



UNIVERSITÄT
LEIPZIG

UNIVERSITÄT LEIPZIG

ABTEILUNG AUTOMATISCHE SPRACHVERARBEITUNG

FORTGESCHRITTENE METHODEN DES INFORMATION RETRIEVAL

Bericht zum Laborprojekt

Jeremy Puchta – Jonathan Lange – Ali Al-Ali

5. März 2019

Inhaltsverzeichnis

1	Einleitung	1
2	Datenakquise und -analyse	1
3	Architektur	2
4	Evaluation	3
5	Zusammenfassung und Ausblick	4

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. Als Beispiele für Suchanfragen können hierfür **reichstagswahl 1930**, **max schmeling** oder **nsdap** genannt werden. 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 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-Formate mit Wortkoordinaten, binarisierte TIFFs als Grundlage der OCR sowie JPEG2000-Bilder für die Anzeige. Insgesamt umfasst der Datensatz 103.771 digitalisierte Seiten.

3 Architektur

4 Evaluation

5 Zusammenfassung und Ausblick

Literatur