einzelne Meldungen hinsichtlich des Tagging Resultates untersucht. Bei der Bewertung der getaggten Passagen wird die Zuverlässigkeit des Algorithmus bewertet. Dabei wird vor allem überprüft, ob die gefundenen Passagen den gleichen Informationsgehalt haben. Neben den Qualitätskriterien wird der Algorithmus bei der Untersuchung ebenfalls auf Probleme analysiert, und die Probleme werden jeweils in der nächsten Iteration behoben.
In diesem Kapitel wird weiter näher auf die einzelnen Iterationen und Begriffe eingegangen, 4.1 erklärt die in den Algorithmen verwendeten Begriffe sowie den Grundalgorithmus. 4.2 beschreibt die dritte und letzte Iteration des Algorithmus, 4.3 beschreibt die zweite Iteration und 4.4 die erste Iteration. Die Reihenfolge wurde so gewählt, um die letzte und erfolgreichste Iteration detailliert zu beschreiben. In 4.3 und 4.4 wird jeweils auf die Unterschiede zur nächsten Iteration eingegangen, und es wird diskutiert, warum eine nächste Iteration gestartet wurde.

4.1 Grundlegende Begriffe

Für das Verständnis der Algorithmen werden hier die verwendeten Begriffe und Objekte erklärt und definiert. Zusätzlich wird der Grundalgorithmus erklärt. Dieser ist während der Entwicklung der Iterationen entstanden und wird von allen Iterationen implementiert. Das Unterkapitel besteht aus drei Abschnitten. Im ersten Abschnitt werden die verwendeten NLP und Spacy Begriffe erklärt. Im zweiten Abschnitt werden die eigenen Konstrukte definiert, und im letzten Abschnitt wird der Grundalgorithmus erklärt.
Für die Aufbereitung der Texte wird Spacy verwendet. Spacy ist ein NLP Werkzeug, das von der deutschen Firma Explosion.ai entwickelt wurde und für NLP

Aufgaben verwendet wird (explosion.ai, 2021). Bei Iterationen 2 und 3 wird Spacy zur Aufbereitung der Texte verwendet. Die Aufbereitung des Textes erfolgt mit einem vortrainierten NLP Model, mit welchem dem Text weitere Informationen hinzugefügt werden.

Spacy verarbeitet den Text mithilfe eines vortrainierten NLP Models in ein Dokument. Ein Dokument besteht aus einer geordneten Liste von Tokens, jedes Token repräsentiert ein Wort, ein Satztrennzeichen oder einen Whitespace (*Library Architecture · SpaCy API Documentation*, o. J.). Das Token enthält neben dem Wort wie es im Text vorkommt weitere Informationen: So kann zum Beispiel die genaue Position des Wortes im Text oder die Wortart als Attribut des Tokens abgerufen werden. Weiter können verschiedene Formen des Wortes abgerufen werden:

- Wortlemma oder Grundform
- Normalform

Abbildung 25 zeigt einen verarbeiteten Text in vier Formen. Die erste Form zeigt den Text wie er an Spacy übergeben wird. Die zweite Form zeigt das Resultat der Tokenisierung, die Wörter und die Satztrennzeichen wurden getrennt und in einer Liste gespeichert. Die dritte Form zeigt ebenfalls die Liste der Tokens, allerdings wird hier nicht das Token aus dem Text, sondern die Wortgrundform, das Lemma, angezeigt. Die letzte Form zeigt die Wortarten, in Englisch Part-of-Speech, der Tokens.

```
text = Alle wirtschaftlich und technisch leistungsfähigen Firmen, die zudem die
nachfolgenden Eignungsnachweise erbringen, sind aufgerufen, ein Angebot in CHF
zu unterbreiten.

tokens = [Alle, wirtschaftlich, und, technisch, leistungsfähigen, Firmen, ,,
die, zudem, die, nachfolgenden, Eignungsnachweise, erbringen, ,, sind,
aufgerufen, ,, ein, Angebot, in, CHF, zu, unterbreiten, .]

lemmas = ['all', 'wirtschaftlich', 'und', 'technisch', 'leistungsfähig',
'Firma', ',', 'der', 'zudem', 'der', 'nachfolgend', 'Eignungsnachweise',
'erbringen', ',', 'sein', 'aufrufen', ',', 'einen', 'Angebot', 'in', 'CHF',
'zu', 'unterbreiten', '.']

POS = ['DET', 'ADV', 'CCONJ', 'ADV', 'ADJ', 'NOUN', 'PUNCT', 'PRON', 'ADV',
'DET', 'ADJ', 'NOUN', 'VERB', 'PUNCT', 'AUX', 'VERB', 'PUNCT', 'DET', 'NOUN',
'ADP', 'PROPN', 'PART', 'VERB', 'PUNCT']
```

Abbildung 25 Text, tokenisiert, lemmatisiert und als Wortart

Der Algorithmus übernimmt eine Meldung sowie eine Abfrage und erstellt daraus eine getaggte Meldung als Resultat. In diesem Abschnitt werden die verwendeten Inputs, Output sowie die relevanten Zwischenergebnisse definiert.

Inputs:

- Meldung: Die Meldung wird im Preprocessing aus einer Simap XML erstellt. Für jede Meldung gibt es zwei Attribute: Die Meldungsnummer, mit der die Meldung identifiziert wird und der Text, der getaggt werden soll.
- Abfrage: Besteht aus einem Text und einem Tag, für die erleichtere Auswertung wird zusätzlich eine Farbe definiert. Beim Text handelt es sich um die gesuchte Wortkette. Die Abfrage aus Abbildung 26 sucht nach Referenzen auf die Ausschreibungsunterlagen.

```
{
  "text": "Gemäss Ausschreibungsunterlagen",
  "tag": "xx_ref_unterlagen_0",
  "color": "0,255,0"
}
```

Abbildung 26 Beispiel einer Abfrage als Input

Output:

- Getaggte Meldung: Die getaggte Meldung ist das Resultat. Neben den Attributen der Meldungen gibt es eine Liste mit den getaggten Passagen. Für eine einfachere Auswertung wird die Liste der getaggten Passagen mit den nicht getaggten Passagen erweitert. Damit sollte die summierte Länge der Passagen gleich der Länge des Meldungstextes sein.

Zwischenergebnis:

- Kandidat: Falls eine Übereinstimmung zwischen einer Passage im Text und der Abfrage besteht, wird ein Kandidat erstellt. Der Kandidat besteht aus der übereinstimmenden Passage und hat als weiteres Attribut den Tag der Abfrage. Für jede Abfrage und Text wird ein oder mehrere Kandidaten definiert. Die Kandidaten verschiedener Abfragen können sich überschneiden. Da eine Passage nur einen Tag enthalten kann, muss der beste Kandidat vom Algorithmus bestimmt werden.

Jede Iteration befolgt den gleichen Grundalgorithmus, welcher hier beschrieben wird. Die Iterationen variieren in der Umsetzung des Algorithmus und erreichen damit verschiedene Resultate.

1. Lade Textliste aus File
2. Lade Abfrageliste aus File
3. Initialisiere Resultatliste
4. Für jeden Text der Textliste
    a. Initialisiere Kandidatenliste
    b. Für jede Abfrage der Abfrageliste
        i. Finde alle Kandidaten für diese Abfrage und diesen Text und füge sie der Kandidatenliste hinzu
    c. Bestimme die besten Kandidaten aus der Kandidatenliste
    d. Erstelle aus den besten Kandidaten getaggte Passagen
    e. Erstelle aus den getaggten Passagen und dem Text das Resultat und füge es der Resultatliste hinzu
5. Speichere die Resultatliste

4.2 Dritte und letzte Iteration

Die Auswertung von 4.3 zeigt, dass die Menge der falsch identifizierten Passagen hoch ist. Diese Passagen bestehen lediglich aus *welche, die* usw. und haben meistens keine Nomen oder Verben. Die Ursache dafür ist, dass der Vergleich der lemmatisierten Tokens sehr viele Übereinstimmungen bei den Füllworten und Pronomen ergibt. Die hohe Anzahl der falsch getaggten Passagen hat zwei Konsequenzen, welche das Ergebnis negativ beeinflussen: Die Auswertung der getaggten Meldungen wird erschwert, da vielen Meldungen falsche Tags zugeordnet werden. Die zweite Konsequenz kommt aus der Zuordnung des richtigen Tags: Da die Passagen jeweils nur einen Tag haben können, kann es sein, dass der falsche Kandidat als der beste Kandidat bestimmt wurde.

Bei dieser Iteration wird die Menge der falsch identifizierten Passagen reduziert, indem die Kandidaten nur aus Tokens mit den Wortklassen *Adverb, Adjektiv, Nomen, Eigennamen und Verb* generiert werden. Die Wortklassen gehören zu den offenen Wortklassen (*Universal POS tags*, o. J.). Abbildung 27

zeigt eine Textpassage, welche auf die offenen Wortklassen reduziert wurde. Dabei hat sich die Anzahl der Tokens von 24 auf 12 reduziert.

```
Alle wirtschaftlich und technisch leistungsfähigen Firmen, die zudem die
nachfolgenden Eignungsnachweise erbringen, sind aufgerufen, ein Angebot in CHF
zu unterbreiten.

['wirtschaftlich', 'technisch', 'leistungsfähig', 'Firma', 'zudem',
'nachfolgend', 'Eignungsnachweise', 'erbringen', 'aufrufen', 'Angebot', 'CHF',
'unterbreiten']
```

Abbildung 27 Text mit gefilterten Tokens

Der Ausschluss der anderen Wortklassen führt dazu, dass der Satz für den Leser schwerer zu verstehen ist. Der Algorithmus findet weniger Kandidaten, wobei die einzelnen Kandidaten jetzt Worte mit Bedeutung, also Tokens, welche zu den offenen Wortklassen gehören, beinhalten müssen. Dementsprechend haben die gefundenen Kandidaten auch eine höhere Ähnlichkeit zur Abfrage. Der Ausschluss einiger Wortklassen bedeutet allerdings, dass aus den gefundenen Tokens in einem Zwischenschritt eine zusammenhängende Passage gebaut werden muss.

In dieser Iteration wird der Kandidat angepasst: Jeder Kandidat beschreibt nicht nur eine oder mehrere zusammenhängende Passagen, sondern auch eine Liste der übereinstimmenden Tokens. Um die Kandidaten zu ermitteln werden folgende Schritte unternommen:

1. Reduktion Tokenliste auf definierte Wortklassen
2. Ermittlung übereinstimmender Tokens
3. Erstellen von Passagen basierend auf den übereinstimmenden Tokens
4. Erweitern der Passagen basierend auf der Text-Struktur

Die Ermittlung der übereinstimmenden Tokens kann nicht ausschliesslich durch den Vergleich des Lemmas stattfinden, eine Übereinstimmung existiert falls eine der Bedingungen erfüllt werden:

1. Überstimmung Lemma
2. Übereinstimmung Normalform
3. Übereinstimmung kleingeschriebenes Wort
4. Falls eines der Token nicht ausschliesslich aus Buchstaben besteht: Ähnlichkeit der Zeichen mehr als 80% gemäss der Python Sequence

Matcher Library (*difflib — Helpers for computing deltas — Python 3.9.4 documentation*, o. J.).

Abbildung 28 zeigt die Identifizierung eines Kandidaten anhand eines freien Beispiels:

```
text  = Es gibt ein Buch mit dem Titel "Mein Name ist Eugen"
query = Mein Name ist Eugen

# reduzierte Tokens
red_tokens_text  = ['geben', 'Buch', 'Titel', 'Name', 'Eugen']
red_tokens_query = ['Name', 'Eugen']

# Übereinstimmung mit Position
resultat = ['Name', 'Eugen'] mit Positionen [9, 11]

# Erstellung Passage: Token in der Mitte wird verwendet
passage = ['Name', 'ist', 'Eugen'] mit Positionen [9, 10 , 11]

# Erweiterung Passage basierend auf Textstruktur.
passage = ['"', 'Mein', 'Name', 'ist', 'Eugen', '"'] mit Positionen [7,8,9,10,11]
```

Abbildung 28 Identifizierung Passage anhand Tokens und Textstruktur

Nachdem die übereinstimmenden Tokens gefunden wurden, werden aus den vereinzelten Tokens zusammenhängende Passagen erstellt. Dazu werden im ersten Schritt Wortketten gebildet und diese im zweiten Schritt validiert. Für die Bildung der Wortkette wird der Abstand zwischen zwei aufeinanderfolgenden übereinstimmenden Tokens bestimmt, wenn dieser kleiner als oder gleich wie der maximale Abstand ist, wird das nächste Token mit allen Wörtern dazwischen der Wortkette hinzugefügt. Falls der Abstand grösser ist, wird eine neue Wortkette gestartet und die alte wird der Wortkettenliste hinzugefügt. Der maximale Abstand wird als 5 definiert und folgt aus der Auswertung der vorgängigen Iterationen. Nachdem aus allen übereinstimmenden Tokens Wortketten gebildet wurden, werden die Wortketten validiert. Damit eine Wortkette valid ist, müssen zwei Bedingungen erfüllt werden: Die Wortkette muss eine minimale Länge aufweisen und die Position jedes Wortes in der Wortkette muss höher sein als die Position des vorhergehenden Wortes.

Die Wortketten werden anschliessend am Anfang und am Ende erweitert, indem der Algorithmus die Struktur des Textes in der Nähe der Wortkette prüft. Dafür wird eine bestimmte Anzahl der vorherigen und darauffolgenden Tokens untersucht. Falls innerhalb der untersuchten Tokens ein Zeilenumbruch oder

ein Punkt gefunden wird oder der Anfang des Textes oder das Ende des Textes gefunden wird, wird die Wortkette bis zu diesem Token erweitert.

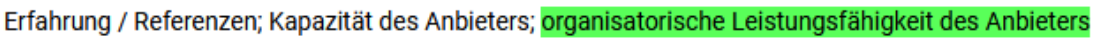


Abbildung 29 Beispiel für eine Erweiterung basierend auf der Struktur des Textes, die Übereinstimmung organisatorische Leistungsfähigkeit wird erweitert.

Abbildung 29 zeigt eine, in einer Meldung gefundene, Passage, die bis zum Ende erweitert wurde. Die Abfrage für die Passage lautete: *Organisatorische Leistungsfähigkeit*. Da sich die Passage fast am Ende des Textes befindet, wird sie bis zum Schluss erweitert. Die validen Wortketten werden mit dem Tag der Abfrage versehen und als Kandidaten zurückgegeben. Die Evaluation der besten Kandidaten wird durch die restriktivere Ermittlung der Kandidaten vereinfacht. Wenn sich zwei Kandidaten überschneiden, besteht ein Konflikt zwischen diesen beiden Kandidaten. Zur Konfliktauflösung wird die Ähnlichkeit der Kandidaten mit ihrer entsprechenden Abfrage ermittelt. Die Ähnlichkeit wird mit einer Methode von Spacy berechnet, welche auf dem in 3.2 erklärten Cosinus Ähnlichkeitsmass basiert.
Sobald alle Konflikte aufgelöst sind, kann das Resultat erstellt werden. Dazu werden die Passagen zwischen den identifizierten Passagen mit ungetaggten Passagen ergänzt, und es wird eine getaggte Meldung erstellt. Ausserdem werden zusätzliche Attribute wie zum Beispiel die Abdeckung des Textes berechnet. Abbildung 10 zeigt ein Beispiel, wie das Resultat für einen Text aussehen könnte.
Die Berechnung der Abdeckung wird in der folgenden Formel dargestellt.

$$\text{Abdeckung} = \frac{\sum Token_{Type=Alphabetic,Tag=untagged}}{\sum Token_{Type=Alphabetic}}$$

Es werden alle Wörter in den getaggten Passagen mit der Gesamtzahl der Worte in der Meldung verglichen. Die Verwendung der Worte führt dazu, dass Aufzählungszeichen wie zum Beispiel *E1* nicht für die Berechnung der Abdeckung verwendet werden.

4.3 Zweite Iteration

Die zweite Iteration verwendete ebenfalls Spacy für die Vorbereitung der Texte. Im Gegensatz zu der dritten und letzten Iteration wurden hier sämtliche

Wortklassen ausser den Satztrennzeichen für die Bestimmung der Kandidaten verwendet. Allerdings erfolgte die Ermittlung der Übereinstimmung nur mit dem Lemma und nicht mit verschiedenen Formen. Übereinstimmungen wurden nur als Position des Wortes in eine Liste eingetragen. Die Liste der Übereinstimmung bestand in diesem Fall nur aus Zahlen und nicht aus Tokens.
Aus der Liste der Übereinstimmungen wurden zusammenhängende Passagen gebildet. Falls zwei Übereinstimmungen zu weit auseinander liegen, werden zwei separate Passagen gebildet. Jede Passage wird auf ihre Gültigkeit geprüft. Gültige Passagen haben eine Mindestanzahl an Tokens, Wörter aus der Abfrage und die Wörter sind in einer Reihenfolge.
Die gültigen Kandidaten wurden ähnlich zur dritten Iteration anhand ihrer Ähnlichkeit verglichen. Im Gegensatz zur dritten Iteration wurden aber nicht zuerst sämtliche Konflikte ermittelt, sondern die Kandidaten wurden nach der Position des ersten Wortes aufsteigend sortiert und die Konflikte zwischen den aufeinanderfolgenden Kandidaten gelöst. Aufgrund der weniger restriktiven Ermittlung der Kandidaten mussten mehr Konflikte gelöst werden. Zusätzlich gab es viele Kandidaten mit wenigen Worten und wenig Informationsgehalt.

4.4 Erste Iteration

Die erste Iteration verwendete kein Spacy für das Preprocessing. Stattdessen wurden die Texte an ihren Leerzeichen getrennt und so tokenisiert. Um einen Kandidaten zu ermitteln wurden gewichtete N-Gramme verwendet. N-Gramme sind N aufeinanderfolgende Wörter. Zum Beispiel bildet "Technische Leistungsfähigkeit" ein 2-Gramm oder bi-Gramm. Nachdem alle übereinstimmenden N-Gramme ermittelt wurden, wird jedem Token die Zahl N zugewiesen, wobei N die Länge des grössten N-Gramms ist, welches dieses Token beinhaltet. Abbildung 30 zeigt die gleiche Beispielabfrage wie in 4.2 verwendet wurde.

```
query = Mein Name ist Eugen
text = Es gibt ein Buch mit dem Titel "Mein Name ist Eugen"
4gram = ('Mein', 'Name', 'ist', 'Eugen')
3gram = ('Name', 'ist', 'Eugen')
2gram = ('ist', 'Eugen')
```

Abbildung 30 Abfrage mit übereinstimmenden N-Grammen

Das Resultat zeigt, dass es ein 4-Gramm, ein 3-Gramm, ein 2-Gramm gibt. Die N-Gramme wurden ermittelt indem der Abfragetext als Moving Frame über den Text läuft und die übereinstimmenden N-Gramme speichert. Dabei wurden N-Gramme welche direkt in das nachfolgende N-Gramm passen, verworfen. Zum Beispiel passt das 3-Gramm *('Mein', 'Name', 'ist')* direkt in das 4-Gramm *('Mein', 'Name', 'ist', 'Eugen')* und wird verworfen, da keine zusätzlichen Erkenntnisse daraus gewonnen werden können.

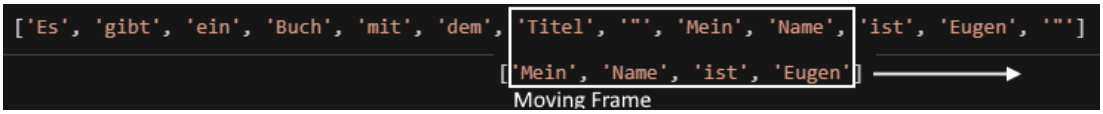


Abbildung 31 Darstellung Moving Frame

Abbildung 31 zeigt eine schematische Darstellung des Moving Frame, wobei die Abfrage *Mein Name ist Eugen* den Moving Frame definiert. Das Resultat der Abbildung 31 wäre das 2-Gramm oder Bigramm *('Mein', 'Name')* welches verworfen wird, weil es in *('Mein', 'Name', 'ist')* passt. Die Idee des Moving Frame stammt von (Salton et al., 1993). Salton et al. haben in ihrer Arbeit ein 'Moving-Window' über n-Sätze ihrer Texte gelegt, um bessere Ergebnisse für ihre IR Abfragen zu bekommen. Nachdem alle N-Gramme für eine Abfrage bestimmt wurden, werden alle Tokens gewichtet. Die Gewichtung eines Tokens wird durch das grösste N-Gramm bestimmt, zu welchem das Token gehört. Im Beispiel bekommen die Tokens *Mein, Name, ist, Eugen* jeweils eine Gewichtung von 4, da sie alle zum 4-Gramm *('Mein', 'Name', 'ist', 'Eugen')* gehören.

Abbildung 32 zeigt die Gewichtung aller Tokens im Beispieltext.

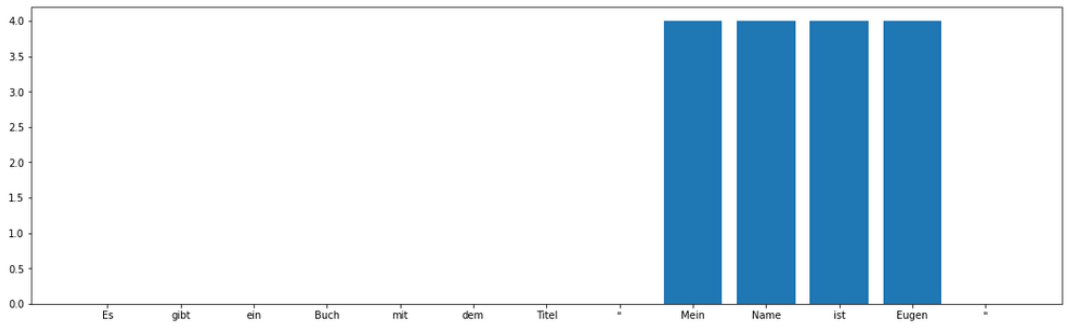


Abbildung 32 Gewichtung der Tokens

Das hier gezeigte Beispiel führt zu einem klaren Resultat. Allerdings wurden bei der Auswertung der EK mit diesem Algorithmus einige Probleme entdeckt:

1. Die Tokenisierung anhand der Leerzeichen im Text führte dazu, dass Satztrennzeichen zum Wort dazugefügt wurden.
2. Durch die fehlende Lemmatisierung wurden nicht sehr viele Übereinstimmungen gefunden, weil nicht das Lemma, sondern die Wortausprägung verglichen wurde.
3. Die Performance war aufgrund der vielen Schleifen sehr schlecht.

Die Probleme 1 und 2 führten dazu, dass mehr Abfragen als eigentlich notwendig erfasst werden müssen. Dies hatte zur Folge, dass das Problem 3 den Ausschlag gab, die Entwicklung abzubrechen und die zweite Iteration zu beginnen. In der zweiten Iteration wurde Fokus auf ein Preprocessing mit Lemmatisierung gelegt, damit mit weniger Abfragen eine höhere Abdeckung erreicht werden kann.

5 Resultate

Mithilfe des entwickelten Algorithmus können die Passagen in den Texten ermittelt werden. Die Passagen werden in mehreren Schritten in den beiden Datensätzen EK und ZK ermittelt und in jedem Schritt wird das Resultat mit dem Auswertungsclient ausgewertet. Nachdem in beiden Datensätzen eine durchschnittliche Abdeckung von mehr als 50% gefunden wurde, werden Meta-Daten zu den gefundenen Passagen ermittelt. Anschliessend werden die Metadaten mit Informationen zu den Meldungen ergänzt und untersucht. Der Auswertungsclient und die Auswertung der Iterationen werden in 5.1 erläutert. In 5.2 wird die Erhebung der Metadaten definiert. 5.3 enthält die Untersuchung der EK, und in 5.4 werden die ZK untersucht. In 5.5 werden die in 5.3 ermittelten Abfragen mit Elasticsearch auf die Ausschreibungsunterlagen angewendet, um relevante Dateien zu finden.

5.1 Auswertung der Iterationen

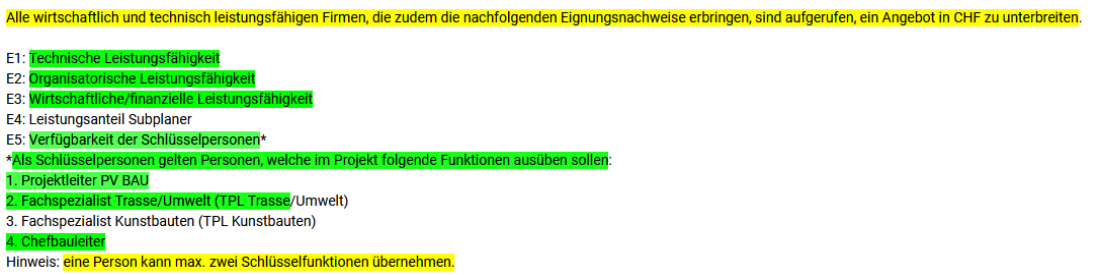
Mit dem Algorithmus können iterativ häufig vorkommende Passagen im Corpus von Hand identifiziert und in den Abfragen definiert werden. Nachdem einige Passagen als Abfragen definiert wurden, wird der Algorithmus neu gestartet, um diese Passagen im Corpus zu identifizieren und mit einem Tag zu versehen.

```
{
  "text": "wirtschaftliche / finanzielle Leistungsfähigkeit",
  "tag": "ek_allg_leistung_finanz_wirtschaftlich",
  "color": "0, 255, 0"
}
```

Abbildung 33 Beispiel einer Abfrage

Abbildung 33 zeigt eine Abfrage. Der Text beschreibt die Passage, welche in den Meldungstexten identifiziert werden soll. Der Tag wird der identifizierten Passage zugeordnet, um nachfolgende Analysen zu ermöglichen, und die Farbe dient zur einfacheren Identifizierung der Passagen im Auswertungsclient. Die Auswahl der Farben beschränkt sich auf Grün und Gelb, wobei Grün für identifizierte Eignungs- oder Zuschlagskriterien steht und Gelb für Passagen, welche identifiziert wurden, aber kein Eignungs- oder Zuschlagskriterium enthalten. Zum Beispiel kommt der Text ‘*Alle wirtschaftlich und technisch*

leistungsfähigen Firmen, die zudem die nachfolgenden Eignungsnachweise erbringen, sind aufgerufen, ein Angebot in CHF zu unterbreiten.' in dieser oder ähnlicher Variation 664-mal in den untersuchten 2000 Texten zu den Eignungskriterien vor, besteht aber mehr aus einem Aufruf zur Angebotsabgabe statt aus einer nützlichen Information für den Anbieter. Abbildung 34 zeigt ein Beispiel eines getaggten Textes der Eignungskriterien mit einem Aufruf(*gelb)* und Eignungskriterien(*grün)*.


Alle wirtschaftlich und technisch leistungsfähigen Firmen, die zudem die nachfolgenden Eignungsnachweise erbringen, sind aufgerufen, ein Angebot in CHF zu unterbreiten.

E1: Technische Leistungsfähigkeit
E2: Organisatorische Leistungsfähigkeit
E3: Wirtschaftliche/finanzielle Leistungsfähigkeit
E4: Leistungsanteil Subplaner
E5: Verfügbarkeit der Schlüsselpersonen*
*Als Schlüsselpersonen gelten Personen, welche im Projekt folgende Funktionen ausüben sollen:
1. Projektleiter PV BAU
2. Fachspezialist Trasse/Umwelt (TPL Trasse/Umwelt)
3. Fachspezialist Kunstbauten (TPL Kunstbauten)
4. Chefbauleiter
Hinweis: eine Person kann max. zwei Schlüsselfunktionen übernehmen.

Abbildung 34 Getaggter Text der EK aus Meldung 790455

Jede Passage kann jeweils mit einem Tag versehen werden. Für die Definition von neuen Abfragen werden also die Texte im Auswertungsclient von Hand ausgewertet, auffällige Passagen können anschliessend in den Abfragen definiert werden, damit sie in der nächsten Iteration durch den Algorithmus identifiziert werden.

Das Resultat des Algorithmus wird als JSON ausgegeben und ist mit einem Beispiel in Abbildung 10 in 2.1 dargestellt. Im Auswertungsclient wird das Resultat mit JavaScript ausgewertet und dargestellt, siehe Abbildung 34. Das Resultat zeigt ebenfalls die Meldungsnummer des Textes, um einen Vergleich mit dem richtigen Meldungstext zu ermöglichen.

Das Resultat enthält also eine Liste mit allen Meldungen der Stichprobe im JSON Format. Dabei besitzt jede Meldung eine Liste mit allen identifizierten Passagen. Nicht identifizierte Passagen werden ebenfalls in der Liste integriert, um die Auswertung zu vereinfachen. Tabelle 3 beschreibt die relevanten Felder des Resultates und deren Bedeutung.

Tabelle 3: Relevante Felder des Resultates

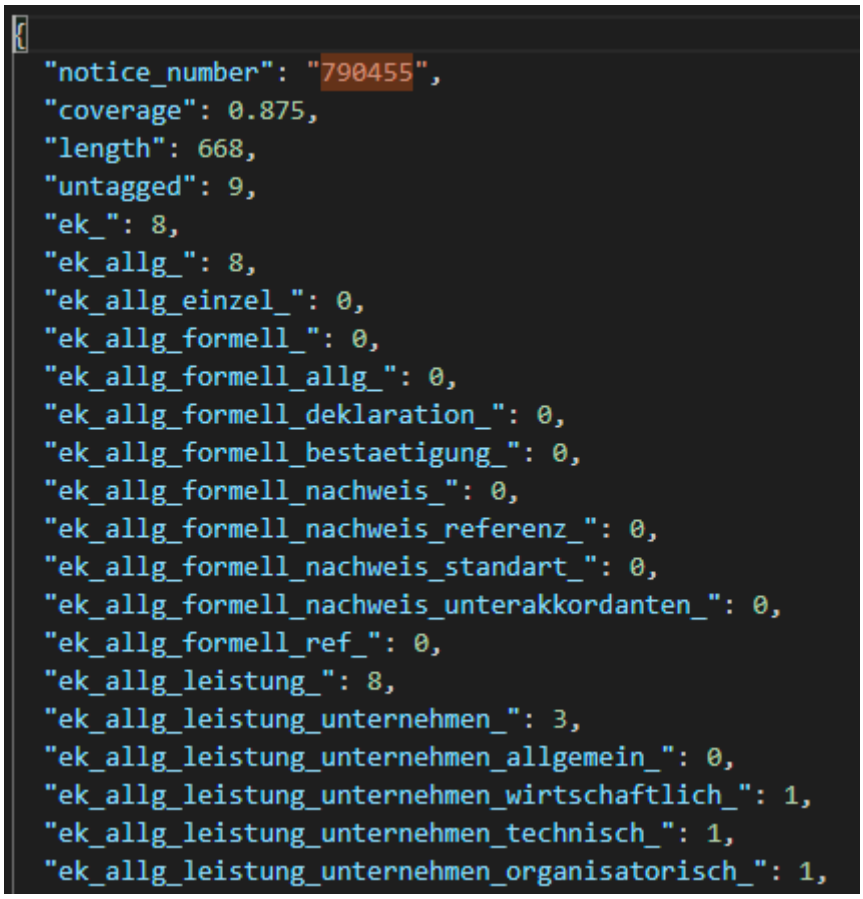
Notice_number	Meldungsnummer, dient dazu die Meldung eindeutig zu identifizieren.

Coverage/Abdeckung	Die Abdeckung zeigt wieviel des Textes identifiziert werden konnte.
Sections	Sammlung aller Passagen aus dem Text, die Passagen können getaggt oder ungetaggt sein.
Section.Similarity	Ähnlichkeit einer identifizierten Passage mit der angewendeten Abfrage. Bei 100% Ähnlichkeit stimmt der Abfragetext exakt mit der Identifizierten Passage überein. Ungetaggte Passage haben immer eine Ähnlichkeit von 0.
Tags	Liste aller Tags, die im Text vorkommen.

Neben den erwähnten Feldern gibt es noch weitere Felder im Resultat, welche vor allem für die Überwachung und Fehlersuche verwendet werden.

5.2 Ermittlung Meta-Daten der Resultate

Der Algorithmus aus 4.2 liefert also als Resultat ein JSON. Um eine effiziente Auswertung des Datensatzes zu ermöglichen, muss das JSON zuerst in einen strukturierten Datensatz transformiert werden. Durch die Strukturierung kann der Datensatz mit pandas (*User Guide — pandas 1.2.4 documentation*, o. J.) ausgewertet und mit matplotlib (*Matplotlib: Python plotting — Matplotlib 3.4.1 documentation*, o. J.) visualisiert werden. Zur Transformation wurde ein JavaScript geschrieben, welches durch die einzelnen Einträge im Resultat iteriert und zu jedem Eintrag Metadaten erfasst. Unter anderem wurde die Anzahl der einzelnen Tags ermittelt.



```
{
  "notice_number": "790455",
  "coverage": 0.875,
  "length": 668,
  "untagged": 9,
  "ek_": 8,
  "ek_allg_": 8,
  "ek_allg_einzel_": 0,
  "ek_allg_formell_": 0,
  "ek_allg_formell_allg_": 0,
  "ek_allg_formell_deklaration_": 0,
  "ek_allg_formell_bestaetigung_": 0,
  "ek_allg_formell_nachweis_": 0,
  "ek_allg_formell_nachweis_referenz_": 0,
  "ek_allg_formell_nachweis_standart_": 0,
  "ek_allg_formell_nachweis_unterakkordanten_": 0,
  "ek_allg_formell_ref_": 0,
  "ek_allg_leistung_": 8,
  "ek_allg_leistung_unternehmen_": 3,
  "ek_allg_leistung_unternehmen_allgemein_": 0,
  "ek_allg_leistung_unternehmen_wirtschaftlich_": 1,
  "ek_allg_leistung_unternehmen_technisch_": 1,
  "ek_allg_leistung_unternehmen_organisatorisch_": 1,
```


Abbildung 35 Ausschnitt aus transformiertem Resultat

Abbildung 35 zeigt den transformierten Eintrag der Meldung 790445 aus Abbildung 34. Die einzelnen Tags wurden für die einfachere Analyse in immer präzisere Kategorien aufgeteilt, wobei die einzelnen Kategorien mit *_* getrennt werden. So gehört der Tag *ek_allg_ref_unternehmen* zu der Kategorie Eignungskriterien, genauer den Allgemeinen Kriterien, genauer der Kategorie Referenzen des Unternehmens. Die Kategorisierung wurde so konzipiert, damit alle *ek_allg_ref_unternehmen* auch zu den Eignungskriterien gezählt werden. In 5.3 und 5.4 wird genauer auf die Tags zu den Eignungs- und Zuschlagskriterien eingegangen.

5.3 Untersuchung Eignungskriterien

Die durchschnittliche Abdeckung der Texte der Eignungskriterien beträgt 55.7% bei 2000 untersuchten Meldungstexten und 107 Abfragen. Es wurden nur 2000 untersucht, um die Rechenzeit für den Algorithmus tief zu halten.
In 5.3.1 werden die ermittelten Abfragen diskutiert und in 5.3.2 wird das Resultat analysiert.

5.3.1 Kategorisierung Eignungskriterien

Die ermittelten Abfragen wurden in verschiedene Gruppen und Kategorien, basierend auf ihrem Inhalt, hierarchisch aufgeteilt. Auf dem höchsten Level wird zwischen Eignungskriterien *ek*, dem Verzicht auf Eignungskriterien *kk* und unrelevanten und sonstigen Informationen *xx* unterschieden. Es wurde insgesamt nur eine Passage in *kk* gefunden: In der Meldung 836757 wurde auf die Eignungskriterien verzichtet.

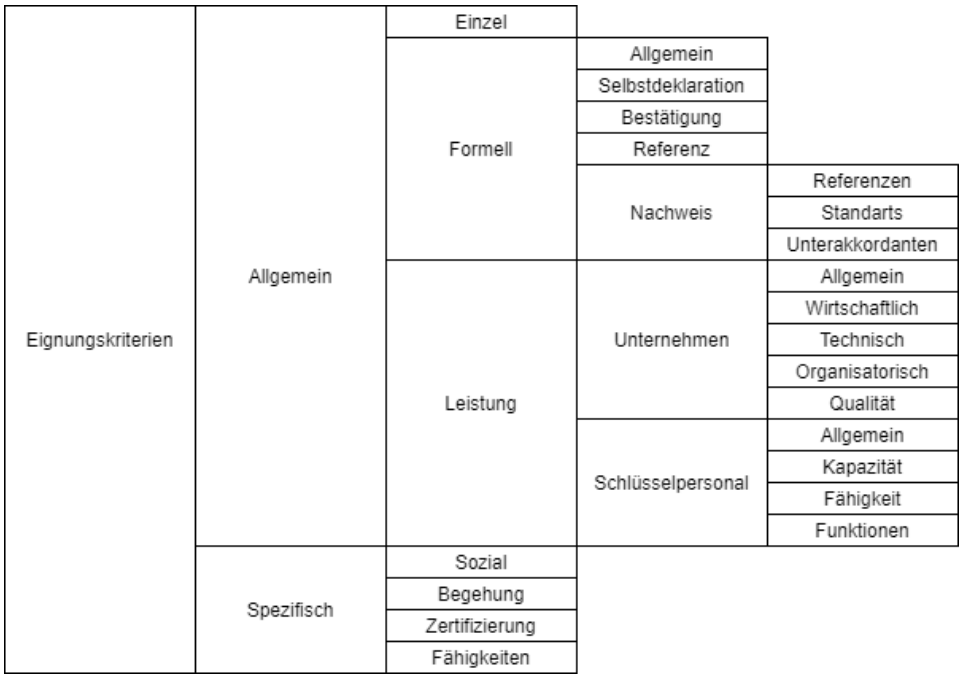


Abbildung 36 Kategorisierung der Eignungskriterien

Abbildung 36 zeigt die Kategorisierung der Eignungskriterien. Bei den Eignungskriterien wird zwischen allgemeinen und spezifischen EK unterschieden. Allgemeine EK wie zum Beispiel *wirtschaftliche / finanzielle Leistungsfähigkeit* müssen präzisiert werden, damit sie verwendet werden können, spezifische EK wie *ein Umweltmanagementsystem nach ISO 14001* sind genügend präzise um als Ausschlusskriterium zu dienen. Bei den Allgemeinen Kriterien wurde zwischen formellen Kriterien wie *Vollständiges und fristgerecht eingereichtes Angebot* und leistungsspezifischen Kriterien wie *Ausreichende personelle Ressourcen* unterschieden, dazu kommen die Einzelkriterien, wo nur einzelne Worte wie *Firmenkultur* definiert werden.

Die leistungsspezifischen Eignungskriterien wurde aufgeteilt in Unternehmensspezifische und Eignungskriterien bezüglich Schlüsselpersonal. Die Unternehmensspezifischen Eignungskriterien konnten wiederum in fünf

Kategorien aufgeteilt werden. Die untenstehende Auflistung zeigt jeweils den Namen der Kategorie sowie ein Beispiel:

- Allgemein (*Hinreichende Befähigung zur Auftragserfüllung*);
- Wirtschaftlich (*Angemessenes Verhältnis von Auftragssumme pro Jahr zum Umsatz der massgebenden Unternehmenseinheit*);
- Technisch (*Fachliche Leistungsfähigkeit der Unternehmung*);
- Organisatorisch (*Gleicher Projektleiter und bauleitender Vorarbeiter für gesamte Bauphase*) ;
- Qualitätsbezogen (*Qualitätssicherung mittels unternehmerbezogenem Qualitätsmanagementsystem*)

Die Eignungskriterien bezüglich des Schlüsselpersonals können in vier Kategorien aufgeteilt werden:

- Allgemein (*Schlüsselpersonen gelten Personen, welche im Projekt folgende Funktionen ausüben sollen:*);
- Kapazität (*Verfügbarkeit Projektleiter und Projektleiter Stv.*);
- Fähigkeit (*Erfahrung Schlüsselpersonen*);
- Funktionen (*Fachspezialist Trasse/Umwelt (TPL Trasse/Umwelt)*) ;

Die formellen Kriterien können in fünf Kategorien aufgeteilt werden, wobei die Kategorie Nachweis drei Unterkategorien hat. Alle fünf Kategorien definieren Produkte, welche der Anbieter mit dem Angebot abgeben muss, oder referenzieren auf weitere Eignungskriterien:

- Allgemein (*Vollständigkeit des Angebots.)*
- Deklaration (*Vollständig ausgefüllte Selbstdeklaration*)
- Bestätigung (*Bestätigung über die Termineinhaltung in schriftlicher Form.*)
- Nachweis:
    - Referenzen (*Nachweis Firmenreferenzen*)
    - Einhaltung Standards (*Nachweis (GAV) und Bestätigung des Anbietenden über die Einhaltung der Arbeitsbedingungen und Arbeitsschutzbestimmungen.*)
    - Unterakkordanten (*Nachweis allfälliger Unterakkordanten*)
- Referenz (*Erfüllung aller MUSS-Anforderungen gemäss Vorgabe Anforderungskatalog Version 2.0 vom 8. November 2017.*)

Bei den spezifischen Eignungskriterien wurden insgesamt vier Kategorien gefunden.

- Sozial (*Lohngleichheit von Mann und Frau*)
- Begehung (*Teilnahme einer autorisierten Vertretung des Anbietenden an der obligatorischen Begehung. Nachweis mittels Eintrags in Präsenzliste.*)
- Zertifizierungen (*QMS und/oder UMS ISO 9001/14001*)
- Fähigkeiten (*Mindestens 3 Jahre Erfahrung in der Bearbeitung von Metall im Heissprägeverfahren*)

<table>
<tr><td rowspan="5">Sonstige</td><td rowspan="2">Aufruf</td><td>Einfach</td></tr>
<tr><td>Mit Referenz</td></tr>
<tr><td rowspan="2">Referenz</td><td>Auf Meldung</td></tr>
<tr><td>Auf Unterlagen</td></tr>
<tr><td>Information</td></tr>
</table>

Abbildung 37 Kategorisierung sonstige Abfragen

Die nicht relevanten und sonstigen Passagen sind einfacher zu kategorisieren. Entweder ist es ein Aufruf zur Angebotsabgabe mit oder ohne Referenz auf die Unterlagen oder die geforderten Nachweise, oder es ist eine Referenz ohne Aufruf. Die Kategorie Information enthält Passagen, die einen informativen Charakter ohne Aufruf und Referenz haben. Die untenstehende Liste zeigt einige Beispiele für nicht relevante Abfragen:

- Aufruf: Einfach (*Alle wirtschaftlich und technisch leistungsfähigen Firmen, die zudem die nachfolgenden Eignungsnachweise erbringen, sind aufgerufen, ein Angebot in CHF zu unterbreiten.)*
- Aufruf: Referenz (*Alle wirtschaftlich und technisch leistungsfähigen Firmen, welche die nachfolgenden Eignungskriterien bzw. -nachweise gemäss Ziffer 3.8 erfüllen, können einen Teilnahmeantrag einreichen.*)
- Referenz: Meldung (*gemäss Kapitel 3.8*)
- Referenz: Unterlagen (*Gemäss Ausschreibungsunterlagen*)
- Info (*Alle Eignungskriterien müssen erfüllt werden.*)

5.3.2 Auswertung Eignungskriterien

Die höchste Abdeckung eines Textes beträgt 100% und wurde von 610 Meldungstexten erreicht, der längste Text mit 100% Abdeckung stammt von der Meldung 980043 und hat 422 Zeichen. Die niedrigste Abdeckung ist 0% und

wurde von 174 Texten erreicht, der längste Text mit 0% Abdeckung stammt von der Meldung 881123 und hat eine Länge von 3400 Zeichen. Abbildung 38 zeigt die beiden Histogramme bezgl. der Länge der Meldungen und der Abdeckung. Es zeigt sich, dass die meisten Texte kurz sind und die Abdeckung entweder gering oder sehr hoch ist.

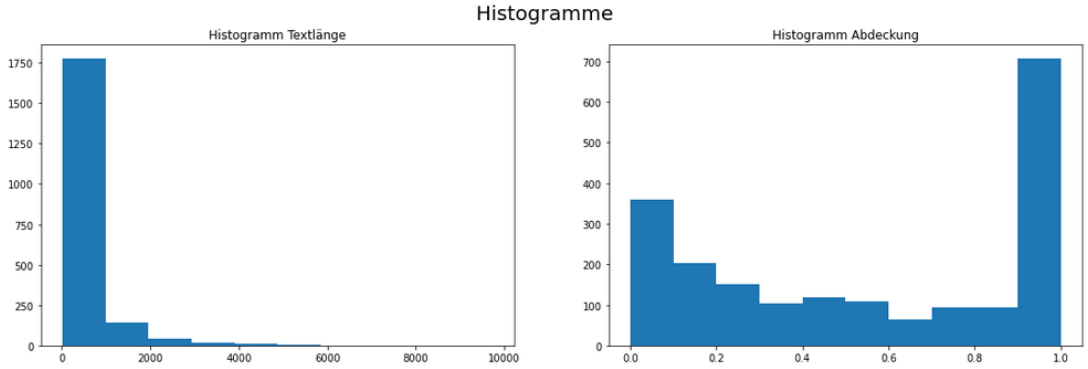


Abbildung 38 Histogramme für Länge und Abdeckung

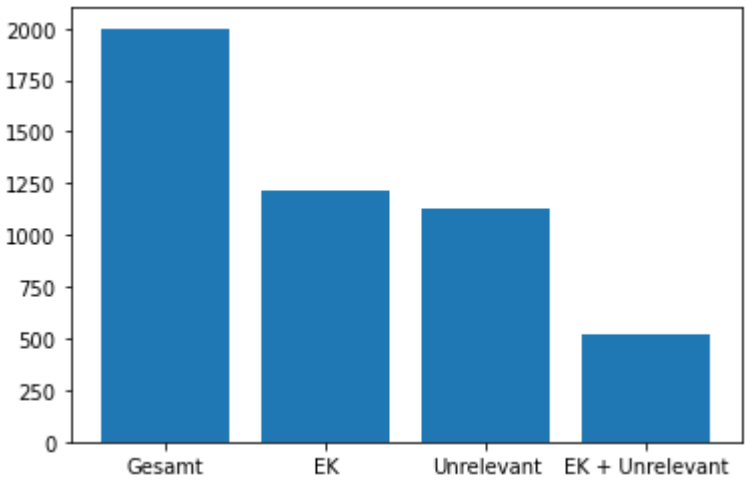


Abbildung 39 Anteil der Gruppen an der Gesamtmenge der Texte

In der Stichprobe von 2000 Meldungstexten wurden 1213 Texte mit insgesamt 2942 identifizierten Passagen mit Eignungskriterien und 1131 Meldungen mit insgesamt 1629 nicht relevanten Passagen gefunden. 519 Texte haben sowohl Passagen mit EK als auch Passagen ohne EK. Die Anteile sind in Abbildung 39 als Balkendiagramme dargestellt. Interessant ist, dass die Menge der Texte sowohl mit EK als auch mit Passagen ohne EK nur etwa die Hälfte der anderen beiden Mengen ausmacht.

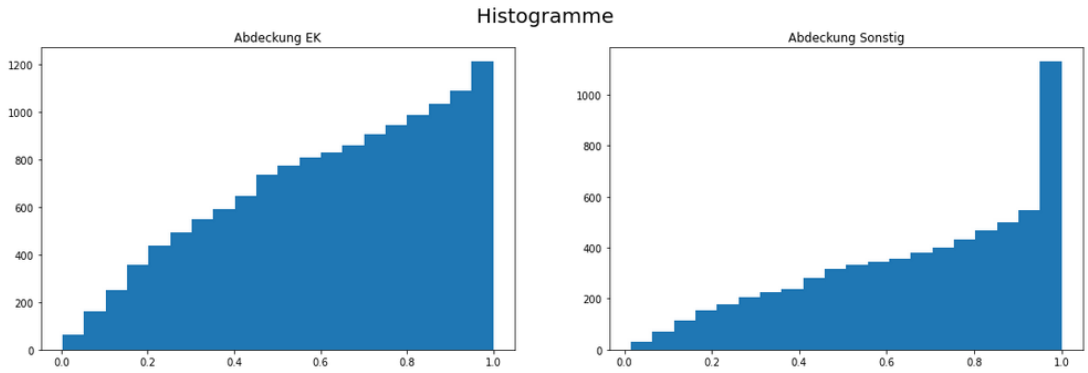


Abbildung 40 Kumulative Histogramme der Abdeckung von EK und sonstig

Abbildung 40 zeigt die kumulierten Histogramme für die 1213 Texte mit EK und 1131 Texte mit Passagen ohne EK. Es fällt auf, dass die Abdeckung der Texte mit Passagen ohne Eignungskriterien rechtsschief ist. Dies kommt unter anderem durch den hohen Anteil an Referenzen auf das *Kapitel 3.8*: von den 549 Texten mit einer Abdeckung von 1 und sonstigen Informationen enthalten über die Hälfte den Text *gemäss Kapitel 3.8* oder etwas ähnliches. 13,9% der 2000 Texte referenzieren ausschliesslich auf die geforderten Nachweise.

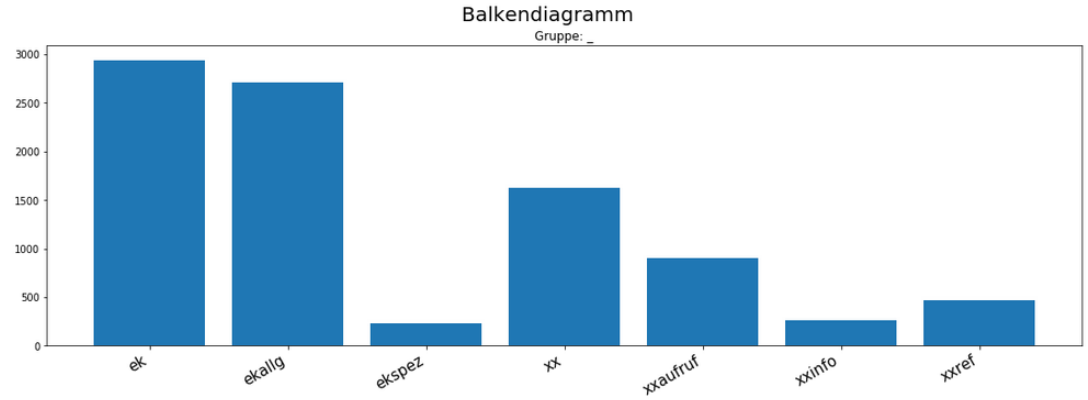


Abbildung 41 Anteil der einzelnen Kategorien

Abbildung 41 zeigt die Anzahl der EK und der nicht relevanten Passagen, hier mit xx deklariert. Bei den EK machen die allgemeinen EK mit 2711 der 2942 identifizierten Passagen den grössten Anteil aus. Spezifische Eignungskriterien wurden nur in 231 Passagen gefunden. Von den 1629 nicht relevanten Passagen entfallen 899 auf einen Aufruf, 473 auf eine Referenz und 257 auf eine Information. Mit 311 Passagen wurde fast doppelt so häufig auf die geforderten Nachweise verwiesen als auf die Unterlagen mit 162 Passagen. Allerdings gibt es für den Verweis auf die Unterlagen das XML-Attribut

DOCUMENTS, welches ausgewählt werden kann und wie in 2.3 dargestellt, in 57’349 Ausschreibungen ausgewählt wurde.

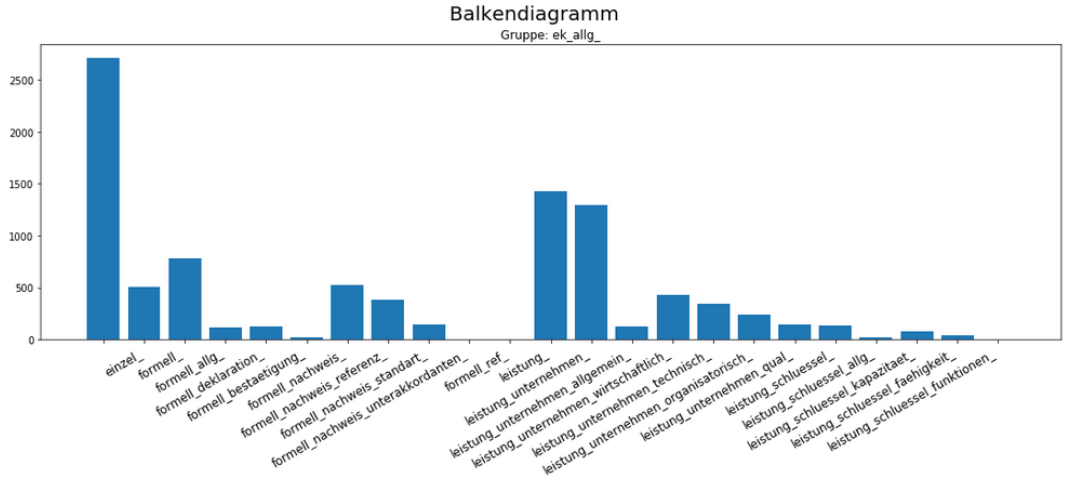


Abbildung 42 Einsatz der Allgemeinen Eignungskriterien

Wie in Abbildung 42 zu sehen ist, kommen mit 1428 Passagen die leistungsbezogene Eignungskriterien am meisten vor. Der grösste Anteil mit 1291 Passagen haben EK der Unterkategorie Leistung-Unternehmen. Formelle Eignungskriterien kommen mit 782 Passagen am zweithäufigsten vor, der grösste Anteil hat die Unterkategorie Formell_Nachweis mit 528 Passagen. 501 Passagen gehören zu der Kategorie Einzel. Tabelle 4 enthält die häufigsten Texte pro Kategorie und Unterkategorie. Bei der Unterkategorie Leistung-Unternehmen führen die Eignungskriterien, welche die wirtschaftliche Leistungsfähigkeit prüfen mit 433 Passagen, gefolgt von technischer und organisatorischer Leistungsfähigkeit mit 347 und 241 Passagen. Bei der Unterkategorie Formell_Nachweis kommen die Eignungskriterien, welche Referenzen verlangen, mit 382 Passagen mehr als doppelt so häufig vor, wie die Eignungskriterien, welche die Einhaltung von Standards verlangen. Tabelle 4 zeigt jeweils die am häufigsten vorkommenden Abfragen.

Tabelle 4 Häufigste Abfragen Allgemeinen EK

Kategorie	Anzahl	Text
Allgemein Einzel	250	*Leistungsfähigkeit*
Formell Nachweis Referenz	285	*Referenzen über die Ausführung von mindestens 3 mit der vorgesehenen Aufgabe*

		vergleichbaren Projekten in den letzten 5 Jahren.
Formell Nachweis Standard	65	*Nachweis (GAV) und Bestätigung des Anbietenden über die Einhaltung der Arbeitsbedingungen und Arbeitsschutzbestimmungen.*
Leistung Unternehmen wirtschaftlich	259	*wirtschaftliche/finanzielle Leistungsfähigkeit*
Leistung Unternehmen technisch	214	*technische Leistungsfähigkeit*
Leistung Unternehmen organisatorisch	100	*Ausreichende personelle Ressourcen*

5.4 Untersuchung Zuschlagskriterien

Bei den Zuschlagskriterien wird mit 95 Abfragen und einer Stichprobe von 2000 Meldungstexten eine durchschnittliche Abdeckung von 51% erreicht. Um diese Abdeckung zu erreichen, wurde im Algorithmus die Option *accept_single_token_candidates* auf Wahr gesetzt. Diese Option ermöglicht, dass auch einzelne Worte wie *Preis* als gültiger Kandidat akzeptiert werden.

5.4.1 Kategorisierung Zuschlagskriterien

Die Zuschlagskriterien können ähnlich den Eignungskriterien in 5.3.1 aufgebaut werden. Auf der höchsten Ebene wird zwischen Passagen mit

Zuschlagskriterien und Passagen ohne Zuschlagskriterien unterschieden.

Zuschlagskriterien	Unternehmen	Referenzen
		Organisation
		Faehigkeiten
		Allgemein
		Nachhaltigkeit
	Angebot	Preis
		Auftrag
		Qualität
		Allgemein
Keine Zuschlagskriterien	Referenz	
	Bewertung	definition
		schema

Abbildung 43 Kategorisierung der Passagen in den Texten über die Zuschlagskriterien

Die Zuschlagskriterien können in zwei Kategorien aufgeteilt werden: unternehmensbezogene und angebotsbezogene Zuschlagskriterien. Die Unternehmensbezogenen Zuschlagskriterien wie zum Beispiel *Vorhandene Kapazität im Unternehmen* bewerten das Unternehmen an sich, während die angebotsbezogenen Zuschlagskriterien wie *Korrigierte Offertensumme* das Angebot bewerten. Die unternehmensspezifischen Zuschlagskriterien können in fünf Unterkategorien aufgeteilt werden. Die Auflistung zeigt wie in 5.3.1 den Namen der Kategorie und ein Beispiel:

- Referenzen (*Erfahrungsnachweis der eingesetzten Schlüsselperson)*
- Organisation (*Personal und Organisation)*
- Fähigkeiten (*Technische Qualitäten)*
- Allgemein (*Institution und Infrastruktur)*
- Nachhaltigkeit (*Ausbildung von Lernenden)*

Die angebotsspezifischen Zuschlagskriterien werden in vier Unterkategorien aufgeteilt:

- Preis (*Beschaffungspreis)*
- Auftrag (*Analyse der Ausgangslage)*
- Qualität (*Inhalt Qualität der eingereichten Unterlagen)*
- Allgemein (*Lieferfrist für Lieferungen)*

Die Passagen ohne Zuschlagskriterien enthalten keine Informationen über Zuschlagskriterien. Allerdings können sie Informationen und Definitionen über die Bewertung von Zuschlagskriterien oder ein konkretes Bewertungsschema

enthalten. Weiter gibt es Passagen, die auf die Ausschreibungsunterlagen oder einen anderen Teil in dem Meldungstext verweisen. Die Auflistung zeigt den Namen und ein Beispiel:

- Referenzen (*Siehe Ausschreibungsunterlagen)*
- Bewertung
    - Definition (*Die Qualität wird anhand der folgenden Kriterien bewertet)*
    - Schema (*Die Bewertung erfolgt immer mit Noten von 0 bis 5:)*

5.4.2 Auswertung Zuschlagskriterien

Die höchste Abdeckung eines Textes beträgt 100% und wurde von 392 Meldungstexten erreicht, der längste Text mit 100% Abdeckung stammt von der Meldung 1062305 und hat eine Länge von 410 Zeichen. 68 Meldungstexte haben eine Abdeckung von 0% und den längsten Meldungstext mit 0% Abdeckung hat die Meldung 1086771 mit einer Länge von 3505 Zeichen. Diese Meldung ist in italienischer Sprache geschrieben und wurde als deutschsprachige Meldung erfasst.

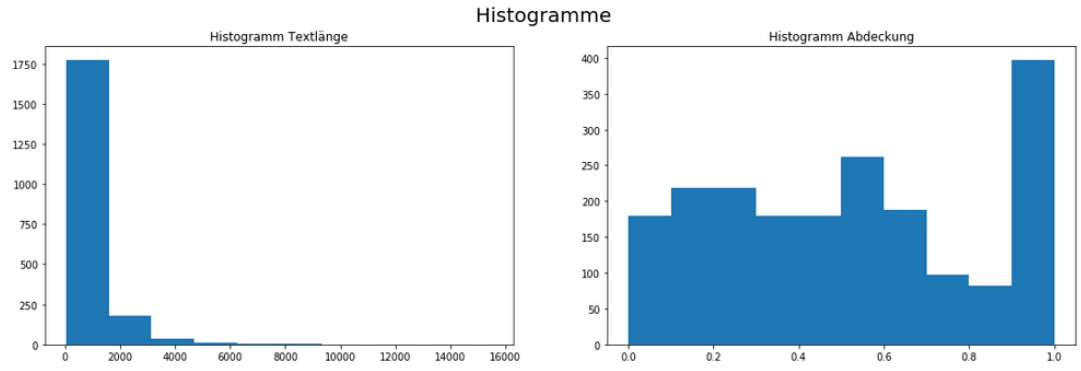


Abbildung 44 ZK: Histogramm für Länge und Abdeckung

Abbildung 44 zeigt, dass die Abdeckung bei den Zuschlagskriterien gleichmässiger verteilt ist als die der Eignungskriterien. Der Anteil der Meldungen mit einer Abdeckung von über 90% ist mit knapp 400 deutlich tiefer als der bei den EK mit 700.

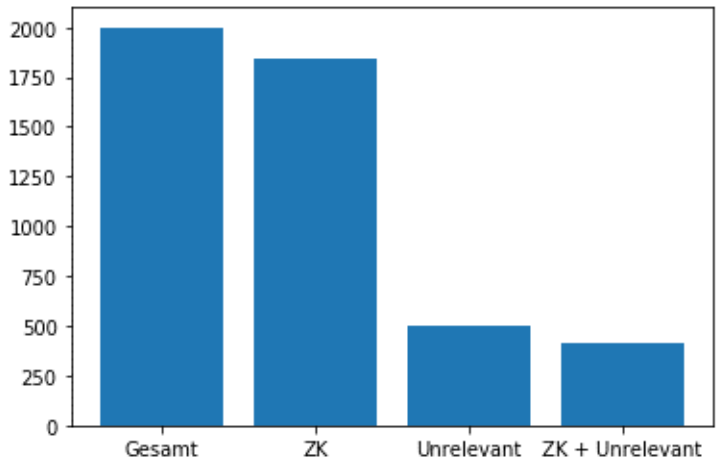


Abbildung 45 ZK: Anteile der Gruppen an der Gesamtmenge der Texte

Ebenfalls zeigen die Anteile der Passagen mit ZK und der Passagen mit unrelevanten Informationen in Abbildung 45 ein anderes Bild: 1838 Meldungstexte enthalten ein oder mehrere identifizierte Passagen über ZK und 504 Meldungstexte enthalten identifizierte Passagen ohne relevante Informationen über ZK. Von den 504 Meldungstexten enthalten 409 sowohl Passagen mit ZK und Passagen ohne ZK. Insgesamt gibt es 4976 Passagen mit Informationen zu ZK und 2851 Passagen ohne Informationen zu ZK.

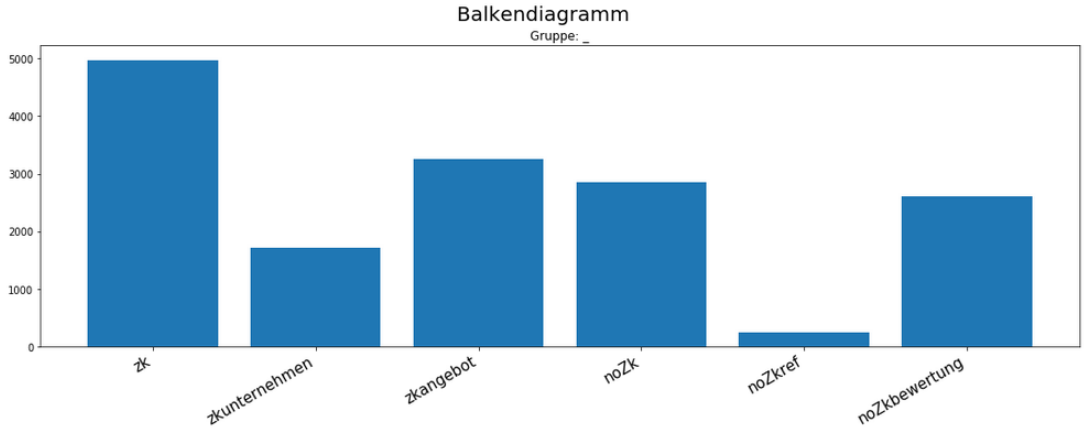


Abbildung 46 ZK: Anteil der einzelnen Kategorien

3260 der 4976 Passagen mit Zuschlagskriterien sind angebotsspezifische Zuschlagskriterien und 1716 Passagen gehören zu den unternehmensspezifischen Kriterien. Bei den Passagen ohne Zuschlagskriterien gibt es 245 Passagen mit Referenz auf die Unterlagen oder einen anderen Teil der Meldung und 2606 Passagen mit Informationen über die Bewertung. In 88 Meldungstexten konnten ausschliesslich Passagen mit einer Referenz identifiziert werden.

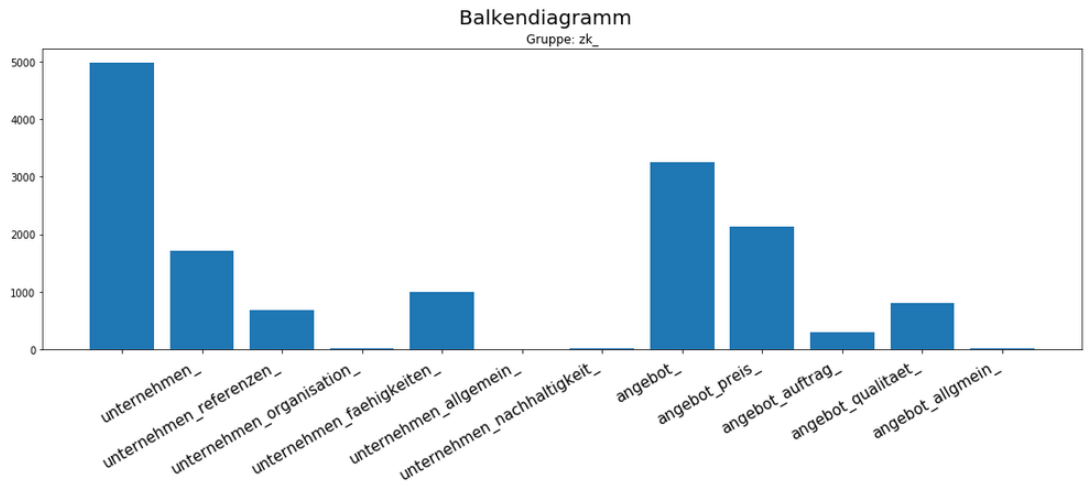


Abbildung 47 Anteil der Unterkategorien in den Zuschlagskriterien

In den 3260 angebotsspezifischen Zuschlagskriterien kommen Passagen aus der Kategorie Preis mit 2134 identifizierten Passagen am häufigsten vor. Die Abfrage mit den meisten Ergebnissen in der Kategorie ist *Preis* mit 1433 identifizierten Passagen. Es gibt viele Passagen mit dem Wort *Preis* da dieses Wort auch häufig in den Bewertungskriterien vorkommt. Die Abfrage mit den zweitmeisten Ergebnissen ist *Angebotspreis,* welche mit 350 identifizierten Passagen deutlich weniger häufig vorkommt. Die Kategorie Qualität kommt mit 806 identifizierten Passagen am zweithäufigsten vor. Die Abfrage *Qualität* macht 589 identifizierten Passagen der grösste Anteil aus und kommt ebenfalls in den Bewertungskriterien vor.

In den 1716 unternehmensspezifischen Zuschlagskriterien gibt es 1002 Passagen aus der Kategorie Fähigkeiten und 682 Passagen aus der Kategorie Referenzen. Die Kategorien Allgemein, Nachhaltigkeit und Organisation kommen mit 3, 14 und 15 Passagen selten vor. In der Kategorie Fähigkeiten gibt es 318 Passagen aus der Abfrage *Erfahrung* und 308 Passagen aus der Abfrage *Schlüsselpersonen*. Bei der Kategorie Referenzen kommt die Abfrage *Referenzen* mit 525 identifizierten Passagen am häufigsten vor.

Von den 2851 identifizierten Passagen ohne Informationen über ZK stammen 2606 aus der Kategorie Bewertung. Die Unterkategorie Schema hat dabei mit 2296 Passagen einen deutlich grösseren Anteil. Die Abfrage *Nicht beurteilbar; keine Angabe* kommt mit 797 identifizierten Passagen am häufigsten vor. Die Unterkategorie Definition hat 310 einen geringeren Anteil. 174 der 310

identifizierten Passagen stammen von der Abfrage *Die Bewertung erfolgt immer mit Noten von 0 bis 5:*.

Die übrigen 245 identifizierten Passagen ohne Information stammen aus der Kategorie Referenz. Davon kommen 114 aus der Abfrage *Submissionsunterlagen.*

5.5 Untersuchung Ausschreibungsunterlagen

Von den in 5.3.1 ermittelten 107 Abfragen können die 87 für die Eignungskriterien relevanten Abfragen in einer ElasticSearch-Bool Query auf die indexierten Ausschreibungsunterlagen angewendet werden, um Dokumente mit einer hohen Ähnlichkeit zu den Abfragen zu finden. Elasticsearch verwendet für das Scoring unter anderem die TFIDF-Darstellung eines Dokumentes(*TFIDFSimilarity (Lucene 7.6.0 API)*, o. J.) Die höchste Score des Suchresultats hat das Dokument '*/170486/bkp_388_ausschreibungsunterlagen_renamed/gehege_ausschreibung_w_os.docx*' mit 352.10 Punkten. Insgesamt wurden 62'001 Dateien in 16'918 Projekten gefunden. Um die Menge an gefundenen Suchergebnissen zu reduzieren werden Schwellenwerte für die minimale Score definiert. Abbildung 48 zeigt die Entwicklung der Anzahl Resultate mit abnehmendem Schwellenwert. Der minimale Schwellenwert wird jeweils relativ zu dem maximalen Score berechnet.

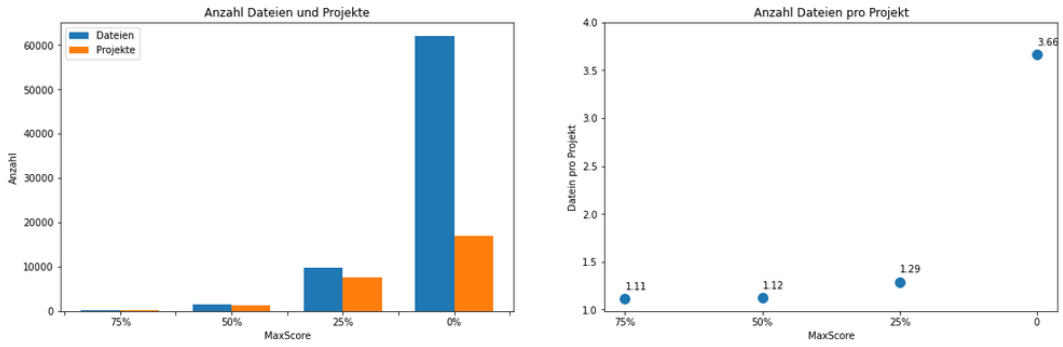


Abbildung 48 Gefundene Dateien und Projekte und Anzahl Dateien pro Projekt für verschiedene Schwellenwerte

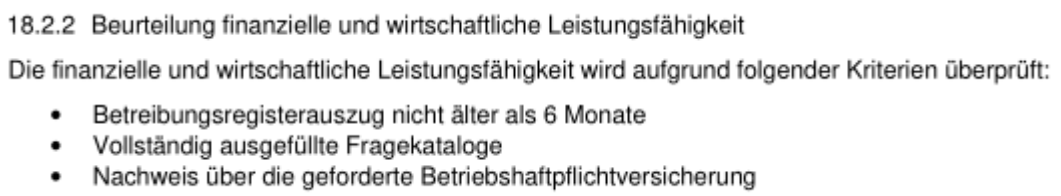
Abbildung 48 zeigt, dass mit einem Schwellenwert meistens ein oder zwei relevante Dokumente pro Projekt identifiziert werden. Eine Minimale Score von 264(*75%*max_score*) liefert 237 Dateien in 264 Projekten.

Die Kombination von Abfragen und einer Suche nach relevanten Dokumenten mit Elasticsearch und einem minimalen Schwellenwert ermöglicht ein zielgerichtetes Durchsuchen der Ausschreibungsunterlagen.

6 Diskussion

6.1 Zusammenfassung und Diskussion

In dieser Arbeit wurde in drei Iterationen ein Algorithmus entwickelt, um in einem Corpus Passagen, welche mit einer Abfrage übereinstimmen zu identifizieren. Mit dem Algorithmus können iterativ neue Abfragen definiert und damit noch mehr Passagen identifiziert werden. Der Algorithmus wurde auf eine Stichprobe der Texte zu Eignungskriterien in den Simap-Meldungen angewendet. In mehreren Iterationsdurchgängen konnten 107 Abfragen ermittelt werden, mit denen eine durchschnittliche Abdeckung von 55,7% erreicht wurde. Die Auswertung des Resultates zeigt, dass in 278 der Texte ausschliesslich auf die geforderten Nachweise verwiesen wird und das häufigste Eignungskriterium die wirtschaftliche Leistungsfähigkeit ist. 92% der 2942 identifizierten Eignungskriterien wurden allgemein formuliert und müssen in den Ausschreibungsunterlagen noch weiter ausgeführt werden. Anhand der Passage *wirtschaftliche/finanzielle Leistungsfähigkeit,* welche 259-mal vorkommt, kann noch kein objektives Ausschlusskriterium definiert werden. Dazu braucht es eine genauere Definition, wie das die Stadtpolizei Stadt Zürich in 18.2.2 der Datei *teil_a_ausschreibungsbestimmungen.pdf* zeigt.



18.2.2 Beurteilung finanzielle und wirtschaftliche Leistungsfähigkeit

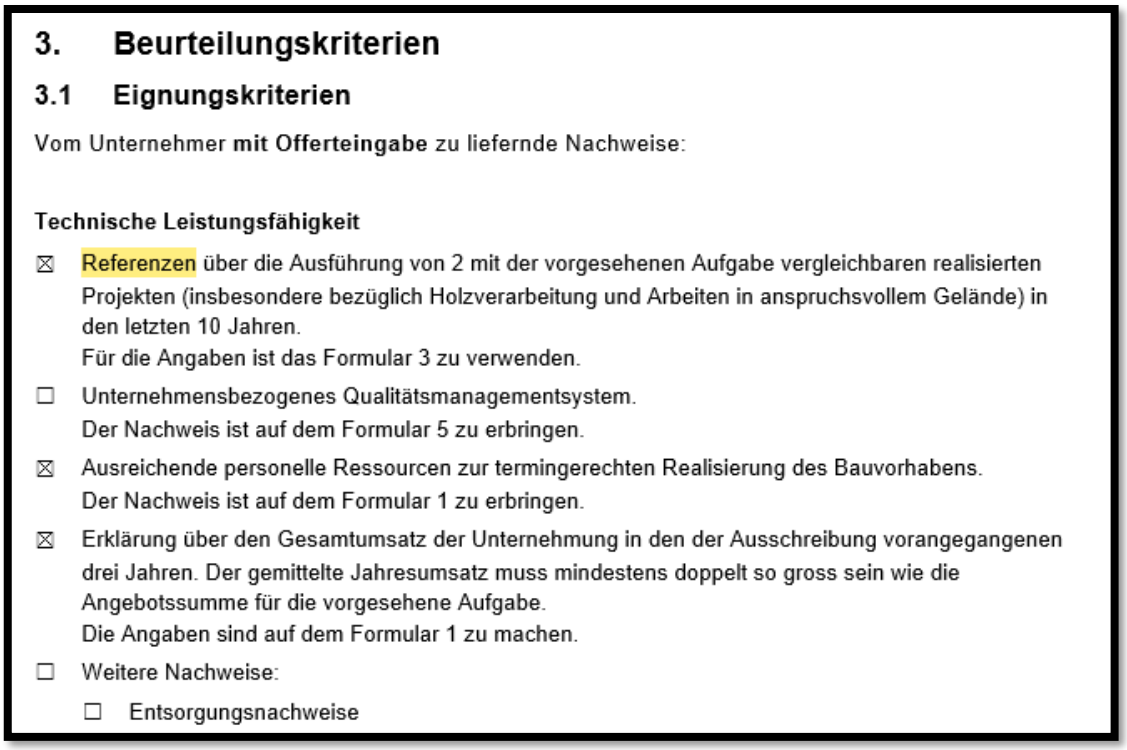
Die finanzielle und wirtschaftliche Leistungsfähigkeit wird aufgrund folgender Kriterien überprüft:

- Betreibungsregisterauszug nicht älter als 6 Monate
- Vollständig ausgefüllte Fragekataloge
- Nachweis über die geforderte Betriebshaftpflichtversicherung

Abbildung 49 18.2.2 Kriterien wirtschaftliche und finanzielle Leistungsfähigkeit(Stadtpolizei Zürich, 2019)

Mit den 87 relevanten der 107 Abfragen konnten in den Ausschreibungsunterlagen mit Elasticsearch 61’930 Dateien in 16’899 Projekten gefunden werden. In diesen Dateien werden wie in Abbildung 49 die Eignungskriterien präzisiert oder definiert. Die Abfragen ermöglichen es auch die Ausschreibungsunterlagen der Projekte, die keine Eignungskriterien in der Meldung angegeben haben, zu durchsuchen und so trotzdem Eignungskriterien für diese zu finden. So konnte für das Projekt 170486, welches in der Ausschreibung 1018615 auf die Ausschreibungsunterlagen verweist, das Dokument

gehege_ausschreibung_w_os.docx gefunden werden. Abbildung 50 zeigt einen relevanten Ausschnitt des Dokuments. Mit der Verwendung von Schwellenwerten in der Elasticsearch Query konnte ausserdem jeweils das relevanteste Dokument pro Projekt gefunden werden.



3. Beurteilungskriterien

3.1 Eignungskriterien

Vom Unternehmer **mit Offerteingabe** zu liefernde Nachweise:

Technische Leistungsfähigkeit

☒ Referenzen über die Ausführung von 2 mit der vorgesehenen Aufgabe vergleichbaren realisierten Projekten (insbesondere bezüglich Holzverarbeitung und Arbeiten in anspruchsvollem Gelände) in den letzten 10 Jahren.
Für die Angaben ist das Formular 3 zu verwenden.

☐ Unternehmensbezogenes Qualitätsmanagementsystem.
Der Nachweis ist auf dem Formular 5 zu erbringen.

☒ Ausreichende personelle Ressourcen zur termingerechten Realisierung des Bauvorhabens.
Der Nachweis ist auf dem Formular 1 zu erbringen.

☒ Erklärung über den Gesamtumsatz der Unternehmung in den der Ausschreibung vorangegangenen drei Jahren. Der gemittelte Jahresumsatz muss mindestens doppelt so gross sein wie die Angebotssumme für die vorgesehene Aufgabe.
Die Angaben sind auf dem Formular 1 zu machen.

☐ Weitere Nachweise:

☐ Entsorgungsnachweise

Abbildung 50 Ausschnitt: Teil der EK im Dokument(ETH Zürich Abteilung Immobilien, 2018)

Das in 1.3 skizzierte Vorgehen(Abbildung 8) konnte damit validiert werden. Die ermittelten Abfragen können nicht nur verwendet werden, um ähnliche Passagen in den Meldungstexten zu ermitteln, sondern auch um relevante Dateien in den Ausschreibungsunterlagen zu finden.

Die gewählte Vorgehensweise verwendet keine spezifischen ML-Techniken wie Klassifizierung oder Clustering, sondern setzt auf einen regelbasierten Algorithmus zur Identifikation der Passagen. Für weitere Informationen bezüglich regelbasierter Extraktion von Informationen kann (Chiticariu et al., 2013) konsultiert werden.

6.2 Ausblick

Für ein weiteres Vorgehen können verschiedene Richtungen eingeschlagen werden: Die Weiterentwicklung des Algorithmus, Anwendungen auf andere Textsammlungen und Erweiterung der bestehenden Abfragen und deren

Kategorisierung. Ausserdem kann der Algorithmus in Applikationen implementiert werden.

Der Algorithmus kann weiterentwickelt werden, um die manuelle Erstellung der Abfragen zu unterstützen. Zum Beispiel kann der Algorithmus so erweitert werden, dass er die häufigen Passagen selbst identifiziert und der Anwender nur ein Kategorisierungsschema erstellen muss. Eine andere Möglichkeit wäre die Definition von Mustern anstelle von konkreten Texten mit Regex oder einer Query Language in den Abfragen, damit die Anzahl der Abfragen geringgehalten werden kann. Weiter kann die Performance optimiert und die Benutzerfreundlichkeit des Algorithmus erhöht werden.

Die Anwendung des Algorithmus auf andere Textsammlungen bedingt kleinere Anpassungen am Skript, welches der Algorithmus aufruft und mit Daten versorgt. Anschliessend können die Nutzer ihre Abfragen definieren und auswerten.

Die Auswertung der in den Meldungstexten ermittelten Abfragen zu den Eignungskriterien in 5.3.1 zeigt, dass die durchschnittliche Abdeckung bei 55.7% liegt, und einige Abfragen nur selten in einer Passage identifiziert wurden. Die Passage ‘*Genügend Führungs- und Personalkapazität’*(ek_allg_leistung_unternehmen_organisatorisch_2) kommt nur einmal vor und hat damit einen geringen Einfluss auf die Abdeckung. Weitere Arbeiten könnten die Abfragen neu definieren und auch die Kategorisierung überarbeiten. Es wäre spannend herauszufinden, ob es noch weitere spezifische Eignungskriterien gibt, welche noch nicht entdeckt wurden. Zusätzlich können auch die Unterschiede zwischen den einzelnen Auftraggebern oder die Entwicklung der Eignungskriterien über die Zeit untersucht werden.

Schlussendlich kann der Algorithmus in verschiedene Applikationen implementiert werden. Die Simap-Meldungen können mit weiteren Informationen versehen und ausgewertet werden. Damit können neue Erkenntnisse gewonnen werden. Ein Beispiel: Auftraggeber verwenden auf Gemeindelevel häufiger EKs der Kategorie A als andere Auftraggeber. Eine weitere Implementierung wäre das Durchsuchen der Meldungstexte und Ausschreibungsunterlagen auf bestimmte Formulierungen und Zertifikate, um das Nachhaltigkeitsmonitoring (Welz & Stuermer, 2020) zu automatisieren, oder sogar den Auftraggebern eine Möglichkeit zu geben ihre Dokumente auf Vollständigkeit zu

prüfen. Weiter kann der Algorithmus für das Labeling von Dokumenten verwendet werden, um so Trainingsdaten für andere Machine-Learning Applikationen zu generieren.

Anhang

Dank

Ich bedanke mich bei Matthias Stürmer für die Betreuung der Masterarbeit und bei der Forschungsstelle Digitale Nachhaltigkeit für den Zugriff zu den Simap-Daten. Weiter gilt mein Dank Inge, Lea und Jürg Schweizer für diverse Rechtschreibkorrekturen und Verständnisfragen.

Code

Der Code für diese Masterarbeit ist unter https://github.com/UniDomi/ir-in-simap-notices zu finden.

Abbildungsverzeichnis

Abbildung 1 Beschaffungsstatistik.ch: Trend-Ausschreibungen 4
Abbildung 2 XML der Meldung 980029 mit markierten EK 7
Abbildung 3 Anteil Elemente mit Text zu EK, ZK und GN in Meldungen 7
Abbildung 4 Text des EK Elements der Meldung 972723 mit identifizierten Passagen. 9
Abbildung 5 Ausschnitt aus Ausschreibungsbestimmung.pdf zeigt die Eignungskriterien 9
Abbildung 6 Schematische Darstellung Algorithmus 10
Abbildung 7 Ablauf einer Iteration, um neue Passagen zu identifizieren. 11
Abbildung 8 Schematische Darstellung Abfrage in Elasticsearch 11
Abbildung 9 Strukturierter Datensatz 13
Abbildung 10 Resultat des Algorithmus aus 4.2 als JSON 14
Abbildung 11 Boxplot Anzahl publizierter Meldungen pro Tag, Woche und Monat 15
Abbildung 12 Anzahl Meldungen pro Wochentag als Median und Durchschnitt 16
Abbildung 13 Entwicklung der Anzahl der Meldungen pro Woche und pro Monat 17
Abbildung 14 Verteilung der Meldungen nach Typ und Sprache 18

Abbildung 15 Teilgebiete und Überschneidungen von Text Mining („The Seven Practice Areas of Text Analytics“, 2012) 20
Abbildung 16 Darstellung Preprocessing; Vektorisierung und Feature Selection 23
Abbildung 17 BoW: Term-Count Darstellung 24
Abbildung 18 Normalisierte Vektoren: Term Frequency Darstellung(L2) 25
Abbildung 19 TF-IDF: Term Frequency- Inverse Document Frequency 25
Abbildung 20 IR System (Bassil, 2012) 26
Abbildung 21 Berechnung der Ähnlichkeit(Similarity) mit dem Skalarprodukt der beiden Vektoren 27
Abbildung 22 Ähnlichkeits-Matrix der in Abbildung 16 definierten Texte 27
Abbildung 23 Text des Elements EK der Meldung 980029 28
Abbildung 24 Text des Elements EK der Meldung 980029 mit identifizierten Passagen 28
Abbildung 25 Text, tokenisiert, lemmatisiert und als Wortart 30
Abbildung 26 Beispiel einer Abfrage als Input 31
Abbildung 27 Text mit gefilterten Tokens 33
Abbildung 28 Identifizierung Passage anhand Tokens und Textstruktur 34
Abbildung 29 Beispiel für eine Erweiterung basierend auf der Struktur des Textes, die Übereinstimmung organisatorische Leistungsfähigkeit wird erweitert. 35
Abbildung 30 Abfrage mit übereinstimmenden N-Grammen 36
Abbildung 31 Darstellung Moving Frame 37
Abbildung 32 Gewichtung der Tokens 37
Abbildung 33 Beispiel einer Abfrage 39
Abbildung 34 Getaggter Text der EK aus Meldung 790455 40
Abbildung 35 Ausschnitt aus transformiertem Resultat 42
Abbildung 36 Kategorisierung der Eignungskriterien 43
Abbildung 37 Kategorisierung sonstige Abfragen 45
Abbildung 38 Histogramme für Länge und Abdeckung 46
Abbildung 39 Anteil der Gruppen an der Gesamtmenge der Texte 46
Abbildung 40 Kumulative Histogramme der Abdeckung von EK und sonstig 47
Abbildung 41 Anteil der einzelnen Kategorien 47

Abbildung 42 Einsatz der Allgemeinen Eignungskriterien 48
Abbildung 43 Kategorisierung der Passagen in den Texten über die Zuschlagskriterien 50
Abbildung 44 ZK: Histogramm für Länge und Abdeckung 51
Abbildung 45 ZK: Anteile der Gruppen an der Gesamtmenge der Texte 52
Abbildung 46 ZK: Anteil der einzelnen Kategorien 52
Abbildung 47 Anteil der Unterkategorien in den Zuschlagskriterien 53
Abbildung 48 Gefundene Dateien und Projekte und Anzahl Dateien pro Projekt für verschiedene Schwellenwerte 54
Abbildung 49 18.2.2 Kriterien wirtschaftliche und finanzielle Leistungsfähigkeit(Stadtpolizei Zürich, 2019)..... 56
Abbildung 50 Ausschnitt: Teil der EK im Dokument(ETH Zürich Abteilung Immobilien, 2018) 57

Tabellenverzeichnis

Tabelle 1: Meldungstypen mit Beschreibung und Kategorie 17
Tabelle 2 Teilgebiete von Text Mining nach(Allahyari et al., 2017) 21
Tabelle 3: Relevante Felder des Resultates 40
Tabelle 4 Häufigste Abfragen Allgemeinen EK 48

Literaturverzeichnis

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *arXiv:1707.02919 [cs]*. http://arxiv.org/abs/1707.02919

Bassil, Y. (2012). A Survey on Information Retrieval, Text Categorization, and Web Crawling. *arXiv:1212.2065 [cs]*. http://arxiv.org/abs/1212.2065

BBL. (2019). *Faktenblatt_SBez Auswertung_2018_DE_V1.pdf*.

Beschaffungsstatistik.ch. (2021, April 18). https://beschaffungsstatistik.ch/uebersicht

Chen, S. (2020, Mai 26). *Getting Started with Text Vectorization*. Medium. https://towardsdatascience.com/getting-started-with-text-vectorization-2f2efbec6685

Chiticariu, L., Li, Y., & Reiss, F. R. (2013). *Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!* 6.

Difflib—Helpers for computing deltas—Python 3.9.4 documentation. (o. J.). Abgerufen 18. April 2021, von https://docs.python.org/3/library/difflib.html

Dridan, R., & Oepen, S. (2012). *Tokenization: Returning to a Long Solved Problem A Survey, Contrastive Experiment, Recommendations, and Toolkit*. 5.

Elastic. (2021). *Was ist Elasticsearch?* Elastic. https://www.elastic.co/de/what-is/elasticsearch

Elasticsearch: Die offizielle Engine für verteilte Suche und Analytics. (o. J.). Elastic. Abgerufen 18. April 2021, von https://www.elastic.co/de/elasticsearch

Endtner, J. (2019). *Machine Learning in Ausschreibungsunterlagen*.

Endtner, J., & Stürmer, M. (2019). *Extraction of Suitability Criteria from Tender Documents Using Machine Learning*. https://doi.org/10.13140/RG.2.2.35360.12808

ETH Zürich Abteilung Immobilien (Hrsg.). (2018). *Ausschreibung Forschungsstation Früebüel, Umbau und Erweiterungen Neubau Gehegeanlage für Wild- und Nutztiere*.

explosion.ai. (2021). *SpaCy*. https://spacy.io/

Gunawan, R., Rahmatulloh, A., Darmawan, I., & Firdaus, F. (2019). *Comparison of Web Scraping Techniques: Regular Expression, HTML DOM and Xpath*. 283–287. https://doi.org/10.2991/icoiese-18.2019.50

Henrich, A. (2008). *Information Retrieval 1*. 421.

Hotho, A., Nurnberger, A., Paaß, G., & Augustin, S. (2005). *A Brief Survey of Text Mining*. 37.

IntelliProcure—Intelligence im öffentlichen Beschaffungswesen. (2018, August 31). https://intelliprocure.ch/dashboard

K. Dalal, M., & A. Zaveri, M. (2011). Automatic Text Classification: A Technical Review. *International Journal of Computer Applications*, *28*(2), 37–40. https://doi.org/10.5120/3358-4633

KPS Solutions. (2020a, März 12). *SOAP Schnittstelle von Simap*. simap.ch. https://www.simap.ch/DE/PDF/COMMON/soapservice_simap.pdf

KPS Solutions. (2020b). *SimapDtd.zip*. https://www.simap.ch/shab-forms/doc/dtd/SimapDtd.zip

Library Architecture · spaCy API Documentation. (o. J.). Library Architecture. Abgerufen 18. April 2021, von https://spacy.io/api

Manning, C., Raghavan, P., & Schuetze, H. (2009). *Introduction to Information Retrieval*. 581.

Matplotlib: Python plotting—Matplotlib 3.4.1 documentation. (o. J.). Abgerufen 19. April 2021, von https://matplotlib.org/

Mitra, M., & Chaudhuri, B. B. (2000). Information Retrieval from Documents: A Survey. *Information Retrieval*, *2*(2), 141–163. https://doi.org/10.1023/A:1009950525500

Ryan, M., & Doubleday, A. (2007). *Evaluating 'Throw Away' Prototyping for Requirements Capture*. 9.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 49–58. https://doi.org/10.1145/160688.160693

Singhal, A. (2001). *Modern Information Retrieval: A Brief Overview*. 9.

Sint, R., Stroka, S., Schaffert, S., & Ferstl, R. (2009, Januar 1). *Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis.*

sklearn. (o. J.). *Sklearn.metrics.pairwise.cosine_similarity*. Abgerufen 22. April 2021, von https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html#sklearn.metrics.pairwise.cosine_similarity

Sklearn.feature_extraction.text.CountVectorizer. (o. J.). Abgerufen 22. April 2021, von https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Solariz/german_stopwords. (o. J.). GitHub. Abgerufen 21. April 2021, von https://github.com/solariz/german_stopwords

Stadtpolizei Zürich. (2019). *Teil_a_ausschreibungsunterlagen*. https://intellipurocure.ch/download?browser=true&file=/185479/teil_a_ausschreibungsbestimmungen.pdf

TFIDFSimilarity (Lucene 7.6.0 API). (o. J.). Abgerufen 22. April 2021, von https://lucene.apache.org/core/7_6_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

The Seven Practice Areas of Text Analytics. (2012). In G. Miner, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (S. 29–41). Elsevier. https://doi.org/10.1016/B978-0-12-386979-1.00002-5

Tiefbaumamt Kanton Bern. (2010). *Eignungs- und Zuschlagskriterien*. *1*, 7.

Universal POS tags. (o. J.). Abgerufen 19. April 2021, von https://universaldependencies.org/u/pos/all.html#al-u-pos/PROPN

User Guide—Pandas 1.2.4 documentation. (o. J.). Abgerufen 19. April 2021, von https://pandas.pydata.org/docs/user_guide/index.html

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Verein Simap. (2009, Juni 29). *Simap-Medien-Info2*. https://www.simap.ch/DOWNLOADPART/portalFileInformation/DE/MEDIA_ENTRY_0_ASSOCIATION_2_UPLOAD_1317630867082.pdf

Verein Simap. (2011). *10-Jahre-Simap-Medienmitteilung*. https://www.simap.ch/DOWNLOADPART/portalFileInformation/DE/MEDIA_ENTRY_0_ASSOCIATION_13_UPLOAD_1336394923655.pdf

Welz, T., & Stuermer, M. (2020). Sustainability of ICT hardware procurement in Switzerland: A status-quo analysis of the public procurement sector. *Proceedings of the 7th International Conference on ICT for Sustainability*, 158–169. https://doi.org/10.1145/3401335.3401352

Yogish, D., Manjunath, T. N., & Hegadi, R. S. (2019). Review on Natural Language Processing Trends and Techniques Using NLTK. In K. C. Santosh & R. S. Hegadi (Hrsg.), *Recent Trends in Image Processing and Pattern Recognition* (S. 589–606). Springer. https://doi.org/10.1007/978-981-13-9187-3_53

Selbständigkeitserklärung

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe o des Gesetzes vom 5. September 1996 über die Universität zum Entzug des aufgrund dieser Arbeit verliehenen Titels berechtigt ist.“

Langnau i.E, 26.04.2021 Dominic Schweizer