Knjigica – Eine Libretto-Suchmaschine

Lukas Neef
Martin Luther University Halle-Wittenberg
Halle (Saale), Germany
lukas.neef@student.uni-halle.de

Jan Heinrich Reimer
Martin Luther University Halle-Wittenberg
Halle (Saale), Germany
jan.reimer@student.uni-halle.de

ZUSAMMENFASSUNG

Die Recherche und Suche von Libretti – Textvorlagen einer Oper – ist umständlich, da keine umfassende Suchmaschine dazu zur Verfügung steht. Allgemeine Websuchmaschinen können die vielen Details nur unvollständig erfassen, sodass sie als Suchmaschine für Libretto schlecht geeignet sind. Wir entwerfen eine umfassende Suchmaschine für Libretti, die alle im Libretto enthaltenen Informationen abbildet, und dadurch sowohl beim Besuch einer Oper, als auch in der Literaturwissenschaft der Librettologie, Bedeutung findet.

KEYWORDS

search engine, libretto, opera

1 EINLEITUNG

Die Textvorlage einer Oper, das Libretto, ist eine wichtige Quelle, sowohl für Besucher einer Oper, sowie für die Literaturwissenschaft. Für die schnelle Recherche von Libretti steht bisher keine Suchmaschine zur Verfügung. Wir entwerfen daher eine Suchmaschine, mit der sich Libretti schnell anhand ihres Textes oder ihrer Eigenschaften finden lassen.

Durch diese neue Suchmaschine, sind zahlreiche Verbesserungen bei der Arbeit mit Libretto zu erwarten: So ist es beispielsweise hilfreich, wenn der Text aus anderen Sprachen übersetzt wird, um ein besseres Verständnis für die Inhalte der Handlung zu bekommen. Auch bekommen Besucher die Möglichkeit, mit Hilfe des Textes der Handlung besser folgen zu können und nicht nur dem Geschehen auf der Bühne zuzuschauen. Dies ist auch der Fall, wenn während der Aufführung ein bestimmter Teil nicht verstanden wurde, weil die Präsentation oder Aussprache undeutlich war, oder durch starke Betonung unverständlich wurde.

Mit Hilfe der Suche im Libretto können des Weiteren auch zusätzliche Informationen zum Autor oder dem Jahr der Erstaufführung recherchiert werden, was in der Vor- und Nachbereitung eines Besuches einer Opernaufführung von Bedeutung ist.

Die Möglichkeiten der Benutzung dieser Suchmaschine sind also vielfältig und können sogar der Forschung in der Librettologie zugutekommen.

2 DATENQUELLEN

Im ersten Schritt der Umsetzung ist eine umfassende Recherche nach Libretto und deren Texten im Internet erfolgt. Als Ergebnis der Suche wurden die in Tabelle 1 aufgeführten Datenquellen ausgemacht.

Zeitgeschichtlich wurde der überwiegene Teil der in den Corpora enthaltenen Libretti im Zeitraum vom 17. bis 20. Jahrhundert uraufgeführt. Die meisten der Corpora enthalten Libretti verschiedener Sprachen, überwiegend jedoch italienische, französische, englische, deutsche oder russische. Auch werden bei einigen Corpora Übersetzungen zur Verfügung gestellt, so werden beispielsweise die Libretti von „Opera Glass“ durch 389 Übersetzungen ergänzt. Bei „Kareol“ sogar die Übersetzung dem Originaltext gegenübergestellt.

Corpus	Anzahl Libretto
Kareol	385
Opera Lib	434
Opera Folio	512
Opera Glass	170
Aria	177
Collection Ulric Voyer	220

Tabelle 1: Libretto-Corpora

3 STRUKTUR

3.1 Literarische Struktur von Libretto

Im geschichtlichen Verlauf der Zeit haben sich Libretti stetig weiterentwickelt. Daher ist es schwierig, für ein Libretto eine feste Struktur festzulegen. Allgemein sin jedoch einige Bestandteile in jedem Libretto vorhanden. So wird anfangs der Titel und die Autoren genannt. Hier werden auch oftmals weitere Informationen, wie Daten zur Premiere, sowie weitere Notizen oder Widmungen des Autors genannt. Nachfolgend werden die im Libretto vorkommenden Rollen mit Namen, teils mit Beschreibung, sowie mit der vorgesehenen Stimmlage, aufgeführt. Anschließend folgt die Handlung des Libretto. So werden die Überschriften, sowie Unterüberschriften genannt. Textpassagen werden, meist mit dem Namen der vortragenden Rolle versehen, einfach aufgelistet. Zwischen Textpassagen stehen gelegentlich Regieanweisungen zur Betonung, Informationen zum Bühnenbild, sowie weitere Instruktionen. Mit der Handlung endet das Libretto.

3.2 Hilfsdatenstruktur

Aus der allgemeinen Struktur von Librettos leiten wir für die Entwicklung der Suchmaschine ein standardisiertes und maschinell lesbares Format ab. Dieses dient dem effizienten Speichern der gerawlten Rohdaten und der Vorbereitung sowie Standardisierung für die spätere erfolgende Indizierung durch die Suchmaschine.

Neben einer Datenstruktur, die ein vollständiges Libretto abbildet, entwerfen wir Klassen, die einzelne Teile der Struktur speichern. In einer Datenstruktur werden alle Informationen zum Autor gespeichert, in einer zweiten Informationen über Widmungen und Notizen des Autors, in einer dritten Informationen zu einer Rolle bzw. einem Darsteller, und in einer letzten werden alle die Handlung betreffenden Informationen abgespeichert. Dabei wird diese letzte in sich in drei Datenstrukturen unterteilt, für Textpassagen, Instruktionen, sowie für Überschriften.

4 TECHNISCHE IMPLEMENTIERUNG

Die Implementierung der Libretto-Suchmaschine unterteilt sich in mehrere Bausteine. Während der Crawler vor allem auf Linux-Bash-Programmen aufbaut, wird in den anderen Bausteinen, so

auch im Parser, Indexer und im Benutzerprogramm auf objektorientierte Programmierung mittels Java/Kotlin gesetzt. Als Framework wird das praxiserprobte Elasticsearch-Framework verwendet. Im Folgenden werden die einzelnen Komponenten der Suchmaschine näher erklärt.

4.1 Crawler

Der von uns entwickelte Crawler hat die Aufgabe, die Corpora in 2 aus dem Internet herunterzuladen, und für die weitere Verarbeitung zu speichern.

Zuerst werden dazu die URLs aller Libretti aus dem Register eines jeden Corpus extrahiert und als Liste zwischengespeichert. Um Probleme mit relativen URLs auszuschließen, werden diese auf die absolute Adresse erweitert.

Danach lädt ein weiteres Skript systematisch die HTML-Dokumente aller erfassten URLs herunter und speichert diese zunächst unverändert ab. Hierbei werden auch verlinkte, verschachtelte Seiten zunächst mit übernommen.

Beide Vorgänge wurden mit Hilfe von Bash-Skripten umgesetzt, die die Linux-Funktionen `wget`, `curl`, sowie die `html-xml-utils` verwenden.

4.2 Parser

Die Inhalte der Internetseite liegen nach dem Crawlen als HTML-Quelltext vor. Damit in der Folge nach enthaltenen Informationen gesucht werden kann, werden die Inhalte mittels Java in ein Objekt geschrieben. Hierbei wurde von uns die Bibliothek „jsoup“ verwendet, welche die Extraktion und Bearbeitung von HTML Inhalten ermöglicht. Entsprechend wird das gecrawlte Material je nach Internetseite so verallgemeinert, dass anhand bestimmter Abfolgen von HTML Elementen die Hilfsdatenstruktur korrekt gefüllt wird. Ein Formatter sorgt abschließend dafür, dass die in der Hilfsdatenstruktur enthaltenen Informationen in ein JSON Dokument umgewandelt werden. Dieses wird von der Suchmaschine als Referenz verwendet.

5 SUCHMASCHINE

Als Suchmaschine verwenden wir die open source Software „Elasticsearch“. Diese beruht auch Apache Lucene und realisiert die Kommunikation mit Klienten über ein RESTful-Webinterface. Das vorher geparste Dokument im JSON Format wird an die Suchmaschine gesendet und wird dort indiziert. Ein entsprechendes Dokument besitzt in unserem Fall die folgende verkürzte Form:

Die Dokumente liegen im Anschluss aufgelistet in Elasticsearch vor und können über einen sogenannten Index von der Suchmaschine erreicht werden. Eine Suchanfrage wird wie die Inhalte zuvor mittels eines JSON Dokuments an Elasticsearch gesendet. Die Anfrage kann verschiedene Parameter enthalten, welche die Resultate mit Filtern versehen und Spezifikationen wie eine bestimmte Ordnung der Inhalte und Beschneidung auf bestimmte Stichworte erlauben. In unserem Fall wurde die Suchanfrage als Boolean umgesetzt, wodurch eine Ausgabe nur dann erfolgt, wenn ein oder mehrere der angegebenen Variablen mit einem Wert gefüllt wurden. Bei diesen Variablen handelt es sich um den Rollennamen, Regieanweisungen, Akte, Szenen und den Text.

6 INDIZIEREN

7 EVALUATION

Die Evaluation wurde anhand des Vorbildes der „Text Retrieval Conference“ (TREC) durchgeführt. In diesem Rahmen wurden 25 sogenannte Topics entworfen, welche die Suchmaschine auf ihre Funktionalität untersucht. Ein etwaiges Topic besteht aus einer zentralen Suchanfrage, aus der weitere Unteranfragen resultieren. Für die Bewertung der Resultate wurde eine Skala von -1 bis 1 verwendet:

-1 nicht relevant, beziehungsweise kein hilfreiches Resultat
0 relevant, erfüllt jedoch nicht den kompletten Umfang der Suchanfrage
1 sehr relevant und erfüllt die Suchanfrage in vollem Umfang

Die Anfragen und Ergebnisse der Evaluation der letzten Version der Suchmaschine werden in folgender Tabelle zusammengefasst:

Die Ergebnisse der Evaluation sind entsprechend folgend verteilt:

-1 11 Anfragen
0 8 Anfragen
1 5 Anfragen

Mit der derzeitigen Version der Suchmaschine werden also bei mehr als 50

8 ZUSAMMENFASSUNG UND AUSBLICK

8.1 Corpus

Obwohl wir bisher nur für den „Opera Lib“-Corpus einen Parser geschrieben haben, lässt sich das entwickelte Framework problemlos auf weitere Corpora erweitern. So planen wir beispielsweise, die

Eine weitere Erweiterungsmöglichkeit besteht darin, Kunstlieder mit in die Suchmaschine einzubeziehen. Der „Kareol“-Corpus stellt dazu 687 Lieder zur Verfügung.

https://aws.amazon.com/de/elasticsearch-service/what-is-elasticsearch/
https://de.wikipedia.org/wiki/Elasticsearch