Institut für Informatik
Der Martin-Luther-Universität Halle-Wittenberg
Naturwissenschaftliche Fakultät III
Big Data Analytics

Bachelorarbeit

Entwicklung und Evaluierung einer Suchmaschine für Polizeipressemitteilungen

zur Erlangung des akademischen Grades
Bachelor of Science
im Studiengang Informatik

Vorgelegt von: Nina Katharina Schwanke
Matrikelnummer: 213211279
am: 7. November 2019

Erstgutachter: Prof. Dr. Matthias Hagen
Zweitgutachter: Prof. Dr. Martin Potthast

Zusammenfassung

Angesichts der Fülle von Informationen die uns im Internet zur Verfügung stehen besteht die dringende Notwendigkeit, Informationen über Textdienste hinweg mit Hintergrundquellen zu vergleichen. Das Ziel der vorliegenden Bachelorarbeit ist es unter der Verwendung von Elasticsearch eine Suchmaschine zu entwickeln, in welche Online-Nachrichtenartikel eingegeben und die dem Nachrichtenereignis zuzuordnenden relevanten Polizeipressemitteilungen ausgegeben werden. Dazu wurden die Pressemitteilungen zunächst von der Website gecrawlt und für die Nutzung in Elasticsearch indexiert. Die aus der URL des Nachrichtenartikels bestehende Anfrage des Nutzenden wird mittels Web-Scraping in Artikeltitel, Artikeltext, Artikeldatum und Artikelort segmentiert. Es wurden zwei Anfrage-Strategien im Rahmen einer an das Cranfield-Paradigma angelehnten Studie verglichen. Hierzu wurde eine Testdokumentensammlung von 105 Online-Nachrichtenartikeln erstellt und evaluiert. Die Ergebnisse der Studie zeigen, dass lange Anfragen, die den Artikeltext beinhalten, deutlich bessere Ergebnis-Rankings erzeugen, als Anfragen, die nur den Artikeltitel berücksichtigen. Nützlich kann die entwickelte und bewertete Suchmaschine beispielsweise im Szenario von widersprüchlicher Berichterstattung sein.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BM Best Match

dpa Deutsche Presse-Agentur GmbH 12

JSON JavaScript Object Notation 12

NDCG Normalized Discounted Cumulated Gain 15

NIST National Institute of Standards and Technology 10

THW Technischem Hilfswerk 12

TREC Text Retrieval Conference 10

URL Uniform Resource Locator

1. Einleitung

Ziel dieser Bachelorarbeit sind der Entwurf, die Implementierung und der Test einer Suchmaschine, die nach Eingabe eines Online-Nachrichtenartikels, Pressemitteilungen der Polizei als Ausgabe zurückliefert, die ebenfalls das im Online-Artikel beschriebene Ereignis betreffen. Mit Hilfe der zu entwickelnden Suchmaschine können dann offizielle Hintergrundinformationen der Polizei, als Ergänzung oder auch Überprüfung des Online-Artikels, aufgerufen werden. Eine solche Suchmaschine ist relevant, da die meisten Menschen Nachrichten online verfolgen und viele die Inhalte anzweifeln.

Nachrichten im Fernsehen verfolgen immer noch 70 % der Menschen, Radio und Printmedien verwenden nur noch 32 % [1]. Insbesondere junge Menschen nehmen Printmedienangebote kaum noch wahr, wenngleich diese häufig ein umfangreicheres Nachrichtenverständnis versprechen [2]. Nachrichten werden primär digital über Smartphone, Computer oder Tablet konsumiert, und das nicht nur von jungen Menschen, denn auch 77 % der über 55-Jährigen, lesen regelmäßig Online-Nachrichtenartikel [1]. Insgesamt lesen weltweit 82 % aller Menschen Online-Nachrichtenartikel.

Leider erweisen sich Online-Artikel jedoch nicht immer als verlässliche Quelle, da diese gelegentlich falsche Informationen enthalten [3]. Das kann versehentlich zustande kommen, geschieht jedoch auch intentional. In Online-Nachrichtenartikeln in sozialen Medien werden vermehrt intentional Falschinformationen publiziert [4]. Sind Anwendende unsicher und möchten Hintergründe überprüfen, kann eine Suchmaschine die offizielle Pressemitteilungen zum Thema liefert hilfreich sein.

Ein anschauliches Beispiel sind Artikel zu einer Geiselnahme am 23. Juni 2016 in einem Kino in Viernheim bei Mannheim. Das Beispiel eignet sich deshalb gut, weil in mehreren archivierten und auch internationalen Online-Artikeln zum Vorfall falsche Informationen publiziert wurden und eine zugehörige Pressemeldung der Polizei vorhanden ist. Am 23.06.2016 wurde der Geiselnehmer von einem Sondereinsatzkommando der Polizei am Tatort erschossen. Dieses spektakuläre Ereignis wurde schnell von Medien aufgegriffen und zeitnah auch online verwertet. Im Frankfurter Regionalteil der BILD-Zeitung wurde beispielsweise noch am selben Tag reißerisch getitelt „Sprengstoff-Mann in hessischem Kino – Geisel Almir (16): So überlebten wir

SPRENGSTOFF-MANN IN HESSISCHEM KINO

Geisel Almir (16): So überlebten wir den Amoklauf!

23.06.2016 - 16:35 Uhr

Viernheim – **Groß-Alarm wegen einer Geiselnahme in einem Kino in Südhessen – jetzt erzählt Geisel Almir Halilovic (16), wie er den Horror überlebte!**

Kurz nach 15 Uhr: Gerade beginnen die ersten Filmvorstellungen des Tages, es laufen das „Dschungelbuch", „Alice im Wunderland" und die Komödie „Central Intelligence".

Da betritt ein Bewaffneter das Kinopolis am Rhein-Neckar-Zentrum! Er ist vermummt, hat eine Schusswaffe in der Hand, trägt einen Patronen-Gürtel um die Schulter. Mit einer Langwaffe feuert er viermal in die Luft. Ob es sich um scharfe Munition handelt ist noch unklar. Der Verbrecher hatte aber auch eine Schreckschuss-Faustfeuerwaffe dabei. Mit der schoss er ebenfalls einmal in die Luft. Die Gaswolke breitete sich schnell aus. Sorgte für Augenreizungen bei vielen Kinogästen.

Schreiend flüchten dutzende Besucher aus dem Kino, sechs Kino-Angestellte führen sie nach draußen. Die Managerin (47): „Ich habe gemerkt, dass etwas passiert ist. Habe die Polizei alarmiert und das Personal angewiesen, sofort das Haus zu verlassen."

Der Waffen-Mann verschanzt sich, nimmt Geiseln. Almir Halilovic (16) war eine von ihnen: **„Der Täter überraschte uns auf der Toilette, sprach gebrochen deutsch. Er zischte uns zu: „Legt Euch hin, wenn Euch euer Leben lieb ist!" Wir waren ungefähr 17 Geiseln."**

Abbildung 1: Ausschnitte der Berichterstattung zur Geiselnahme im Viernheimer Kino am 23.06.2016. Links: Online-Artikel der „BILD"-Zeitung vom 23.06.2016.[1] Rechts: Ausschnitt der „BILD"-Zeitung vom 24.06.2016.[2]

den Amoklauf!" (vgl. Abbildung 1 links). Im Artikel selbst tauchen weitere Details auf, obwohl zum Erscheinungszeitpunkt kaum etwas bekannt war („Bewaffneter", „Patronen-Gürtel um die Schulter", „Waffen-Mann"). Auch in der gedruckten Ausgabe der BILD-Zeitung am Folgetage steht der „Amoklauf im Kino" groß im Titel (vgl. Abbildung 1 rechts). Sogar international wurde über die Ereignisse berichtet: etwa in einem Tweet der BBC, in dem von mindestens 20 Verletzten die Rede ist (vgl. Abbildung 2 links),[3] während in der BILD die Anzahl der Geiseln auf 17 geschätzt wurde. Glücklicherweise sind beide Angaben falsch gewesen. Tatsächlich handelte es sich nicht um einen Amoklauf. Während die konkreten Motive bis heute unbekannt sind, ist sicher, dass lediglich Waffenattrappen verwendet wurden. Entsprechend wurden auch keine 20 Personen verletzt. Der Kinosaal am Tatort war an diesem Nachmittag, nach Angaben der Staatsanwaltschaft Darmstadt, nämlich auch nur mit 15 Gästen besucht, die alle unverletzt blieben.[4] Das große Interesse der Medien hat in diesem Beispiel zur weiten Verbreitung von Fehlinformationen beigetragen.

[1] https://web.archive.org/web/20190828162752/https://uebermedien.de/wp-content/uploads/bild_viernheim.jpg

[2] https://web.archive.org/web/20190726130314/https://www.bild.de/regional/frankfurt/frankfurt-aktuell/kino-viernheim-bewaffneter-mann-schuesse-im-kino-center-46456070.bild.html

[3] https://web.archive.org/web/20160623153633/https://www.bbc.com/news/uk-36610068

[4] https://staatsanwaltschaften.hessen.de/sta-darmstadt?rid=HMdJ_15/STA_Darmstadt_Internet/sub/724/7244c20a-7f84-551d-0648-712ae8bad548 [Zugriff 25.08.2019]

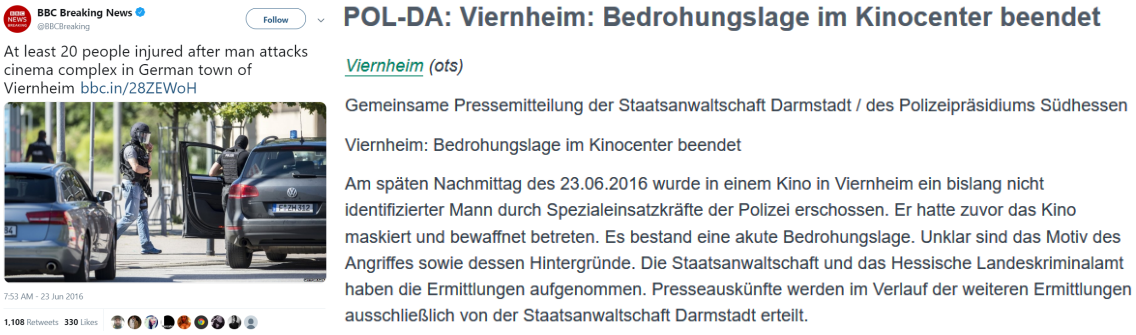

Abbildung 2: Links: Tweet der BBC am 23.06.2016 zur Bedrohungslage im Kinocenter in Viernheim.[5] Rechts: Pressemitteilung des Polizeipräsidiums Südhessen am 23.06.2016 zu den Ereignissen im Viernheimer Kino am selben Tag.[6]

Beispiele von fehlerhaften Meldungen lassen Nachrichtenkonsumeierenden vermehrt Fakten und Aussagen der Medien hinterfragen. In Deutschland ist 38% der Befragten unklar, welche Inhalte im Internet Fehlinformationen enthalten.Weltweit vertrauen 2019 durchschnittlich nur noch 42 % den Nachrichteninhalten im Allgemeinen, 2 Prozentpunkte weniger als im Vorjahr [1]. Nur noch die Hälfte traut den Inhalten der Nachrichtenangebote, die er selbst nutzt [1]. Die meisten Menschen nutzen Online-Nachrichtenartikel vertrauen deren Inhalten aber nicht.

Wird das Misstrauen von Nutzenden zu groß, kann dies einen vorübergehenden oder langfristigen Interessenverlust an Nachrichtenmeldungen zur Folge haben. Damit verringert sich allerdings auch die gesellschaftliche und politische Teilhabe, was nicht demokratiefördernd ist [5]. Insbesondere für junge Menschen ist es wichtig, dass Nachrichten unabhängig, sachlich und vertrauenswürdig sind [6].

Im Falle, dass Inhalte von Online-Nachrichten hinterfragt werden, könnte ein Ausweg die Unterstützung einer Überprüfung anhand zuverlässiger Quellen sein. Das Finden der Primärquelle befriedigt das Informationsbedürfnis des Fragenden. Den dafür erforderlichen Rechercheaufwand würde eine passende Polizeipressesuchmaschine verringern helfen. Zu einer Online-Nachricht sollte es also möglichst einfach sein, Quellen zu identifizieren, die zum einen verlässlich sind in der Richtigkeit ihrer

[5] https://web.archive.org/web/20190824131706/https:/twitter.com/BBCBreaking/status/745993273660088320/photo/1

[6] https://web.archive.org/web/20160625131945/https://www.presseportal.de/blaulicht/pm/4969/3361168

Angaben, und zum anderen die Informationen bereitstellen, über die in nachgefragten Online-Nachrichtenartikeln berichtet wird.

Das Informationsbedürfnis des Menschen ist auf sein Umfeld konzentriert [7]. Daraus resultiert auch das Interesse an Nachrichtenmeldungen im Allgemeinen, zum Beispiel als Gesprächsthema. Über viele lokale Ereignisse berichten die Pressemitteilungen der örtlichen Polizei, die beispielsweise auf der Website Presseportal zugänglich sind. Das Interesse an diesen Informationen wächst – verdeutlicht etwa durch die ständig wachsenden Zugriffszahlen dieser Internetseite.[7] Darüber hinaus stellen Polizeipressemitteilungen eine Nachrichtenquelle dar, der gemeinhin vertraut wird [8]. Deshalb eignet sich die Website Presseportal als Ausgangspunkt, um eine Suchmaschine zu entwickeln, die es möglichst einfach macht, zu einem Nachrichtenartikel eine vertrauenswürdige Primärquelle zu finden (etwa die gegebenenfalls existierenden offizielle Polizeipressemitteilung). Selbstverständlich verfügt auch die Website Presseportal über eine Suchfunktion. Diese stellt sich leider jedoch als unzureichend heraus, falls eine Mitteilung konkret zu einem Ereignis gesucht wird. Insbesondere dann, wenn das Ereignis bereits mehr als 14 Tage zurück liegt.

Das dieser Bachelorarbeit zugrundeliegende Projekt soll dazu beitragen, Hintergrundinformationen zu Online-Nachrichtenartikeln digital und bedienungsfreundlich zugänglich zu machen. Dazu wird eine Suchmaschine für Polizeipressemitteilungen mit dem Framework Elasticsearch entwickelt. Die Pressemitteilungen wurden zunächst von der Website „Presseportal“ heruntergeladen und gespeichert. Sie standen für diese Bachelorarbeit bereits vorab zur Verfügung. Für die Nutzung in Elasticsearch ist die Anpassung der Daten an eine spezielle Indexdatenstruktur notwendig. Auch die Eingabe der Anwendenden (die Suchanfrage) in die Suchmaschine muss entsprechend angepasst werden. Für den Suchalgorithmus wird das Retrieval-Modell BM25 verwendet. Der Erfolg der Suche ist maßgeblich von der Schnittstelle zwischen der Eingabe von Anwendenden und Elasticsearch bzw. dem dahinter stehenden Retrieval-Modell abhängig. Diese Schnittstelle ist deshalb Kern des Projekts. Die Suchmaschineneingabe durch Nutzende soll möglichst einfach sein und deshalb nur die Internet-Adresse (URL, Uniform Resource Locator) des Nachrichtenartikel selbst sein. Ausgabe sind die relevanten Pressemitteilungen zum angefragten Artikel. Damit können Interessierte ihr Wissen aus dem Online-Nachrichtenartikel, um

[7] `https://www.presseportal.de/about` [Zugriff 02.09.2019]

die prägnanten Informationen aus den relevanten Pressemitteilungen ergänzen. Wie relevant eine einzelne Pressemitteilung bezüglich eines Nachrichtenartikels ist, wird nach dem bei der TREC-Konferenz etablierten *Cranfield-Paradigma* bewertet. Kern der Bewertungsmethode ist der Aufbau einer Testsammlung, welche dann wiederholt in Experimenten verwendet wird. Die Testsammlung setzt sich aus einer Dokumentensammlung, Topics und Relevanzbewertungen zu den Topics und Dokumenten zusammen [9].

Die Hauptbeiträge dieser Bachelorarbeit sind die Erstellung einer Sammlung von Online-Nachrichtenartikeln (Topics) und Relevanzbewertungen von Polizeipressemitteilungen (Dokumentensammlung) bezüglich dieser Topics, also die Zusammenstellung einer Testsammlung nach dem Cranfield-Paradigma. Um diesen Bewertungsmaßstab sinnvoll anwenden zu können, müssen mindestens zwei vergleichbare Suchmaschinenanfragen umfangreich ausgewertet werden, indem für beide die Ergebnisausgaben evaluiert werden. Dies geschieht durch Relevanzbewertungen. Basierend auf der erstellten Testsammlung soll die folgende Forschungsfrage beantwortet werden:

- Wie kann zu einem Online-Nachrichtenartikel automatisch eine Anfrage an eine Elasticsearch-basierte Suchmaschine so formuliert werden, dass die Retrieval-Qualität höher ist als bei einer einfachen aus dem Artikeltitel bestehenden Anfrage?

Die vorliegende Bachelorarbeit ist wie folgt gegliedert. In Kapitel 2 wird grundlegenden Literatur zum Thema „information retrieval systems“ und „query by document“ vorgestellt und es werden wichtige Begrifflichkeiten erklärt. In Kapitel 3 steht die explorative Datenanalyse der von der Website Presseportal gecrawlten Pressemitteilungen und die Struktur der Topics, die zur Evaluierung der Suchmaschine verwendet werden im Mittelpunkt. In Kapitel 4 wird der Ansatz für das Indexieren der Polizeipressemitteilungsdaten für die Verwendung in Elasticsearch und die Merkmale der Suchanfrage, die aus den Textbausteinen eines Online-Artikels eine Anfrage an Elasticsearch formuliert beschrieben. Kapitel 5 erläutert die konkreten Evaluierungsmaßnahmen nach dem Cranfield-Paradigma und deren Ergebnisse. Kapitel 6 fasst die erzielten Ergebnisse kurz zusammen und gibt einen Ausblick auf noch offen gebliebene Fragen.

2. Verwandte Arbeiten und Begriffsklärung

Dieses Kapitel führt kurz in die für die Entwicklung einer Suchmaschine wesentlichen Konzepte ein und erklärt in diesem Zusammenhang wichtige Begriffe. Die Rahmendaten der genutzten Suchmaschine Elasticsearch werden benannt und die theoretischen Grundlagen der komplexen Anfragestrategie und der Evaluierungsmethode beschrieben.

2.1. Suchmaschinen

Ausgangspunkt der Funktionalität einer Suchmaschine ist der *Index* und das Retrieval-Modell. Der Index ist die Datenstruktur, welche die Informationen der zu durchsuchenden Dokumente beinhaltet. Information Retrieval (IR) ist auf das Abrufen beziehungsweise Auffinden bereits vorhandener Informationen konzentriert. Dabei handelt es sich um computergestütztes Suchen nach komplexen Inhalten. Deshalb liegen Suchmaschinen, die komplexere Suchanfragen unterstützen, Retrieval-Modelle zugrunde.

Die in eine Suchmaschine eingegebenen Suchterme bezeichnet man als *Anfrage*. Die Herausforderung besteht darin, allein unter Verwendung der Anfrage die gesuchten Informationen aus einer vorhandenen Textdokumentensammlung bereitzustellen. Grundlegende Aufgaben von Information Retrieval sind daher die Formalisierung und Modellierung der Entscheidungsprozesse welche Menschen zugrunde legen, wenn sie entscheiden, ob ein Dokument beziehungsweise ein Text ein bestimmtes Informationsbedürfnis erfüllt. Stimmen die Inhalte der Ergebnisdokumente eines Modells mit dem Informationsbedürfnis des Menschen überein, ist das Modell zutreffend. Um die Qualität eines Modells zu bewerten, erfolgt üblicherweise eine Evaluierung der Inhalte der Ergebnisdokumente durch menschliche Annotatoren. Dabei wird ein Dokument, das dem Informationsbedürfnis einer spezifischen Anfrage eines Nutzenden entspricht, als *relevant* eingestuft. Relevanz wird zumeist nur binär unterschieden, kann aber auch mehrstufig differenziert werden [10]. Entsprechend sind für eine sinnvolle Bewertung immer umfangreiche Nutzerstudien erforderlich. Um einen einheitlichen Standard zu wahren existieren hierfür Richtlinien der Text Retrieval Conference (TREC),[3] die wissenschaftliche Vergleichbarkeit gewährleisten sollen.

[3] `https://trec.nist.gov/`

Die wichtigen Merkmale von Suchmaschinen beinhalten auch die wesentlichen Punkte des Information Retrieval: effektive Ranking-Algorithmen, Evaluierung und Benutzerinteraktion. Der Ranking-Algorithmus hängt dabei direkt vom zugrundeliegenden Retrieval-Modell ab. Das Ranking sollte möglichst *effizient* und *effektiv* sein. Die Effizienz des Rankings ist von der Laufzeit und dem Speicherbedarf des Index abhängig. Die Effektivität hingegen beruht auf dem jeweils verwendeten Retrieval-Modell und kann durch Evaluierungsmethoden eingeschätzt werden [11]. Dabei basiert die Wahl des zu bevorzugenden Retrieval-Modells unter anderem auf der Struktur und Länge der Querys im Verhältnis zu den Dokumenten im Korpus. Ein *Korpus* beschreibt die Dokumentensammlung verwendet in Information Retieval Experimenten. Das Korpus beinhaltet eine Sammlung von Textdokumenten, typischen Anfragen und *relevance judgements*. Unter Relevanzbeurteilungen ist eine Liste der relevanten Dokumente für jede Query zu verstehen.

2.2. Retrieval-Modell BM25

Okapi BM25.[4] ist ein verbreitetstes Retrieval-Modell, das in vielen Bereichen Anwendung findet. Das Modell basiert prinzipiell zunächst auf der Gewichtung von Anfrage Termen bezüglich eines Dokuments, in welchem diese enthalten sind. Dabei erfolgt bei erhöhtem Auftreten eine stärkere Gewichtung. Das Modell der Vorkommenshäufigkeit wird „term frequency“ (TF) genannt und bietet die Grundlage für die Zuordnung von Dokumenten zu einer Suchtrefferliste. Damit nicht die im allgemeinen Sprachgebrauch am häufigsten verwendeten Wörter und Textbausteine wie beispielsweise Artikel, Präpositionen und Relativpronomen der Anfrage Terme das Suchergebnis bestimmen, wird TF um die inverse Dokumenthäufigkeit (IDF) erweitert.

Dieses Maß bestimmt die spezifische Häufigkeit eines Terms für die Gesamtmenge der betrachteten Dokumente, denn ein in der Dokumentensammlung häufig vorkommender Term verfügt über weniger Aussagekraft, um ein Dokument auf der Suchtrefferliste zu positionieren, als ein seltener. Das probabilistische Retrieval-Modell Okapi BM25 modelliert nun zusätzlich die Relevanz der Dokumente unter Berücksichtigung ihrer jeweiligen Gesamtlänge im Verhältnis zu allen anderen Dokumentlängen in der Dokumentensammlung. Ein kürzeres Dokument mit der gleichen Termhäufig-

[4] BM steht für Best Match, und 25 ist ein Nummerierungsschema

keit wird einem langen Dokument gegenüber vorgezogen. Im Zusammenhang dessen ergibt sich die Berechnung des *Scores* eines Dokuments. Eine Suchmaschine berechnet für alle Dokumente im Korpus einen Score zu einer Anfrage. Das Dokument mit dem höchsten Score wird als erster Treffer der Suche ausgegeben. Der Score eines Dokuments D und einer Anfrage Q berechnet sich nach BM 25 im Wesentlichen wie folgt.

$$score(D,Q) = \sum_{i=1}^{n} \log \frac{N - n(q_i) + 0,5}{n(q_i) + 0,5} \cdot \frac{f(q_i,D) \cdot (k_1 + 1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \quad (1)$$

Dabei sind die $q_1, ..., q_n$ die Terme der Anfrage. Die Termhäufigkeit des Terms q_i im Dokument ist $f(q_i, D)$ und $|D|$ die Länge des Dokumentes D in Wörtern. Die durchschnittliche Dokumentlänge in der Textsammlung ist durch $avgdl$ beschrieben. Die freien Parameter k_1 und b sind, falls nicht anwendungsspezifisch optimiert, üblicherweise $k_1 \in [1,2; 2,0]$ und $b = 0,75$ ($b \in [0..1]$). Die Berücksichtigung des IDF-Modells wird im ersten Faktor der Formel ersichtlich, denn N ist die Gesamtanzahl der Dokumente in der Sammlung und $n(q_i)$ die Anzahl der Dokumente die q_i enthalten. Es gibt Abweichungen und Anpassungen dieser Formel, die im Zusammenhang mit BM25 genutzt werden. Es besteht beispielsweise die Möglichkeit auch die Länge der Query mit zu berücksichtigen.

Problematisch am BM 25 Modell kann sein, dass die Terme die in mehr als der Hälfte der Dokumente vorkommen, nicht berücksichtigt werden. Der Nachteil der alleinigen Verwendung des TF und IDF überwiegt jedoch, da hier vermehrt auftretende Wörter den Score beeinflussen können. BM25 eignet sich deshalb in der Form besser, dass häufige Terme einen geringen Einfluss auf die Ausgabe im Ergebnisranking haben und die Dokumentlänge im Verhältnis zur Dokumentensammlung Beachtung findet [11, 10].

2.3. Elasticsearch

Lucene[5] ist eine bekannte Java basierte Open-Source-Suchmaschine, die in einer Vielzahl von kommerziellen Anwendungen eingesetzt wird. Elasticsearch basiert auf der Lucene-Bibliothek und ist unter der Apache-Lizenz 2.0[6] freigegeben. Oka-

[5] http://lucene.apache.org/

[6] https://github.com/elastic/elasticsearch/blob/master/LICENSE.txt

pi BM 25 wird seit der Apache-Lucene™ Version 6.0.0 als Standard Retrieval-Modell zur Verfügung gestellt.[7] Dieser Standard wurde dann auch in Elasticsearch integriert. In dieser Bachelorarbeit wird Elasticsearch Version 7.2.0 basierend auf Lucene Version 8.0.0 verwendet. Die browserbasierte Open-Source-Analyseplattform Kibana ist ebenfalls in Version 7.2.0 mit enthalten. Für den Betrieb der Suchmaschine standen die Hardwareressourcen des Webis-Teams zur Verfügung. Für die Verwendung in Elasticsearch müssen die Dokumente des Korpus, in denen gesucht werden soll, indexiert werden, denn Strukturierung der Daten im Index ermöglicht ein schnelles Suchen. Wie der Index konkret gestaltet wird, hängt vom Anforderungsprofil des Retrival-Modells und dem jeweiligen Korpus ab. Die Indexierungsstrategie ist deshalb anwendungsspezifisch und wird in Kapitel 3 näher beschrieben.

2.4. Indexierung

Die Implementierung einer Suchmaschine setzt die Verfügbarkeit einer Dokumentensammlung voraus. Eine Sammlung von Textdokumenten muss dann so gespeichert sein, dass ein Zugriff durch die Suchfunktion (Retrieval-Algorithmus) der Suchmaschine möglich ist. Schnittstelle für diesen Zugriff ist die Indexierung der Dokumente. Indexverarbeitung fasst im Wesentlichen drei Schritte zusammen: Texterfassung, Texttransformation und Indexerstellung. Die Texterfassung ist die Identifizierung und Bereitstellung der zu durchsuchenden Dokumente. Häufig handelt es sich um eine bestehende Dokumentensammlung, wie beispielsweise im vorliegenden Projekt die Polizeipressemitteilungen. Sind die gewünschten Dokumente wie in diesem Fall im Web verfügbar, können sie heruntergeladen werden, man spricht von einem *Crawl*. Der Begriff Crawl beschreibt den Prozess eines Skript- oder Softwareprogramms, das eine Website besucht und Inhalte und Links erfasst. Dieser web-crawl dient der Erstellung eines Datenspeichers, indem Text und Metadaten aller Dokumente gespeichert sind. Der zweite Schritt der Indexverarbeitung ist die Texttransformation. Ziel ist es, die Bestandteile eines Dokuments, die zum Suchen gebraucht werden, in *Index-Terme* beziehungsweise Terme aufzuteilen. Alle Terme eines Index bilden das *Index-Vokabular*. Ein einzelner Term, kann z.B. ein einzelnes Wort, Datum, ein Satz oder eine Formulierung sein. Die Ausgabe der Texttransformation bildet den Index. Der Index sollte effizient erweiterbar sein, indem er schnell um neue Dokumente ergänzt werden kann [11].

[7] `http://mail-archives.apache.org/mod_mbox/www-announce/201604.mbox/<CABM+u9LWQhLxSyg=vfQ4=_gcLOPLe6FJrejH0jjjFwFhBRivBQ@mail.gmail.com>` [Zugriff 25.09.2019]

2.5. Query by Document

Wird ein Textdokument als Anfrage an eine Suchmaschine gestellt und verwandte Dokumente aus einer anderen Textsammlung als Ergebnis ausgegeben, bezeichnet man diese Funktionalität als „Query by Document“ [12]. Es werden aus einem Dokument mehrere Querys beziehungsweise Anfragen generiert und konjunktiv verknüpft. Abhängig davon, wie gut die einzelnen Querys mit Textteilen eines Dokuments in der Dokumentensammlung übereinstimmen, wird ein Score ermittelt. Für die Verknüpfung aller Querys die aus dem Anfragedokument ermittelt wurden ergibt sich ein Gesamt-Score für das jeweilige Dokument [13].

2.6. Cranfield Paradigma

Die Text Retrieval Conference (TREC) ist eine Evaluierungskampagne zur Untersuchung der Effizienz von Information Retrieval-Methoden. Seit 1992 wird TREC durch das National Institute of Standards and Technology (NIST) und das U.S.-Handelsministerium durchgeführt. Ziel ist es, Forschung auf dem Gebiet Information Retrieval zu fördern, indem die Infrastruktur für groß angelegte Bewertungen von Text Retrieval-Methoden bereitgestellt wird. Hierzu werden große und praxisnahe-Textsammlungen und Fragen zur Verfügung gestellt. Teilnehmende Teams bei TREC stellen die zur Verfügung gestellten Fragen an ihr eigenes Retrieval-System und reichen eine Liste der höchstplatzierten Dokumente ein. NIST bündelt dann diese Einzelergebnisse und bewertet die Relevanz der Dokumente [14]. Zur Evaluierung der Ergebnisse hat sich das *Cranfield Paradigma* bewährt. Hierbei handelt es sich um einen Test, der sich aus folgenden drei Grundbausteinen zusammensetzt [9].

- Dokumentensammlung
- Topics
- Relevanzbewertungen zu den Topics und Dokumenten

Die Dokumentensammlung fasst die Dokumente zusammen, die durchsucht werden können. Die Topics sind die Anfragen, die an diese Dokumentensammlung gestellt werden können. Dabei wird mit einem Topic ein Informationsbedürfnis eines Suchenden verknüpft, welches es zu erfüllen gilt. Die Relevanzbewertungen geben an, ob ein Dokument dem erwünschten Informationsbedürfnis eines Topics genügt. Sind die Informationen im Dokument enthalten wird es als relevant bewertet. Diese drei

Teile bilden die Testsammlung und ermöglichen eine wiederholte Evaluierung verschiedener Retrieval-Modelle anhand der gleichen Topics. Damit können verschiedene Systeme (verschiedene Retrieval-Modelle) in Bezug auf die gleiche Testsammlung untersucht werden, was eine hohe Vergleichbarkeit gewährleistet. Die Testsammlung ist eine bestehende Ressource, die für nachfolgende Untersuchungen genutzt werden kann [9]. Diese Bewertungsmethode hat sich im Rahmen der TREC entwickelt und etabliert.

Ein TREC-Zyklus endet mit einer Konferenz, die dem Erfahrungsaustausch der Teilnehmenden dient. Dadurch wird der Austausch von Forschungsideen und Entwicklungen gefördert und ein wissenschaftlich vergleichbarer Standard für die Bewertung von Retrieval-Systemen etabliert. Die TREC-Testsammlungen und Bewertungssoftware stehen der gesamten Retrieval-Forschungsgemeinschaft zur Verfügung, sodass die Evaluierung eigener Retrieval-Systeme jederzeit möglich ist. So sind geeignete und aktuelle Bewertungstechniken leicht zugänglich. Das beschleunigt insbesondere auch den Technologietransfer von Forschungsresultaten zu kommerziellen Produkten, einem weiteren Ziel der Konferenz TREC [14].

3. Korpuserstellung und -analyse

Die verwendete Dokumentensammlung besteht aus deutschsprachigen Pressemitteilungen von Landes- und Bundespolizei, Feuerwehr, Bundesbehörden, Staatsanwaltschaft, Zoll und Technischem Hilfswerk (THW). Diese werden auf der Website Presseportal verwaltet.[8] Der Begriff Polizeipressemitteilungen fasst in dieser Bachelorarbeit einfachheitshalber alle Pressemeldungen der genannten Einrichtungen zusammen, da die meisten Mitteilungen des verwendeten Dokumentensammlung der Polizei zuzuordnen sind. Die Pressemitteilungen werden von *news aktuell*, einem Tochterunternehmen der Deutsche Presse-Agentur GmbH (dpa), zur Verfügung gestellt. Eine einzelne Pressemeldung der Polizei wird von der zuständigen Pressestelle, beziehungsweise den lokalen Vertretern für Öffentlichkeitsarbeit (synonym Public Relations, PR) der jeweiligen Polizeidienststelle, an die dpa übermittelt. Alle PR-Inhalte werden auf der Website Presseportal veröffentlicht und zusätzlich an verschiedene Medienformate weitergeleitet.[9]

Der in dieser Arbeit verwendete Crawl umfasst insgesamt 1.172.703 Pressemeldungen die im Zeitraum 01.01.2001 bis 05.07.2018 auf der Website Presseportal veröffentlicht wurden. Da nur der Textteil einer Meldung in der Suchmaschinenausgabe angezeigt werden soll, wird auf Videos und Bilder verzichtet und lediglich der Link zum Originalartikel, welcher zusätzliches Bild- und Tonmaterial enthalten kann, mit angegeben.

Eine einzelne Pressemitteilung liegt dabei nach dem Crawlen im offenen Standarddateiformat JavaScript Object Notation (JSON) vor. Die einzelnen Attribute sind in Tabelle 1 beispielhaft für eine Pressemitteilung der Polizei dargestellt. Meldungen anderer Dienststellen, beispielsweise der Feuerwehr, weisen die gleiche Struktur mit entsprechend anderen Spezifikationen auf (z.B. Dienststellenkürzel: Polizei „POL“ und Feuerwehr „FW“).

3.1. Erstellung der Dokumentensammlung

Der erste wichtige Schritt zur Entwicklung der Suchmaschine ist die Vorverarbeitung der Polizeipressemitteilungen, damit diese als strukturierte Daten die Dokumenten-

[8] `https://www.presseportal.de/blaulicht/` [Zugriff 02.09.2019]
[9] `https://www.dpa.com/de/unternehmen/dpa-gruppe/news-aktuell/` [Zugriff 02.09.2019]

Tabelle 1: Bestandteile einer Polizeipressemitteilung im JSON-Format

Attribut	Beschreibung	Beispielwert
_id	ergänzte Identifikationsnummer, zusammengesetzt aus officeID und ID	6013-3697313
URL	url der Pressemitteilung, enthält zuletzt die officeID und die ID	https://www.presseportal.de/blaulicht/pm/6013/3697313
officeID	ID der Polizeidienststelle	6013
title	Dienststellenkürzel: Titel der Pressemitteilung	POL-MFR: Unbekannter Fußgänger gesucht
officeName	Name der Polizeidienststelle	Polizeipräsidium Mittelfranken
keywords	Schlüsselbegriffe, immer mindestens Polizei meistens auch Kriminalität	„Polizei“, „Kriminalität“
body	Haupttext der Pressemitteilung und Kontaktdaten der Dienststelle	Fürth (ots) - Die Verkehrspolizei Fürth sucht ...
ID	Identifizierungsnummer der Pressemitteilung	3697313
officeURL	url der Pressemitteilungen der einen Dienststelle	https://www.presseportal.de/blaulicht/nr/6013
published	Veröffentlichungsdatum	2017-07-31T11:35:53+0200

sammlung bilden. Die Pressemitteilungen müssen dafür erfasst, dann transformiert und zuletzt indexiert werden. Dieser Prozess ist in diesem Unterkapitel näher beschrieben. Die Pressemitteilungsdokumente wurden im Juli 2018 gecrawled und lagen somit bereits vor Beginn dieser Bachelorarbeit im JSON Lines Format vor.[10,11]

Elasticsearch bietet die Möglichkeit, im Rahmen der Indexierung der Dokumente für die Suche Bild- und Tondaten zu entfernen. Darüber hinaus können die einzelnen Dokumente um zusätzliche Informationen ergänzt werden. In diesem Fall ist jedem Dokument eine eindeutige Identifikationsnummer (ID) zugeordnet worden. Eine eindeutige Kennung ist für die später folgende Evaluierung zwingend erforderlich. Die gewählte ID (hier _id) setzt sich aus der Identifikationsnummer der Pressemitteilung und der Identifikationsnummer der herausgebenden Dienstelle zusammen, da eine dieser Nummern zur Identifizierung einer einzelnen Pressemitteilung nicht ausreichend ist. Ein weiterer Vorteil dieser beiden Nummern ist, dass beide in

[10] `http://jsonlines.org/` [Zugriff 02.09.2019]

[11] Der Crawl wurde durch Studenten (A. D. Hakimi, A. Susheva, B. Nicholson, C. Pfeiffer, C. Traser und D. Sturm) der Universität Leipzig im Rahmen eines Moduls im Sommersemster 2018 durchgeführt.

der URL einer Pressemitteilung enthalten sind (https://www.presseportal.de/ blaulicht/pm/**OfficeID/ID** ,vgl. Tabelle 1). Das ermöglicht einen einfacheren Vergleich des Ergebnisrankings der Polizeipressesuchmaschine mit Ergebnisrankings anderer externer Suchportale, welcher im Rahmen der Evaluierung angestrebt wird. Diese Anpassung der Dokumente geschieht mittels eines Analyzers der standardmäßig von Elasticsearch zur Verfügung gestellt wird, und Konfigurationsmöglichkeiten bietet.[12]

Eine weitere Modifikation ist die Verwendung einer Stoppwortliste während des Indexierens.[13] Stoppwörter sind Füllwörter beispielsweise „und“, „an“, „die“, „ich“, „auch“ et cetera. Diese Wörter enthalten selbst kaum Information über den Kontext und ermöglichen für sich genommen keine adäquate Unterscheidung zwischen den Dokumenten. Deshalb ist es sinnvoll diese Wörter aus dem Index zu entfernen. Ohne Stoppwörter ist das Indexvokabular von geringerem Umfang und lässt ein schnelleres Durchsuchen des Indexes zu. In dem für die Polizeipressemitteilungen konfigurierten Analyzer werden neben den Stoppwörtern auch Leerzeichen entfernt. Zusätzlich werden die einzelnen Wörter gestemmt. Darüber hinaus werden alle Wörter normalisiert, das bedeutet es wird ausschließlich Kleinschreibung verwendet.

3.2. Selektion der Topics

Eine Pilotstudie[14] entwickelte und evaluierte zwei Anfragestrategien. Für die Evaluierung wurden die Relevanz von insgesamt über 500 Pressemitteilungen zu 46 Onlineartikeln bewertet. Diese Nachrichtenartikel sind 6 verschiedenen Themenkategorien zugeordnet. Dabei betreffen nicht alle Online-Artikel unterschiedliche Ereignisse.

Die Merkmale beider Anfragestrategien werden nun kurz erläutert. Die erste Strategie besteht im Suchen des Onlineartikeltitels im Titel und im Haupttextteil (auch Body genannt) der Pressemitteilungen. Die zweite Anfrage ist diffiziler gestaltet. Der Titel des Artikels wird im Titel der Pressemitteilung und der Body des Artikels im Body der Pressemitteilung gesucht. Weiter wird der Ort des Artikels im Body, im Titel oder in den Adressangaben der Pressemitteilung gesucht. Zusätzlich wird nach dem Erscheinungsdatum des Artikels im Zeitraum von zwei Wochen davor bis acht Wochen danach gefiltert.

[12] `https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-analyzers.html`
[13] Deutsche Standardstoppworte von Elasticsearch
[14] Sommerschule Studienstiftung Greifswald 2018

Verglichen wird die Ranking-Qualität der beiden Anfragestrategien anhand des Normalized Discounted Cumulated Gain (NDCG). Für die erste Anfragestrategie ergab sich $nDGC@3 = 0,485$ und $nDGC@3 = 0,698$ für die komplexere Version.

Die Ergebnisse der Pilotstudie können für die Ermittlung der Stichprobengröße genutzt werden. Um diese zu berechnen kann ein statistischer Test genutzt werden. Bei einem Hypothesentest wird einem Beobachtungsergebnis (Hypothese) eine mathematisch begründete Entscheidung über dessen Gültigkeit zugeordnet. Die Hypothese kann entweder gültig oder ungültig sein. Dabei nimmt man eine erste Hypothese, Nullhypothese genannt an und überprüft die Gültigkeit einer zweiten Alternativhypothese. In diesem Anwendungsfall ist die einfache Anfragestrategie die Nullhypothese und die komplexere Strategie die Alternativhypothese deren Gültigkeit zu Überprüfen ist. Für die Beantwortung der Forschungsfrage dieser Bachelorarbeit ist es interessant zu erfahren wie groß eine Stichprobe gewählt werden muss, um auszuschließen, dass es sich bei den erhobenen Daten um eine zufällige Verteilung handelt. Zur Berechnung des notwendigen Stichprobenumfangs wurde ein Zweistichproben-t-Test mit dem Programm G*Power Version 3.1.9.4 durchgeführt [15]. Grundlage der Berechnung sind $nDGC@3$ beider Anfragesrategien, denn der t-Test fordert zwei Mittelwerte. Daraus ergibt sich eine geforderte minimale Stichprobengröße von 12. Diese Mindestanforderung wird überschritten und 15 Artikel pro Themenkategorie gewählt, damit die nDCG-Werte pro Kategorie eine statistisch signifikante Aussagekraft versprechen.

Alle Topics sind einer Kategorie zugeordnet, hierdurch kann die Auswirkung der Anfragestrategien auf verschiedenen Themengebiete analysiert werden. Einerseits liegt eine hohe thematische Varianz zwischen den Kategorien vor, sodass sichergestellt wird, dass eine weitreichende Anwendung durch Nutzende die unterschiedlichste Hintergründe erfahren möchten, möglich ist. Andererseits weisen die Kategorien nahe Anknüpfungspunkte auf, um zu testen, ob Ereignisse über die häufig in ähnlicher Art und Weise berichtet wird differenzierbar sind und durch die gewählte Anfragestrategie entsprechend in der Ausgabe platziert werden. Eine thematische Eingrenzung erfolgte dabei schon zwingend durch die Dokumentensammlung, die nur eine begrenzte Anzahl an Pressemitteilungen umfasst. Diese beziehen sich ohnehin nur auf spezifische Informationen, zumeist über Kriminalität oder lokale Unglücksfälle. Deshalb ist eine stichprobenartige Untersuchung der Dokumentensammlung erfolgt,

um zu ermitteln zu welchen Themenbereichen viele oder auch wenige Pressemitteilungen enthalten sind. Dabei ist auch ein subjektiver Einblick über die Struktur und den Aufbau der Pressemitteilungen gelungen. Ein weiterer wichtiger Ansatzpunkt für die Bildung von Kategorien ist das zu erwartende Informationsbedürfnis der Nutzenden. Deshalb sind unter Berücksichtigung der deutschen Kriminalstatistik von 2018 sieben Kategorien entstanden, die repräsentativ für die Bedeutung polizeilicher Vorfälle sind [16]. Dabei sind Verbrechen ohnehin eines der häufigsten Themengebiete in Online-Nachrichtenartikeln weltweit [17]. Die Kategorien sind Mord, Diebstahl, Migration, Sportveranstaltungen, Unwetter, Verkehrsunfälle und allgemeine Kapitalverbrechen. Jeder Kategorie sind 15 Online-Nachrichtenartikel zugeordnet. Die einzelnen Kategorien werden nachfolgend detaillierter erläutert.

Kategorie - Mord. Diese Kategorie umfasst Mord und ähnliche Tötungsdelikte und die damit einhergehenden Untersuchungen. Hier liegt der absolute Fokus auf Kriminalität. Mord ist als Kapitaldelikt im Verhältnis zu Massendelikten wie beispielsweise Diebstahl eine Seltenheit, in der Schwere aber nicht zu übertreffen [16]. Insofern dienen beide Delikte als absolute Kontrastobjekte im gemeinsamen Bereich der Straftaten. Interessant ist hier auch, dass bei laufenden Untersuchungen durchaus eine hohe Anzahl von Pressemitteilungen der Polizei zu einem einzelnen Mordfall auftauchen können. Dies stellt eine besondere Herausforderung bei der Suche nach den relevantesten Meldungen dar. So ist die Pressemitteilung, die die Ergreifung eines Täters beschreibt relevanter als der öffentliche Fahndungsaufruf.

Kategorie – Diebstahl. Diese Kategorie befasst sich mit allen Ereignissen, die im direkten Zusammenhang mit einem Diebstahl stehen. Dabei ist Diebstahl als klassisches Delikt schon aufgrund seiner Häufigkeit geeignet, eine passende Obergruppe für eine Vielzahl von Meldungen darzustellen [16]. Dies zeigt sich auch in der Dokumentensammlung, denn das Wort „Einbruch“ kann in 13,6 % der Pressemeldungen gefunden werden.

Kategorie – Migration. Die Kategorie Migration befasst sich mit Streitigkeiten zwischen Personen verschiedener Nationen und Meldungen in Bezug auf Asylsuchende und andere Flüchtlinge. Ebenso sind Verhaftung von im Ausland gesuchten Straftätern und Abschiebungen enthalten. Diese Kategorie befasst sich mit einem Thema, das durchaus auch politische Bedeutung aufweist. Hier besteht ein hohes

Informationsinteresse der Bevölkerung, da Kriminalität im Zusammenhang mit Migration wiederholt, vor allem auch in den letzten Jahren, ein Thema im öffentlichen Diskurs war. Pressemitteilungen Migration betreffend sind in das politische Informationsbedürfnis einzuordnen. Politische Themen sind vorherrschend in Nachrichtenmeldungen [17, 18].

Kategorie – Sportveranstaltungen. Sport ist ein Thema das in deutschen Artikel im Vergleich zu USA, Russland, UK und Frankreich besonders häufig erwähnt wird [17]. Eine Inhaltsanalyse von Onlineartikeln der Süddeutschen und Spiegel Online aus dem Jahr 2008 ergab, dass „Sport“ in mehr als 11 % der Online-Nachrichtenartikel auftaucht [17]. In der Dokumentensammlung enthalten knapp 32.000 Dokumente den Begriff Sport. Diese wichtige Kategorie umfasst Meldungen im Zusammenhang mit großen Sportereignissen, wobei es sich vornehmlich um Fußballspiele handelt. Dabei erfolgt eine besondere Häufung in Bezug auf Welt- und Europameisterschaft. Mitteilungen über körperliche Auseinandersetzungen, Unfälle aber auch friedlich verlaufene Großveranstaltungen sind hier eingeordnet. Die Besonderheit dieser Kategorie ist, dass die Ausgestaltung der konkreten Meldung vielfältig ist, die Gemeinsamkeit jedoch darin besteht, dass der mediale Fokus oft auf den Ausgang des sportlichen Events konzentriert ist.

Kategorie - Unwetter. Diese Kategorie beinhaltet Ereignisse im Zusammenhang mit Unwettern wie Stürmen, Hochwassern und Winterkatastrophen. Darunter fallen z.B. wetterbedingte Absagen von Großveranstaltungen, aber auch Unfälle und Verkehrsbeeinträchtigungen. Diese Kategorie betrifft Menschen die Hintergründe über Einsatzmaßnahmen, Geschädigte, Hilfegesuche, Verkehrsumleitungen und wetterbedingte Gefahrenmeldungen erhalten möchten. Es existiert ein reges Informationsinteresse im Zusammenhang mit dem Themengebiet Unwetter. Solche Wettererscheinungen betreffen große Bereiche und damit eine Vielzahl von Menschen und sind daher in der Tagespresse oft behandelt. In dieser Kategorie kann daher Informationsinteresse in hohem Maße durch mittelbare Betroffenheit folgen, was ihre Anwendung als Analysekategorie stützt. Dies zeigt sich auch in stetig ansteigender Sendezeit für Wetter in den letzten Jahren [18].[15] Darüber hinaus ist „Unwetter“, mit 700 tags, das häufigste Keyword in der Dokumentensammlung, denn 54 % der Pressemitteilungen die den Begriff enthalten wurden unter diesem Keyword vermerkt.

[15] In den folgenden Nachrichtensendungen: Tagesthemen, heute-journal, heute, RTL aktuell, Sat.1 Nachrichten

Kategorie - Verkehrsunfälle. Diese Kategorie umfasst hauptsächlich Verkehrsunfälle. Die Besonderheit hier liegt darin, dass sich die Meldungen auf Geschehen beziehen, die mit nur geringen Unterschieden mehrmals täglich auftreten. Verkehrsunfälle sind Alltag und entsprechend schwer zu differenzieren. Der Wortstamm „Verkehr“ ist in 27 % der Meldungen in der Dokumentensammlung enthalten und der Begriff „Verkehrsunfall“ in 16,6 %. Dies zeigt die besondere Häufung des Themas in den Pressemitteilungen.

Kategorie – allgemeine Kapitalverbrechen und Bandenkriminalität. Die letzte Kategorie umfasst häufig Ereignisse im Zusammenhang mit Banden- oder Gruppenkriminalität. Hierunter zählen auch Gruppenstreitigkeiten oder größere Ausschreitungen bei Großveranstaltungen mit Körperverletzungen und Verstößen gegen das Waffengesetz. Die Kategorie umfasst verhältnismäßig wenig Meldungen. Begriffe im Zusammenhang mit dieser Kategorie haben eine Häufigkeit von 1 % oder weniger in den Meldungen. Sie dient damit auch als Kontrast zur häufigen Kategorie Verkehrsunfälle. Darüber hinaus sind Ereignisse mit vielen Mitwirkenden und Zeugen häufig von großem medialen Interesse und werden meist über einen längeren Zeitraum hinweg verfolgt, als beispielsweise ein Verkehrsunfall mit Todesopfern. Viele Banden werden erst nach langer Ermittlungsarbeit gefasst und sind manchmal über Jahre kriminell aktiv gewesen. Online-Nachrichtenartikel zu dieser Kategorie sind daher leichter zu finden, als Artikel beispielsweise zu einzelnen Sportveranstaltungen mit Ausschreitungen.

3.3. Archivierung der Topics

Für alle Kategorien sind jeweils 15 Online-Nachrichtenartikel gefunden worden. Über jedes in einem dieser Online-Nachrichtenartikel beschriebene Ereignis gibt es mindestens eine relevante Polizeipressemitteilung in der Dokumentsammlung. Wichtigster Bestandteil, für die Archivierung nach dem Cranfield Paradigma, ist die URL der Nachrichtenartikel, welche der Suchmaschine als Eingabe dient. Damit die dahinter stehende Website für die darauffolgenden Analysen immer im gleichen Zustand betrachtet werden kann, wird die URL aller verwendeten Artikel im Internet Archive: Wayback Machine gespeichert.[16] Diese persistenten URLs sind für jedes Topic zusätzlich zur jeweiligen URL mit angegeben. Die Sammlung der peristenten URLs

[16] `https://archive.org/web/` [Zugriff 23.09.2019]

Tabelle 2: Schwierigkeitsbewertungen leicht (1), mittel (2) oder schwer (3) für alle Online-Nachrichtenartikel vorab Einschätzungen nach Kategorien.

Kategorie	**Level**	**Anzahl**	**Gesamtschwierigkeit**
Unwetter	1	7	28
	2	6	
	3	2	
Diebstahl	1	9	23
	2	4	
	3	2	
Migration	1	4	28
	2	9	
	3	2	
Sportveranstaltung	1	2	32
	2	9	
	3	4	
Mord	1	4	30
	2	7	
	3	4	
Verkehrsunfälle	1	2	28
	2	13	
	3	0	
Kapitalverbrechen	1	2	28
	2	13	
	3	0	

der Artikel ist auch im Web Archive öffentlich zugänglich.[17] Jedem Topic ist eine Identifikationsnummer zugewiesen, welcher die Nummer der Kategorie voransteht, beispielsweise `5_1` (5 für Kategorie 5-Mord und 1 für den ersten Topic aus 15 dieser Kategorie). Darüber hinaus wurde vorab eine intuitive menschliche Einschätzung des Schwierigkeitsniveaus vermerkt. Alle Topics wurden einer der Schwierigkeitsstufen leicht (1), mittel (2) oder schwer (3) zugewiesen. Die Einschätzung beruht auf der Ähnlichkeit zu einer bekannten Pressemitteilung aus der Sammlung und unterlag dabei folgenden Kriterien:

[17] Für diese Bachelorarbeit erstelltes Web Archive. Verfügbar unter `https://archive.org/details/@sunny_prm?tab=web-archive`[Zugriff 23.09.2019]

1. Das Vokabular des Artikeltitels stimmt überwiegend mit dem der Pressemitteilung überein. Der Artikeltext beschreibt das Ereignis konkret. Datum des Artikels ist nicht mehr als zwei Monate nach oder zwei Wochen vor der Publikation der Pressemeldung.

2. Artikeltitel stimmt in wenigen Wörtern (ohne Stoppworte) mit dem der Pressemitteilung überein. Der Artikeltext beschreibt das Ereignis intuitiv weniger konkret.

3. Der Artikel erläutert sehr malerisch das Ereignis enthält beispielsweise Zeugen/Passanten Interviews die tief berührt, betroffen, schockiert sind. Zusätzlich ist der Artikel insgesamt umfangreich formuliert oder aber auf ein mindestens ein Jahr zurückliegendes Ereignis bezogen, beispielsweise bei Hinweisen auf Gendenkveranstaltungen zu einer zurückliegenden Gewalttat. Oder die entsprechende Pressemitteilung ist in einer „Multimeldung“ enthalten, d.h. es werden mehrere unterschiedliche Ereignisse der gleichen Dienststelle in einer Meldung zusammengefasst.

In Tabelle 2 sind die Anzahlen pro Kategorie zusammengefasst. Die durchschnittliche Schwierigkeitseinschätzung pro Kategorie ist in etwa $1,9$ (mittel). Während die Kategorien Verkehrsunfälle und Kapitalverbrechen/Bandenkriminalität hauptsächlich Topics der Stufe mittel enthalten sind die meisten Artikel der Stufe schwer in den Kategorien Sportveranstaltungen und Mord zu finden. Diebstahl ist die Kategorie mit den meisten als leicht eingestuften Online-Nachrichtenartikeln.
Ein Topic im HTML-Format ist beispielhaft in Abbildung 3 zu sehen. In dieser Formatierung liegen alle 105 Topics vor.

```
<category>Mord</category>

<topicId>5_5</topicId>

<url>https://www.derwesten.de/staedte/duisburg/cafe-vivo-in-duisburg-geschaefts-
fuehrerin-46-wurde-erschossen-angestellte-fand-die-leiche-id210456067.html</url>

<persistentUrl>https://web.archive.org/web/20190905203934/https://www.derwesten.
de/staedte/duisburg/cafe-vivo-in-duisburg-geschaeftsfuehrerin-46-wurde-erschossen
-angestellte-fand-die-leiche-id210456067.html</persistentUrl>

<difficulty>2</difficulty>
```

Abbildung 3: Beispiel eines Topics

4. Erstellung der Suchmaschine

Das folgende Kapitel beschreibt die Schritte die notwendig sind, um die Eingabe von Suchenden automatisch in eine geeignete Anfragestrategie an Elasticsearch mit der indexierten Dokumentensammlung zu übersetzen. Die Struktur und die Bestandteile der Anfragestrategien werden kurz erläutert. Dabei wird die Implementierung der Anfragestrategien in Elasticsearch beschrieben. Es werden zwei Anfragestrategien vorgestellt im späteren Verlauf dieser Bachelorarbeit werden auch Anpassungen dieser Strategien beschrieben. Die hier genannten dienen deshalb dem Verständnis der technischen Realisierung der Anfragestrategien in Elasticsearch. Alle verwendeten Methoden werden an diesen Beispielen erläutert.

4.1. Web-Scraping

Für die technische Umsetzung einer Suchmaschine ist die Schnittstelle zwischen Suchenden und Suchmaschine entscheidend. Die Query muss so formuliert werden, dass eine effektive Suche im Index der Suchmaschine möglich ist. Im Fall des „Query by Document“-Ansatzes können sogar mehrere Querys aus der ursprünglichen Eingabe erzeugt werden. In der Polizeipressemitteilungssuchmaschine wird eine URL als Eingabe verlangt. Um die benötigten Informationen zur Erstellung einer Indexanfrage aus der Eingabe-URL zu erhalten, wird Web-Scraping[18] verwendet. Hierzu sind die Textbestandteile Artikeltitel, Artikeltext, Erscheinungsdatum und Ort der eingegebenen Website extrahiert worden. Dies ist jedoch nicht für jede Website möglich. Es kann auch vorkommen, dass Informationen fehlen, so ist des Öfteren kein Ort zu erfassen. Die Suche kann dann auch ohne diese Angaben erfolgen. Fehlen sehr viele Angaben, sind unpassendere Suchergebnisse zu erwarten. Die besten Suchergebnisse ergeben sich, wenn alle vier Textbestandteile erfolgreich extrahiert werden können. Es kann in der Regel immer mindestens ein Titel extrahiert werden.

Unabhängig von dieser automatischen Erfassung wurden für den Testdatensatz alle benötigten Textbestandteile manuell ergänzt und falls erforderlich manuell korrigiert, um die Anfragestrategien für die Suchmaschine für ein erfolgreiches Extrahieren der Daten zu evaluieren. Es wäre auch möglich, Nutzenden die Option zu bieten, die Query selbst zu korrigieren, das heißt gegebenenfalls manuell beispielsweise den Ort zu ergänzen, falls kein automatisches Extrahieren gelungen ist.

[18] Auch „web harvesting“ oder „data scraping“, gemeint ist die Extraktion von Daten aus dem Netz

```
1 GET /*/_search
2 { "query": {
3     "bool": {
4         "should":[
5             {"multi_match": { "query": "Artikeltitel", "fields": ["title^2", "body"]}}
6             {"match": {"body": "Artikeltext"}},
7             {"match": {"body": "Artikeldatum"}},
8             {"multi_match": {"query": "Artikelort", "fields": ["officeName", "body^5", "title"]}},
9 { "range": {
10    "timestamp":{
11        "time_zone": "+01:00",
12        "gte": "Artikeldatum-2w/w",
13        "lte": "Artikeldatum+8w/w"
14 }}}]}}}
```

Abbildung 4: Elasticsearch-Query der erweiterten Anfragestrategie.

4.2. Anfragestrategien

Abbildung 4 zeigt die erweiterte Anfrage an Elasticsearch. Es handelt sich dabei um eine sogenannte *Query context* Anfrage, die ermittelt, wie gut ein gegebenes Dokument zu den festgelegten Anfragebedingungen passt. Die *Query context* Anfrage berechnet im Unterschied zur *Filter context* Anfrage einen Score, der angibt, wie gut das Dokument relativ zu anderen Dokumenten die Anfragebedingungen der Query erfüllt.[19] Der Parameter `should` (Zeile 4) bedeutet, dass auch keine Bedingung erfüllt sein kann. Auch dann wird ein Score berechnet. Ein solches Dokument mit minimalem Score beeinflusst jedoch auch den Score der Dokumente, welche die Bedingungen erfüllen. Eine `multi_match` Anfrage erlaubt das Suchen einer Query in mehreren Indexfeldern. Zusätzlich können diese gewichtet werden. Die `multi_match` Anfrage in Abbildung 4 (Zeile 5) gewichtet die Felder Titel und Body im Verhältnis 2:1.[20] In den Zeilen 6 und 7 sind die `match` Anfragen zur Suche des Artikeltextes und des Artikeldatums im Body der Pressemitteilungen. Diese Anfrage die ist Standardabfrage zur Durchführung von Volltextanfragen. Für den Ort wird eine `multi_match` Anfrage verwendet, um den Ort im OfficeNamen, Body und Titel der Pressemitteilung zu suchen. Die vollständige Adresse der Dienststelle ist, sofern angegeben, im Body der Mitteilung enthalten. Zuletzt wird eine Anfrage vom Typ `range` genutzt um das geforderte Zeitfenster aus dem Datum zu konstruieren. Gesucht wird im Zeitraum zwei Wochen vor dem Artikeldatum bis acht Wochen danach. Für die Berechnung muss im `timestamp` Feld eine Zeitzone (Zeile 11) und die untere (Zeile 12) sowie die obere Zeitfenstergrenze (Zeile 13) mit angegeben werden. JSON verfügt über keinen Datentyp *date* deshalb sind Daten in Elasticsearch als *string* angegeben, die

[19] `https://www.elastic.co/guide/en/elasticsearch/reference/6.3/query-filter-context.html` [Zugriff 25.09.2019]

[20] `https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-multi-match-query.html` [Zugriff 25.09.2019]

```
GET /*/_search
{ "query": {
    "bool": {
        "should":[
            {"multi_match": { "query": "Artikeltitel", "fields": ["title^2", "body"]}}]}}}
```

Abbildung 5: Elasticsearch-Query der einfachen Anfragestrategie. Nur der Artikeltitel wird im Titel und im Body des Online-Nachrichtenartikels gesucht.

ein formatiertes Datum enthalten (z.B. “2018-07-05”oder “2018/07/05 11:14:20”).[21]

Die einfache Anfragestrategie ist in Abbildung 5 dargestellt und ist in der komplexen Anfragestrategie mit enthalten. Es handelt sich um eine `multi_match` Anfrage die nur den Artikeltitel im Titel und Body der Pressemitteilung ebenfalls als Anfrage des Typs *Query context* sucht und entsprechend einen Score für alle Dokumente in der Sammlung zurückliefert.

[21] `https://www.elastic.co/guide/en/elasticsearch/reference/current/date.html` [Zugriff 25.09.2018]

5. Evaluierung

Für die Evaluierung der Strategien wurden jeweils die ersten fünf Dokumente betrachtet. Diese fünf Dokumente sind dabei die mit dem höchsten Score absteigend sortiert. Sie wurden von mir als Assessor im Rahmen dieser Bachelorarbeit auf ihre thematische Relevanz in Bezug auf das im Topic beschriebene Ergebnis bewertet. Es wird dabei unterschieden zwischen nicht relevant (0), relevant (1) und höchst relevant (2). Die Benutzerrelevanz die eventuell über die thematische Relevanz hinaus gehen könnte, beispielsweise bei Interesse des Suchenden an ähnlichen Vorfällen, wird in dieser Bachelorarbeit nicht berücksichtigt.

5.1. Generierung der QRELs

Die sogenannten Qrel Dateien, die sich im Rahmen der TREC Evaluierungsgrundsätze entwickelt haben, sind ein Standard, um manuelle Relevanzbewertungen in ein vergleichbares Format zu fassen.[22] Es handelt sich um eine Textdatei, die pro Beurteilung eine Zeile mit der Identifizierung des Dokuments, des Topics und der gegebenen Relevanzbewertung enthält. Alle Ergebnisse von verschiedenen Systemen oder auch verschiedenen Anfragestrategien, die Bewertungen erhalten haben sind in dieser Datei zusammengefasst. Für die Bachelorarbeit existiert eine Qrel Datei, die alle Bewertungen beinhaltet. Ein Ausschnitt sieht dann beispielsweise wie folgt aus.

```
1_3 0 4969-1470606 0
1_3 0 4969-2819606 2
1_3 0 4969-2779247 1
```

In diesem Beispiel sind drei Pressemitteilungen als Ergebnisausgabe zu Topic `1_3`, dem dritten Online-Nachrichtenartikel der Kategorie Unwetter, bewertet worden. Die erste Pressemitteilung mit der ID 4969-1470606 ist als nicht relevant eingestuft. Die zweite ist höchst relevant, während die dritte relevant ist. Die Bewertungen sind in einer zufälligen Reihenfolge vorgenommen worden, um zu verhindern das die Einschätzungen hierdurch beeinflusst werden. Es handelt sich um ein depth-5 pooling [19]. Die zweite Zahl, gibt die Anzahl der Iterationen der Bewertung an. In der Regel ist dies immer eine Null (in dieser Bachelorarbeit immer Null), und wird bei der Evaluierung an keiner Stelle berücksichtigt.

[22] `https://trec.nist.gov/data/qrels_eng/`

5.2. Implementierung der Ergebnisrankings

Ein weiterer Schritt zur Evaluierung ist die automatische Generierung von Dateien, die die Dokument_IDs der Ergebnisdokumente in der entsprechenden Ausgabereihenfolge enthalten. Damit kann zusammen mit der vorhandenen Qrel Datei eine automatische Auswertung einer Anfragestrategie erfolgen. In dieser Bachelorarbeit wird der normalized Discounted Cumulated Gain (nDCG) der ersten fünf Dokumente des Rankings als Vergleichskriterium verwendet. NDCG misst den Nutzen eines Dokuments anhand seiner Position in der Ergebnisliste [20]. Der Nutzen wird von oben nach unten kumuliert (hier nur insgesamt 5 Dokumente), wobei der Gewinn jedes Ergebnisses auf den niedrigeren Rängen diskontiert wird. Ein weiterer Evaluationsparameter in der Auswertung ist *Precision at 1* (P@1 oder Precision@1), welcher die Relevanz des ersten Dokuments angibt. Entscheidend ist bei P@1 also allein, ob das Ergebnis mit dem höchsten Score relevant ist.

5.3. Pilotstudien

Damit die Entwicklungsnotwendigkeit der Pressemitteilungssuchmaschine gerechtfertigt werden kann, werden zunächst die intuitiven bestehenden Alternativen für eine solche Suchmasche näher betrachtet werden. Das ist zum einen die integrierte Suchfunktion der Website Presseportal und zum anderen eine Google-Suche nach einer entsprechenden Pressemitteilung. Beide Optionen wurden zunächst im Rahmen von Pilotstudien untersucht und sind nachfolgend kurz erläutert. Für die Pilotstudien wurde als Testdatensatz zwei Kategorien je 15 Topics ausgewählt. Die Wahl viel auf die Kategorien Mord und Diebstahl, deren Gesamtschwierigkeitslevel das schwerste beziehungsweise leichteste aller Kategorien abbilden.

Suchfunktion Presseportal: Für diese Pilotstudie wurden die Titel der Topics der jeweiligen Kategorien in der Suche des Presseportals in der Rubrik Blaulicht gesucht.[23] Um mit der entwickelten Polizeipressesuchmaschine vergleichbar zu sein, wurde das Datum auf die Zeit, für welche Pressemitteilungen in der Dokumentensammlung enthalten sind (01.01.2001-05.07.2018), eingeschränkt. Leider blieb die Suche für die meisten Topics erfolglos. In der schwierigeren Kategorie Mord generierten allein zwei von 15 Topics überhaupt Treffer, unter diesen Treffern ließen sich allerdings auch relevante Dokumente finden. Die Ergebnisse der als leichter ein-

[23] `https://www.presseportal.de/blaulicht/suche/` [Zugriff 20.09.2019]

Tabelle 3: Titel der Topics der Kategorie Diebstahl. Angegeben ist das vorab eingestufte Schwierigkeitslevel, Ausgabe der Presseportalsuchfunktion (ja oder nein) und die Anzahl der relevanten Ausgaben im Verhältnis zur Ausgaben insgesamt (relevant/gesamt).

Titel	Level	Ausgabe	relevant
Nümbrecht E-Bike aus Schaufenster geklaut	1	ja	1/1
Polizei ermittelt Grabschänder vom Nordfriedhof	2	ja	1/1
Zwei Einbrüche in Lagerhallen Fernseher und Virtual-Reality-Brille gestohlen	1	ja	1/1
Diebstahl auf Firmengelände	2	ja	0/3
Einbruch in Supermarkt	1	ja	0/6
Von Zivilfahndern gefasst Probefahrt endet mit Festnahme	2	nein	0/0
Einbrecher auf Baustelle stehlen Kabel, Werkzeug und Staubsauger	1	nein	0/0
Goldpaste gestohlen	1	nein	0/0
S-Nord/S-Mitte: Porsche für 160.000 Euro und Motorrad gestohlen	1	nein	0/0
Diebe machen fette Beute 5000 Kilogramm Nutella gestohlen	3	nein	0/0
Diebesbande soll 1600 Reifen gestohlen und verkauft haben	1	nein	0/0
Autoreifen und Alufelgen von 11 Neufahrzeugen gestohlen	1	nein	0/0
Polizei Bad Hersfeld schnappt Serientäter-Trio	1	nein	0/0
Polizei fasst Taschendiebe auf der Hannover Messe	2	nein	0/0
Mit Kran oder Klein-Lkw Langfinger betreiben großen Aufwand bei Diebstahl	3	nein	0/0

gestuften Kategorie Diebstahl sind im Vergleich zu den Ergebnissen der Kategorie Mord besser, wenn auch nicht vielversprechend. In Tabelle 3 sind die Titel der Kategorie Diebstahl zusammen mit den Ergebnissen der Suchfunktion des Presseportals aufgelistet. Mutmaßlich sucht die Website nach exakter Übereinstimmung, weshalb oft überhaupt keine Dokumente zurückgegeben werden können. Die Vermutung der exakten Übereinstimmung wird auch durch die folgende Beobachtung bestärkt. Für die Kategorien Mord und Diebstahl wird ein relevantes Dokument genau dann zurückgegeben, wenn auch nur exakt ein Dokument gefunden wurde. Werden hingegen mehrere Dokumente zurückgegeben ist keins davon relevant (vgl. Tabelle 3). Für die meisten Anfragen werden überhaupt keine Dokumente zurück gegeben.

Google-Suche: Um die gleichen Rahmenbedingungen für diese Pilotstudie zu schaffen werden Google-Suchoperatoren verwendet.[24] Die Einschränkungen auf die Website Presseportal.de/Blaulicht gelingt mit dem *site*-Operator. Für die Anpassung des Zeitfensters stehen mehrere Optionen zur Verfügung. Zum einen die manuelle Auswahl eines benutzerdefinierten Datumsbereichs. Verwendet man diese Möglichkeit, kann keine Pressemitteilungen auf eine Anfrage, bestehend aus Artikeltitel eines jeden Topics der Kategorie Mord, gefunden werden. Zum anderen besteht die Option den *daterange*-Operator zu verwenden. Dieser Operator verlangt die Ein-

[24] `https://www.sem-deutschland.de/seo-firma/seo-glossar/google-suchoperatoren/` [Zugriff 20.09.2019]

gabe als julianisches Datum (JD, englisch *Julian Date*).[25] Eine solche Suchanfrage beschreibt sich beispielsweise wie folgt:

```
Einbruch in Supermarkt daterange:2451911.35679-2458315.35679
site:presseportal.de/blaulicht/
```

Auch mit dieser Option werden oft keine Dokumente ausgegeben. Darüber hinaus funktioniert die zeitliche Einschränkung durch den *daterange*-Operator leider nicht und es werden auch neuere Ergebnisse mit ausgegeben. Damit sind diese Ausgaben nicht mit denen der entwickelten Suchanfragestrategien sinnvoll vergleichbar.
Google bietet mit dem großen Datumsbereich noch eine weitere Möglichkeit das Datum für die Suchergebnisse einzugrenzen. Hier ist jedoch nur eine Begrenzung für Jahre möglich, beispielsweise 2001..2018 für die Jahre 2001 bis einschließlich 2018. Aber auch diese Option gibt Pressemitteilungen über das einschränkende Zeitfenster hinaus zurück.

Zusammenfassend muss auf ein Vergleich der entwickelten Suchanfragestrategien mit der Google-Suche verzichtet werden, da keine Möglichkeit mittels Google-Operatoren gefunden werden konnte, um die Suche so einzuschränken, dass die Ergebnisse vergleichbar sind. Die Ergebnisse der Suchfunktion des Presseportals sind hingegen unter vergleichbaren Rahmenbedingungen entstanden. Die Resultate sind jedoch nicht besonders aussichtsreich da oft überhaupt keine Dokumente zurückgegeben werden.

5.4. Ergebnisse

Einige Ergebnisse der Evaluierung der verschiedenen Anfragestrategien auf dem Index der Polizeipressemitteilungen (PM) sind nachfolgend beschrieben. Zunächst ist in Tabelle 4 die Bewertung verschiedener Anfragestrategien nach Kategorien dargestellt. Es fällt auf, dass die Original Suchfunktion des Presseportals (OPP) deutlich schlechtere Suchergebnisse liefert als Anfragen an die entwickelte Suchmaschine. Selbst die einfache Anfragestrategie, die nur den Artikeltitel im Titel und im Body der Pressemitteilung sucht ist der Suchfunktion des Presseportals überlegen. Weiter fällt auf, dass die Suchergebnisse der Suchstrategie Titel und Body der OPP schlechter sind als die alleinige Suche des Titels, während sich die Ergebnisse der ent-

[25] `https://moz.com/blog/mastering-google-search-operators-in-67-steps`[Zugriff 20.09.2019]

Tabelle 4: Evaluation verschiedener Anfragestrategien: Titel (T), Titel und Body (TB), und die Kombination Titel, Body, Ort und Datum (TBOD) für verschiedene Kategorien. Es wird die Polizeipressesuchmaschine (PPS) mit der Original Suchfunktion des Presseportals (OPP) anhand von nDCG@5 und precision@1 verglichen.

		Alle Kategorien		Unwetter		Diebstahl		Migration	
Methode	Zeit	nDCG	P@1	nDCG	P@1	nDCG	P@1	nDCG	P@1
T@OPP	0,7s	0,13	0,14	0,13	0,10	0,27	0,27	0,16	0,20
TB@OPP	0,9s	0,04	0,04	0,00	0,00	0,07	0,07	0,07	0,07
T@PPS	0,6s	0,21	0,21	0,26	0,27	0,24	0,20	0,23	0,20
TB@PPS	9,2s	0,75	0,76	0,52	0,53	0,93	0,93	0,72	0,80
TBOD@PPS	9,1s	0,87	0,88	0,62	0,73	0,90	0,87	0,83	0,87
		Sport		Mord		Verkehrsunfall		Kapitaldelikt	
Methode	Zeit	nDCG	P@1	nDCG	P@1	nDCG	P@1	nDCG	P@1
T@OPP	0,7s	0,00	0,00	0,09	0,13	0,00	0,00	0,31	0,27
TB@OPP	0,9s	0,00	0,00	0,00	0,00	0,00	0,00	0,13	0,13
T@PPS	0,6s	0,10	0,07	0,20	0,33	0,08	0,13	0,34	0,27
TB@PPS	9,2s	0,87	0,90	0,47	0,53	0,72	0,73	1,00	1,00
TBOD@PPS	9,1s	0,95	0,87	0,80	0,87	0,97	1,00	1,00	1,00

wickelten Suchmaschine mit Hinzunahme des Bodys entscheidend verbessern (vgl. Tabelle 4). Für die Kategorien Sport und Verkehrsunfall werden mit der OPP überhaupt keine Dokumente gefunden. Diese beiden Kategorien schneiden auch bei der Titelsuche in der PPS deutlich schlechter ab, als die übrigen Kategorien. Das bestätigt die Vermutung, dass die OPP auf einem Exact-Match-Paradigma basiert, also nur 100%ige Treffer zurück liefert. Dies begründet warum die in der entwickelten Suchmaschine genutzte teilweise Übereinstimmungs Methode (Partial-Match), die von relevant bis weniger relevant abstuft (Score), diese deutlich erhöhte Ranking-Qualität bedeutet. Mit der komplexen Anfragestategie, die Titel, Body, Datum und Ort berücksichtigt, werden die besten Rankings erzeugt und häufig ein relevantes Dokument auf dem ersten Rang zurückgeliefert. Die Kategorie Kapitaldelikte erreicht sogar das auffällig optimale Ergebnis von 1 für nDCG@5 und Precision@1. Ebenso sticht die Kategorie Verkehrsunfälle mit den guten Ergebnissen heraus. Diese beiden Kategorien sind aber auch die, die keine Topics des Schwierigkeitslevels schwer (3) enthalten. Um diese Beobachtung zu analysieren sind in Tabelle 5 nDCG@5 und Precision@1 für die selben Anfragestrategien für die verschiedenen Schwierigkeitsstufen angegeben. In dieser Übersicht bestätigt sich, dass das Schwierigkeitslevel

Tabelle 5: Evaluation der verschiedenen Anfragestrategien: Titel (T), Titel und Body (TB), und die Kombination Titel, Body, Ort and Datum (TBOD) für verschiedene Schwierigkeitsstufen. Es wird die Polizeipressesuchmaschine (PPS) mit der Original Suchfunktion des Presseportals (OPP) verglichen in Bezug auf nDCG@5 und precision@1.

		Alle Level		Level 1		Level 2		Level 3	
Methode	Zeit	nDCG	P@1	nDCG	P@1	nDCG	P@1	nDCG	P@1
T@OPP	0,7s	0,13	0,14	0,28	0,31	0,09	0,10	0,00	0,00
TB@OPP	0,9s	0,04	0,04	0,10	0,10	0,02	0,02	0,00	0,00
T@PPS	0,6s	0,21	0,21	0,44	0,48	0,14	0,11	0,04	0,07
TB@PPS	9,2s	0,75	0,76	0,82	0,86	0,78	0,79	0,47	0,47
TBOD@PPS	9,1s	0,92	0,88	0,93	0,90	0,94	0,98	0,59	0,60

die Suchschwierigkeit annähernd widerspiegelt. Für das schwierige Level 3 sind die ausgegeben Dokumente häufiger nicht relevant, beziehungsweise sind die relevanten Ergebnisse seltener auf dem ersten Rang. Das Kriterium wonach das Vokabular des Artikeltitels der Tpoics des leichten Schwierigkeitslevels überwiegend mit dem der Pressemitteilung übereinstimmt spiegelt sich in den Evaluierungen der Titelsuchanfragen (T@OPP und T@PPS) wieder. Denn relevante Dokumente zu den Level 1 Topics können durch diese Strategie im Vergleich zu Level 2 und 3 mit einer höheren Wahrscheinlichkeit gefunden werden.

Ein weiterer angegebener Vergleichsparameter ist die durchschnittliche Laufzeit pro Suchanfrage. Mit komplexer werdender Anfragestrategie verlängert such die Laufzeit. Positiv fällt auf, dass die Titelsuchanfrage der entwickelten Suchmaschine nicht nur bessere Ergebnisse zurückliefert, sondern auch $0,1$ s schneller ist als die Titelsuche in der originalen Suchfunktion des Presseportals. Die komplexen Anfragestrategien mit ca. 9 s Laufzeit erfordern Geduld bei interessierten Suchenden, lassen sich aber durch die guten Resultate rechtfertigen.

Das eine Anfragestrategie unter Berücksichtigung des Bodys deutlich bessere Ergebnisse im Vergleich zu einer kürzeren Anfrage (nur Titel) bedeutet, ist durch die Resultate in Tabelle 5, also durch die Verbesserung des nDCG@5 und P@1, um 54 % beziehungsweise 55 %, zweifelsfrei gezeigt. Auch die darüber hinausreichende Optimierung um weitere 17 % (nDCG@5) beziehungsweise 12 % (P@1) ist nicht zu

Tabelle 6: Evaluation der verschiedenen Anfragestrategien der Einbindung des Orts. Alle Auswertungen beziehen sich auf die entwickelte Suchmaschine. Anfrage ohne Ort (TBD). Varianten: Ort nur im Title (O@T), Ort nur im Body (O@B), Ort nur im Office Namen (O@ON). Eine Gewichtung der Felder wird im Exponenten angegeben z.B. O@TB^{2}ON für Ort im Titel, Body und Office Namen im Verhältnis 1:2:1. Angeben sind nDCG@5 und Precision@1 für verschiedene Schwierigkeitslevel.

		Alle Level		Level 1		Level 2		Level 3	
Methode	Zeit	nDCG	P@1	nDCG	P@1	nDCG	P@1	nDCG	P@1
TB@PPS	9,2s	0,75	0,76	0,82	0,86	0,78	0,79	0,47	0,47
TBD@PPS	8,6s	0,87	0,91	0,90	0,93	0,93	0,98	0,55	0,53
TBOD@PPS	9,1s	0,91	0,88	0,93	0,90	0,97	0,93	0,60	0,60
O@T	8,8s	0,86	0,90	0,90	0,93	0,97	0,93	0,52	0,53
O@B	8,7s	0,88	0,92	0,90	0,93	0,94	0,98	0,60	0,60
O@ON	8,7s	0,88	0,91	0,90	0,93	0,94	0,98	0,58	0,53
O@B^{2}ON	8,9s	0,88	0,92	0,90	0,93	0,94	0,98	0,61	0,60
O@TB^{2}ON	8,9s	0,88	0,92	0,90	0,93	0,94	0,98	0,61	0,60
O@TB^{5}ON	9,0s	0,87	0,88	0,86	0,90	0,94	0,97	0,60	0,47

verkennen. Allerdings bietet die komplexe Anfragestrategie mit Ort und Datum mehr Modifizierungsmöglichkeiten, weshalb diese Optionen im Folgenden noch detaillierter beschrieben werden.

Suche unter Verwendung des Ortes

In Tabelle 6 sind verschiedene Anfragestrategien untersucht, die den Ort unterschiedlich in die komplexe Suchanfrage integrieren. Insbesondere die Topics des Schwierigkeitslevels 3 bieten noch Optimierungspotential. Eine Suchstrategie ohne Verwendung des Ortes liefert auch sehr gute Ergebnisrankings für die Schwierigkeitslevel leicht und mittel. Die Precison@1 ist für die Schwierigkeitslevel 1 und 2 sogar höher im Vergleich zur komplexen Anfrage mit Ort (vgl. Tabelle 6 TBOD@PPS). Für Schwierigkeitslevel 3 sind die Ergebnisse jedoch sowohl für nDCG und P@1 schlechter. Möglich ist es den Ort im Titel, im Body oder im Office Namen der Pressemitteilungen zu suchen. Deshalb sind diese drei Fälle probiert und evaluiert worden (vgl. Tabelle 6). Für die Suche des Ortes im Titel der Pressemitteilung verschlechtert sich die Ranking-Qualität für die Level 2 und 3 Topics, und damit auch insgesamt, im Vergleich zur Anfragestrategie ohne Ort. Die Ergänzung um die Suche des Orts im Body der Pressemitteilung verbessert das Ergebnis hingegen bedeutend für die

schweren Topics. Für die Topics der mittleren Schwierigkeitsstufe kann eine Verbesserung des nDCG im Vergleich zur Strategie ohne Ort erreicht werden. Die letzte Möglichkeit besteht in der Suche des Ortes im Office Namen (Dienststellen Adresse) der Pressemitteilungen. Auch diese Option verbessert das Ergebnis-Ranking. In diesem Fall erhöht sich der nDCG für die mittleren und schweren Topics (um 1 %), während alle anderen Parameter keine Änderung anzeigen.

Die soeben betrachtete Evaluierung hat gezeigt, dass vor allem die Suche des Ortes im Body eine Verbesserung zur Folge hat. Aber auch die Suche im Office Namen wirkt sich geringfügig positiv auf die Ranking-Qualität aus. Deshalb ist nachfolgend eine stärkere Gewichtung des Ortes im Body bei gleichzeitiger Suche des Ortes im Office Namen in Tabelle 6 beschrieben. Ebenfalls wurde diese doppelte Gewichtung des Ortes im Body bei gleichzeitiger Suche im Titel und im Office Namen evaluiert. Ein Vergleich lässt keine Unterscheidung zu.

Suche unter Verwendung des Datums

Das Datum wird in zwei unterschiedlichen Formen in die komplexe Suchanfragestrategie integriert. Zum einen wird das Datum als String, wie es von der Website der Anfrage-URL extrahiert werden konnte im Body der Pressemitteilung gesucht. Zum anderen wird das Datum in ein von Elasticsearch gefordertes Datumsformat (ISO 8601) konvertiert und ein Zeitfenster um diesen Bereich gewählt. Es werden dann die Dokumente der Sammlung bevorzugt, welche im Rahmen dieses Zeitfensters veröffentlicht wurden. Gesucht wird hierfür im Feld *published* der Dokumente.

Die bisher vorgestellten Evaluierungen haben immer mit der Bezeichnung Datum (D) beide Angaben (Datum und Zeitfenster: Zeitraum zwei Wochen davor bis acht Wochen danach) zusammengefasst verwendet. Im Folgenden wird dies differenzierter untersucht. In Tabelle 7 sind die Evaluierungsparameter verschiedener Anfragestrategien die das Datum berücksichtigen angegeben. Wird nur das Datum oder nur das Zeitfenster berücksichtigt sind die Resultate für Topics der leichten und schweren Kategorie ähnlich. Für die Topics der mittleren Kategorie unterscheiden sie sich jedoch geringfügig, so ist die P@1 für die Datumssuche erfolgreicher als die des Zeitfensters. Umgekehrt wird mit dem Zeitfenster ein bessere Ranking-Qualität erreicht, denn der nDCG@5 ist hier höher. Dafür gibt es eine einfache Begründung. Die Suche des Datums im Body der Pressemeldung ist vergleichbar mit einer Exact-Match-Anfrage,

Tabelle 7: Evaluation der verschiedenen Anfragestrategien der Einbindung des Datums (D) und des Zeitfensters (Z). Für das Zeitfenster sind immer die Wochenanzahlen von/bis bezüglich des Datums angegeben. Alle Strategien berücksichtigen Titel (T), Body (B) und Ort (O). Angegeben wird nDCG@5 und Precision@1 für alle Schwierigkeitslevel. Es wird nur die Polizeipressesuchmaschine betrachtet.

		Alle Level		Level 1		Level 2		Level 3	
Methode	Zeit	nDCG	P@1	nDCG	P@1	nDCG	P@1	nDCG	P@1
TBO	9,2s	0,87	0,91	0,88	0,93	0,93	0,97	0,60	0,60
TBOZ2/8	8,5s	0,87	0,91	0,89	0,93	0,97	0,93	0,60	0,60
TBOD	8,8s	0,88	0,92	0,90	0,93	0,94	0,98	0,60	0,60
TBODZ2/2	8,8s	0,88	0,92	0,90	0,93	0,94	0,98	0,60	0,60
TBODZ2/0	8,8s	0,88	0,92	0,90	0,93	0,94	0,98	0,60	0,60
TBODZ2/8	9,1s	0,91	0,88	0,93	0,90	0,97	0,93	0,60	0,60

die entweder trifft oder fehlschlägt. Kann das Datum im Body gefunden werden ist die Wahrscheinlichkeit ein relevantes Dokument gefunden zu haben hoch. Im Unterschied dazu ist das angegebene Zeitfenster eher mit einer Partial-Match-Anfrage vergleichbar, welche mehrere Möglichkeiten zulässt. Deshalb ist es nachvollziehbar, dass in diesem Fall der nDCG@5 das höhere Ergebnis erzielt.

Tabelle 7 gibt auch die Auswertungen für verschiedene Zeitfenster an. Auch hier zeigt sich der zuvor beschriebe Effekt, wird das Zeitfenster kleiner steigt P@1 und der nDCG@5 verringert sich. Für die Topics des schweren Level 3 haben die verschiedenen Strategien des Datums und des Zeitfensters keine erkennbaren Auswirkungen auf die Ranking-Qualität.

6. Zusammenfassung und Ausblick

Im Rahmen der vorliegenden Bachelorarbeit konnte gezeigt werden, dass unter der Verwendung von Elasticsearch eine Suchmaschine entwickelt werden kann, die eine deutlich höhere Ranking-Qualität aufweist als die Suchfunktion der Website Presseportal. In die entwickelte Suchmaschine können leicht zu bedienen Online-Nachrichtenartikel eingegeben und die dem Nachrichtenereignis zuzuordnenden relevanten Polizeipressemitteilungen ausgegeben werden. Dazu wurden die Pressemitteilungen zunächst von der Website gecrawlt und für die Nutzung in Elasticsearch indexiert. Die aus der URL des Nachrichtenartikels bestehende Anfrage des Nutzenden wird mittels Web-Scraping in Artikeltitel, Artikeltext, Artikeldatum und Artikelort segmentiert. Es verschiedene Anfrage-Strategien im Rahmen einer an das Cranfield-Paradigma angelehnten Studie verglichen. Hierzu wurde eine Testdokumentensammlung von 105 Online-Nachrichtenartikeln erstellt und evaluiert. Die Ergebnisse der Studie zeigen, dass lange Anfragen, die den Artikeltext beinhalten, deutlich bessere Ergebnis-Rankings erzeugen, als Anfragen, die nur den Artikeltitel berücksichtigen. Auch ohne Verwendung des Orts können gute Ergebnisse für Topics erzielt werden, deren Vokabular und Gesamtwörteranzahl intuitiv überwiegend mit dem der zum Ereignis relevanten Pressemitteilungen übereinstimmt. Schwieriger sind Topics zu finden die stark vom gebrauchten Vokabular der Polizeipressemitteilungen abweichen. Hier verbessert die Integration von Datum und Ort die Ranking-Qualität. Interessant wäre eine weiterführende systematische Analyse der Topics der schweren Level, um die Anfragestrategie in diesem Bereich weiter zu verbessern. Hier verbirgt sich das größte Potential. Dafür müsste die Auswahl der Topics dieser Schwierigkeitsstufe erweitert werden (bisher insgesamt 15 Level 3 Topics). Weitere Optionen zur Verbesserung bietet die Laufzeit, die sich derzeit auf ca. 9s beläuft und vermutlich noch nicht der angestrebten Benutzungsfreundlichkeit genügt.

Literatur

[1] Antonis Kalogeropoulos, Rasmus Kleis Nielsen, Nic Newman und Richard Fletcher. Reuters Institute Digital News Report 2019. Reuters Institute for the Study of Journalism, 2019.

[2] David Tewksbury und Scott L. Althaus. Differences in Knowledge Acquisition among Readers of the Paper and Online Versions of a National Newspaper. *Journalism & Mass Communication Quarterly*, 77(3):457–479, 2000.

[3] Edson C. Tandoc Jr., Zheng Wei Lim und Richard Ling. Defining "Fake News". *Digital Journalism*, 6(2):137–153, 2018.

[4] Suhang Wang, Jiliang Tang, Kai Shu, Amy Sliva und Huan Liu. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explorations Newsletter*, 19(1):22–36, September 2017.

[5] Markus Prior. Any Good News in Soft News? The Impact of Soft News Preference on Political Knowledge. *Political Communication*, 20(2):149–171, 2010.

[6] Irene Costera Meijer. The Paradox of Popularity. *Journalism Studies*, 8(1):96–116, 2007.

[7] Kim Christian Schrøder. What Do News Readers Really Want to Read About? how Relevance Works for News Audiences. Digital News Publications, 2019. `http://www.digitalnewsreport.org/publications/2019/news-readers-really-want-read-relevance-works-news-audiences/`.

[8] European Commission. Wie sehr vertrauen Sie der Polizei? Statista, 2019. `https://de.statista.com/statistik/daten/studie/377233/umfrage/umfrage-in-deutschland-zum-vertrauen-in-die-polizei/`.

[9] Donna Harman. *TREC-Style Evaluations*, pages 97–115. Springer Berlin Heidelberg, 2013. In: Agosti M., Ferro N., Forner P., Müller H., Santucci G. (eds) Information Retrieval Meets Information Visualization.

[10] Christopher D. Manning, Prabhakar Raghavan und Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[11] W. Bruce Croft, Donald Metzler und Trevor Strohman. *Search Engines – Information Retrieval in Practice*. Pearson Education, 2009.

[12] Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas und Dimitris Papadias. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 34–43, New York, NY, USA, 2009. ACM.

[13] Ali Dasdan, Paolo D'Alberto, Santanu Kolay und Chris Drome. Automatic Retrieval of Similar Content Using Search Engine Query Interface. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 701–710, New York, NY, USA, 2009. ACM.

[14] National Institute of Standards and Technology. Overview. `https://trec.nist.gov/overview.html`, April 2019. [Zugriff 24.09.2019].

[15] Franz Faul, Edgar Erdfelder, Axel Buchner und Albert-Georg Lang. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160, 2009.

[16] *Polizeiliche Kriminalstatistik*, volume 1. Bundeskriminalamt, 2018. S.12 `https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/PolizeilicheKriminalstatistik/2018/pks2018Jahrbuch1Faelle.pdf?__blob=publicationFile&v=6` [Zugriff am 23.09.2019].

[17] Thorsten Quandt. (no) news on the world wide web? *Journalism Studies*, 9(5):717–738, 2008.

[18] IFEM. InfoMonitor 2018. `https://www.ard-werbung.de/fileadmin/user_upload/media-perspektiven/pdf/2019/0219_Krueger_Zapf-Schramm_2019-02-13.pdf`, 2019. [Zugriff am 24.09.2019].

[19] J. Shane Culpepper, Xiaolu Lu und Alistair Moffat. The effect of pooling and evaluation depth on IR metrics. *Information Retrieval Journal*, 19(4):416–445, June 2016.

[20] Kalervo Järvelin und Jaana Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

A. Allgemeine Ergänzungen

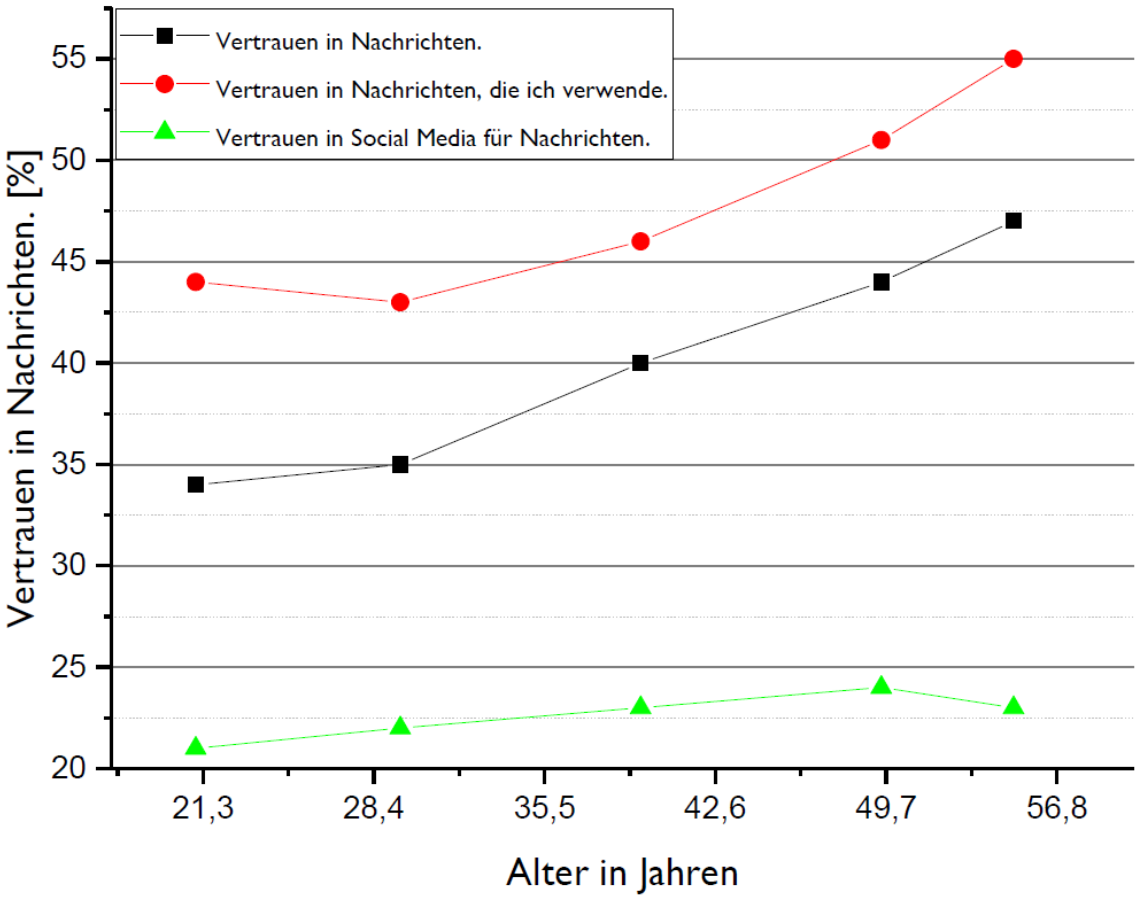

Abbildung 6: Anzahl der durch das Reuters Institut befragten Menschen in % die den Aussagen „Ich vertraue in Nachrichten“ (schwarz), „Ich vertraue in Nachrichten die ich verwende.“ (rot) oder „Ich vertraue in Social Media für Nachrichten“ (grün) zustimmen oder nachdrücklich zustimmen.[1]

B. Erklärung

Hiermit erkläre ich, Nina Katharina Schwanke, dass ich die Arbeit selbstständig verfasst habe, sie in gleicher oder ähnlicher Fassung noch nicht in einem anderen Studiengang als Prüfungsleistung vorgelegt habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.
Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach anderen gedruckten oder im Internet verfügbaren Werken entnommen sind, habe ich durch genaue Quellenangaben kenntlich gemacht.

Halle, 07.11.2015 N. Schwanke

Ort, Datum Nina Katharina Schwanke

C. Danksagungen

An dieser Stelle möchte ich mich bei allen bedanken, die mich bei der Anfertigung dieser Bachelorarbeit unterstützt, begleitet und motiviert haben. Ich möchte mich vor allem bei meinem Betreuer Herrn Prof. Dr. Matthias Hagen für die Ermöglichung dieser Bachelorarbeit, die Hinweise und das Beantworten meiner vielen Fragen bedanken. Zudem möchte ich mich bei Herrn Maik Fröbe für die Hilfe, Betreuung und praktischen Hinweise während der Gestaltung dieser Bachelorarbeit bedanken. Weiterhin möchte ich Prof. Dr. Martin Potthast für das Erstellen eines Zweitgutachtens, seine Anregungen und die Bereitstellung des Pilotprojektes, auf dem diese Bachelorarbeit aufbaut, danken. Ich danke auch den Studierenden der Universität Leipzig die dieses Projekt bearbeitet haben.

Mein besonderer Dank gilt meiner Familie, die mir mein Studium ermöglicht und mich in meinen Entscheidungen unterstützt hat. Weiterhin geht mein Dank an meine Freunde: Vielen Dank für euer offenes Ohr, Anregungen, Zuversicht und konstruktive Kritik.