



UNIVERSITÄT
LEIPZIG

UNIVERSITÄT LEIPZIG

ABTEILUNG AUTOMATISCHE SPRACHVERARBEITUNG

FORTGESCHRITTENE METHODEN DES INFORMATION RETRIEVAL

Bericht zum Laborprojekt

Jeremy Puchta – Jonathan Lange – Ali Al-Ali

11. März 2019

Inhaltsverzeichnis

1	Einleitung	1
2	Technologiestack	1
3	Datenakquise und -analyse	2
3.1	Datenverarbeitung	2
3.2	Textstatistiken	3
4	Architektur	4
4.1	Indizierungsprozess	4
4.2	Suchprozess	4
5	Vorstellung der Suchmaschine	4
5.1	Landing Page	4
5.2	Auflistung der Suchergebnisse	5
5.3	Detailansicht	5
6	Evaluation	5
7	Zusammenfassung und Ausblick	5

1 Einleitung

Der vorliegende Bericht erläutert den Aufbau sowie die Funktionsweise der Suchmaschine *Historical News Search*, welche im Rahmen des Laborprojektes innerhalb des Moduls *Fortgeschrittene Methoden des Information Retrieval* im Wintersemester 2018 / 2019 erstellt wurde. Ziel des Laborprojektes war es, die vermittelten Vorlesungsinhalte zu vertiefen und diese bei der Erstellung einer domänenspezifischen Suchmaschine umzusetzen. Die vorgestellte Suchmaschine dient der Exploration von historischen Nachrichten. Verschiedene Szenarien sind als Anwendungsfälle für eine solche Suchmaschine denkbar. Ein Nutzer kann nach der Berichterstattung zu Personen, Gruppen, Orten oder bestimmten historischen Ereignissen suchen. Eine weitere Möglichkeit der Nutzung stellt die Ahnenforschung dar, bei welcher ein Nutzer in Zeitungsberichten nach Informationen über seine Vorfahren suchen kann und somit mehr über seine Familienhistorie erfahren kann.

Im folgenden Kapitel wird zunächst näher auf den Datensatz eingegangen und dieser mit Methoden der deskriptiven Statistik analysiert. Die Architektur und der Aufbau der Suchmaschine werden in Kapitel 3 näher dargestellt. Dazu zählt insbesondere die Erstellung des Indizes sowie die Umsetzung des Suchprozesses über alle beteiligten Komponenten des Systems. In Kapitel 4 werden zunächst die Evaluationsmethodik sowie die Resultate der Evaluation erläutert. Weiterhin wird dargestellt, welche Anpassungen vorgenommen wurden, um die Effektivität der Suchmaschine zu verbessern. Abschließend erfolgt in Kapitel 5 eine Zusammenfassung des Projektes und es wird ein Ausblick in Bezug auf eine Weiterentwicklung der Suchmaschine geliefert.

2 Technologiestack

Zur Umsetzung des Laborprojektes wird die Open-Source-Suchmaschine *Elasticsearch* verwendet. *Elasticsearch* ist in Java geschrieben und basiert auf der Java-Bibliothek *Lucene*. Außerdem ist *Elasticsearch* gebaut für den Einsatz auf verteilten Systemen und die Echtzeitverarbeitung von großen Datenmengen, was unter anderem zur Volltextsuche eingesetzt wird und *Elasticsearch* somit zu einer ausgezeichneten Option für die Umsetzung einer Dokumentensuchmaschine macht. *Elasticsearch* stellt sämtliche Funktionen über eine programmiersprachenunabhängige REST-Schnittstelle zur Verfügung. Um die Dokumente durchsuchbar zu machen, legt *Elasticsearch* die Datenstruktur des *Invertierten Index* an, in welcher für jeden Term gespeichert wird in welchem Dokument dieser auftritt.

Im Backend des Systems kommt das Python-Framework *Flask* zum Einsatz. Bei *Flask* handelt es sich um ein Mikroframework, welches eine einfache, jedoch ebenfalls erweiterbare und robuste Möglichkeit bietet APIs zu erstellen. Im Frontend wird mit *Angular* ein Framework verwendet, welches Entwickler durch sein Komponentensystem darin unterstützt modularen Quellcode zu produzieren, der sich einfach warten lässt. Außerdem bringt es bereits viele Funktionen, die häufig im Frontend zum Einsatz kommen mit, wie zum Beispiel Routing. Zur Vereinfachung des Deployments und zur Optimierung

des Entwicklungsprozesses kommt die Containervirtualisierungs – Technologie *Docker* zum Einsatz.

3 Datenakquise und -analyse

Die *Staatsbibliothek zu Berlin* besitzt ein breites Spektrum von historisch bedeutsamen digitalisierten Zeitungen. Diese werden im hauseigenen Zeitungsinformationssystem namens *ZEFYS* kostenfrei bereitgestellt. Die Digitalisate, Volltexte und Metadaten der *Berliner Volks-Zeitung (BVZ)* dienen als Korpus für die im Rahmen des Laborprojektes entwickelte Suchmaschine. Bei der *Berliner Volks-Zeitung* handelt es sich um eine von 1904 bis 1944 veröffentlichte regionale deutsche Tageszeitung aus Berlin. Sie besitzt große Bedeutung für die Forschung im Bereich der Kulturwissenschaften, da sie im Gegensatz zu den meisten linken Parteizeitungen, über ein gutes Feuilleton verfügt. [QUELLE] Der Zeitraum der Digitalisate beläuft sich auf die Jahre 1890 bis 1930, wobei einige Jahre stärker abgedeckt sind als andere. Die Datensets umfassen jeweils strukturierte Metadaten im METS-XML-Containerformat für jede Ausgabe, per OCR erzeugte Volltexte im ALTO-XML-Format mit Wortkoordinaten, binarisierte TIFFs als Grundlage der OCR sowie JPEG2000-Bilder für die Anzeige. Insgesamt umfasst der Datensatz 103.771 digitalisierte Seiten.

3.1 Datenverarbeitung

Mithilfe des ALTO-XML-Formates wird die Möglichkeit geschaffen Digitalisate vollständig und originalgetreu zu replizieren. Aus diesem Grund sind neben Zeitungsinhalten Informationen zum Layout der Seite und die exakten Koordinaten der Wörter enthalten. Dokumente werden von *Elasticsearch* im JSON-Format repräsentiert. Daher ist es erforderlich die im Datensatz vorhandenen XML-Dateien in das JSON-Format zu konvertieren und dabei relevante Informationen zu extrahieren. Relevante Informationen stellen beispielsweise das Veröffentlichungsdatum der Zeitung, weitere Informationen wie die Ausgabe und den gesamten Text einer Zeitungsseite dar. Listing 1 zeigt ein Beispiel für die konvertierte JSON-Repräsentation einer Zeitungsseite. Interessant ist insbesondere der Schlüssel "Text", der ein dreifach geschachteltes Array bestehend aus den Texten einer Seite enthält. Innerhalb des Arrays der ersten Ebene befinden sich die einzelnen Artikel einer Seite, welche wiederum Arrays mit den Wörtern der Zeilen eines Artikels enthalten. Durch diese Repräsentation ist man in der Lage die einzelnen Artikel statt einer gesamten Seite als Ergebnis einer Suchanfrage auszugeben.

```

1 {
2   "Year": "1930",
3   "Month": "01",
4   "Day": "03",
5   "NewspaperNumber": "001",
6   "PageNumber": "004",
7   "Edition": "0",
8   "Issue": "059",
9   "Text": "[ // Gesamter Zeitungstext
10             [ // Artikel
11               ["Nach","Schluss","der","Redaktion","
12                 eingetroffene","Depeschen."],
13               ["Dortmund","31.","Maerz.","Wie","die
14                 ","Rheinisch-Westfaelische"]
15             ],
16             [
17               (...)
18             ]
19           ]"
20 }

```

Listing 1: JSON-Repräsentation einer Zeitungsseite

Bei näherer Betrachtung der Daten lässt sich feststellen, dass sich eine große Anzahl von OCR-Scanartefakten in den Texten befindet. Diese rühren von der Interpretation des Scanners der altdeutschen Schrift. Es lässt sich feststellen, dass die Daten weniger OCR-Scanartefakte enthalten je jünger die Zeitungen sind. Die vollständige Auflösung dieser Scanartefakte stellt eine unverhältnismäßig große Herausforderung dar, weshalb sich gegen die Durchführung der Artefaktsauflösung entschieden wird. Stattdessen erfolgt bei der Auswahl der Referenzmenge für das Repository eine stärkere Ausrichtung an den jüngeren Digitalisaten.

3.2 Textstatistiken

Um ein in sich geschlossenes Projekt im öffentlichen Repository zu präsentieren, wird der gesamte Datensatz auf eine geringere Größe reduziert. Tabelle 1 zeigt ausgewählte Textstatistiken für diese reduzierte Dokumentkollektion.

Gesamtanzahl Dokumente	6,144
Gesamtanzahl Worte	44,337,561,540
Vokabulargröße	1,663,267
Wörter mit Vorkommen > 1000	1,461,311
Wörter mit einmaligen Vorkommen	499

Tabelle 1: Textstatistiken für Dokumentkollektion

4 Architektur

In diesem Kapitel wird zunächst der Technologiestack des Systems beschrieben. Außerdem erfolgt die detaillierte Vorstellung des User Interfaces mit Erklärung der einzelnen Sichten bevor abschließend der Indizierungs- und der Suchprozess erläutert werden. Abbildung 1 visualisiert die einzelnen Komponenten des Systems mittels UML-Komponentendiagramm.

// UML-KOMPONENTENDIAGRAMM ARCHITEKTUR DER SUCHMASCHINE

4.1 Indizierungsprozess

Die im Rahmen der Datenvorverarbeitung erstellten JSON-Dokumente werden mit Elasticsearch indiziert. Jede Seite einer Zeitung wird dabei als eine JSON-Datei repräsentiert und enthält neben dem Veröffentlichungsdatum und weiteren Metadaten (Ausgabe, Seitenzahl, ...) die Zeitungsartikel. Diese stellen den wichtigsten Bestandteil für die Indizierung dar, da in diesem relevante Terme enthalten sind, die von einem Nutzer gesucht werden können. Wie jedoch bereits im Kapitel 2.1 dargelegt wurde, ist die Qualität der Daten durch eine Vielzahl von OCR-Scanartefakten sehr gering.

4.2 Suchprozess

5 Vorstellung der Suchmaschine

Nachdem die Architektur des Systems in Kapitel 3 vollumfänglich erläutert wurde, erfolgt in diesem Kapitel die Vorstellung der Suchmaschine, insbesondere des User Interfaces. Das User Interface besteht aus drei verschiedenen Sichten, die sich wie folgt darstellen:

1. Landing Page mit Eingabefeld für eine Suchanfrage
2. Seite zur Auflistung der Suchergebnisse
3. Seite zur Darstellung von Detailinformationen zu einer Zeitungsseite inklusive des zugehörigen Zeitungsscans

Bei der Beschreibung der Sichten erfolgt eine Erklärung der Prozesse die im System ablaufen.

5.1 Landing Page

Bei der *Landing Page* handelt es sich, wie der Name schon sagt, um die Seite, die ein Nutzer sieht sobald dieser die Suchmaschine verwendet. Auf dieser befindet sich neben einem thematisch passenden und ansprechenden Hintergrundbild ein Eingabefeld, in welcher die Nutzer eine Suchanfragen stellen können. Abbildung 1 zeigt die *Landing Page* mit dem Eingabefeld für Suchanfragen.



Abbildung 1: Landing Page der Historical News Search

5.2 Auflistung der Suchergebnisse

Nachdem ein Nutzer seine Suchanfrage in das Eingabefeld auf der *Landing Page* eingegeben hat, wird diese an die `search`-Methode im Backend übergeben. Wie bereits in Kapitel 3.2 beschrieben sucht diese in der Dokumentkollektion nach Zeitungsseiten, welche die Terme aus der Suchanfrage enthalten und gibt diese zurück. Die zurückgegebenen Zeitungen werden, nach ihrem Gewicht sortiert, in einer Liste von *Cards* dargestellt. Dabei wird eine Zeitungsseite von je einer *Card* repräsentiert. Diese enthalten neben einer Überschrift und dem Veröffentlichungsdatum ein kurzes Snippet, um dem Nutzer eine Vorschau auf die Artikel der Zeitungsseite zu bieten und diesen somit bei der Suche nach relevanten Ergebnissen zu unterstützen. Weiterhin ist zur Gewährleistung der Übersichtlichkeit des User Interfaces eine *Paginierung* implementiert. Jede Seite enthält besitzt zehn Resultate. Abbildung 2 illustriert die Auflistung der Suchergebnisse für eine Suchanfrage nach `max schmeling`.

// ABBILDUNG EINFÜGEN!

5.3 Detailansicht

Klickt ein Nutzer auf eine der *Cards* aus der Ergebnisliste, werden ihm sämtliche Artikel angezeigt, die auf der entsprechenden Zeitungsseite stehen.

// WIRD NACH UMBAU DER DETAILANSICHT FORTGEFÜHRT

// ABBILDUNG EINFÜGEN SOBALD FERTIGGESTELLT

6 Evaluation

7 Zusammenfassung und Ausblick

Literatur