Knjigica – Eine Libretto-Suchmaschine

LUKAS NEEF, Martin Luther University Halle-Wittenberg
JAN HEINRICH REIMER, Martin Luther University Halle-Wittenberg

To do (1)

Additional Key Words and Phrases: search engine, libretto, opera

1 EINLEITUNG

Im Zuge der Lehrveranstaltung „Websuche- und Information Retrieval“ haben wir in der spezifizierten Suche von Libretto festgestellt, dass diese ausbaufähig ist. Durch die Umsetzung besserer Möglichkeiten der Suche von Libretto sind vielseitige Vorteile als Resultat 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 Menschen mit einem schlechten Gehör die Möglichkeit, mit Hilfe des Textes der Handlung besser folgen zu können und nicht nur dem Geschehen auf der Bühne zuschauen zu müssen. Dies ist auch anderweitig der Fall, wenn während der Aufführung ein bestimmter Teil nicht verstanden wurde, weil die Präsentation oder Aussprache sehr undeutlich war oder stark betont wurde.

Mit Hilfe der Suche im Libretto können des Weiteren auch sämtliche Informationen wie Autor oder das Jahr der Erstaufführung recherchiert werden, was in der Vor- und Nachbereitung eines Besuches der Aufführung von Bedeutung ist. Die Möglichkeiten der Benutzung dieser Suchmaschine sind also vielfältig und können sogar der Forschung 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 folgende Datenquellen ausgemacht:

Webseite	Status	Anzahl Libretto	Sonstiges
Kareol	Gecrawlt	385 Libretto, 687 Lieder	
Opera Lib	Indiziert	434 Libretto	Stücke mit Prämieren in den Jahren 1600-1929
Opera Folio	Noch nicht benutzt	512 Libretto	Italienische, französische, englische, deutsche Libretto vom 17. bis 2
Opera Glass	Noch nicht benutzt	170 Libretto	65 Komponisten, 389 Übersetzungen
Aria	Noch nicht benutzt	177 Libretto	
Collection Ulric Voyer	Noch nicht benutzt	220 Libretto	

3 DATENSTRUKTUR

3.1 Struktur von Libretto

Im Verlauf der Zeit haben sich Libretto stetig weiterentwickelt. Daher ist eine feste Struktur der Texte nicht bestimmbar. Grob lässt sich der Aufbau eines Librettos jedoch mit folgender Abfolge definieren:

(1) Prolog
(2) Erster Akt

Authors' addresses: Lukas Neef, Martin Luther University Halle-Wittenberg, Universitätsring 3, Halle (Saale), Germany, 06108, lukas.neef@student.uni-halle.de; Jan Heinrich Reimer, Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, Halle (Saale), Germany, 06120, jan.reimer@student.uni-halle.de.

(3) Intermezzo
(4) Zweiter Akt
(5) Epilog

In den uns vorliegenden Quellen erfolgt zudem vor Beginn des Textes eine Angabe zusätzlicher Informationen wie Erstaufführung, Sprache, Zusammenfassung der Handlung oder veröffentlichte Medien, welche das Libretto beinhalten. Des Weiteren wird zu Beginn der jeweiligen Handlung die Rollen vorgestellt, welche das Libretto begleiten und meist auch in welcher Stimmlage diese singen. Nicht notwendigerweise wird die Handlung auch von einem Erzähler begleitet.

3.2 Hilfsdatenstruktur

Aus der allgemeinen Struktur von Librettos wurde für die Suchmaschine ein standardisiertes Format abgeleitet. Dieses dient dem effizienten speichern der Daten und der Vorbereitung und Standardisierung für die folgende Suche. Zu diesem Zweck wurden mehrere Klassen entworfen, welche die Informationen der Librettos auf Kerninformationen eingrenzen. Der Einfachheit halber ist eine Unterteilung in Haupt- und Nebenklassen erfolgt. Zu den Nebenklassen gehören der Autor, die Annotationen, der Plot, Informationen zur Premiere und den agierenden Rollen. Vereint werden diese in der Hauptklasse „Libretto“.

4 TECHNISCHE IMPLEMENTIERUNG

Mit Hilfe welcher Tools haben wir die im letzten Kapitel angegebene Hilfsdatenstruktur erreicht? Die Implementierung der Suchmaschine ist mit Hilfe mehrere Schritte erfolgt, welche im Folgenden näher erklärt werden sollen.

4.1 Crawler

Um an die Inhalte beider Seiten zu gelangen und somit eine Grundlage für die spätere Suche zu schaffen, wurde ein Crawler konzipiert, welcher systematisch die HTML-Seiten aller Libretti erfasst und diese zunächst unverändert abspeichert. Dieser Vorgang wurde mit Hilfe von Bash Skripten umgesetzt und basiert somit auf dem kopieren der im Internet aufgeführten HTML Quelltexten. Da die von uns recherchierten Internetseiten meist über ein Register verfügen, werden die in den Quelltexten aufgeführten Links mit der Kennzeichnung „href“ extrahiert und in einer neuen Datei abgespeichert. Um Probleme mit Links aufgrund relativer Adressen auszuschließen, werden diese um die der absoluten Adresse erweitert. Nachdem alle auf der Seite befindlichen Inhalte erfasst wurden, wird mit Hilfe des Programms „wget“ jede referenzierte Seite heruntergeladen.

4.2 Parser

Die Inhalte der Internetseite liegen nach dem Crawlen nun in 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:

```
{
  "title": "Die Entführung aus dem Serail",
  "subtitle": "Deutsche Singspiel.",
  "language": "de",
  "authors": [
    {
      "name": "BRETZNER",
      "fullName": "Christoph Friedrich BRETZNER",
      "scopes": [
        "TEXT"
      ]
    },
    // ...
    {
      "name": "MOZART",
      "fullName": "Wolfgang Amadeus MOZART",
      "scopes": [
        "MUSIC"
      ]
    }
  ],
  "annotations": [],
  "roles": [
    {
      "name": "SELIM",
      "description": "SELIM Bassa"
    },
    {
      "name": "KONSTANZE",
      "description": "KONSTANZE Geliebte des Belmonte",
      "voice": "SOPRANO"
    }
  ],
  "plot": [
```

```
35    //...
36    {
37      "type": "SECTION",
38      "data": {
39        "section": "Erster Auftritt",
40        "level": "SCENE"
41      }
42    },
43    {
44      "type": "INSTRUCTION",
45      "data": {
46        "instruction": "Platz vor dem Palast des Bassa am Ufer des Meeres. ↩
      Belmonte allein."
47      }
48    },
49    {
50      "type": "TEXT",
51      "data": {
52        "roleName": [],
53        "text": "Hier soll ich dich denn sehen;\nKonstanze! dich mein Glück↩
      !\nLaß Himmel es geschehen!\nGib mir die Ruh zurück.\nIch duldete der ↩
      Leiden\nO Liebe! allzuviel!\nSchenk mir dafür nun Freuden\nUnd bringe ↩
      mich ans Ziel!"
54      }
55    },
56    //...
57    {
58      "type": "SECTION",
59      "data": {
60        "section": "Zweiter Auftritt",
61        "level": "SCENE"
62      }
63    },
64    {
65      "type": "TEXT",
66      "data": {
67        "roleName": [
68          "OSMIN"
69        ],
```

```
70          "text": "Wer ein Liebchen hat gefunden,\nDie es treu und redlich ↩
      meint,\nLohn' es ihr durch tausend Küsse,\nMach ihr all das Leben süße,\↩
      nSei ihr Tröster, sei ihr Freund.\nTrallalera, trallalera.",
71          "instruction": "mit einer Leiter, welche er an einen Baum vor der ↩
      Tür des Palastes lehnt, hinaufsteigt und Feigen abnimmt"
72        }
73      }
74      // ...
75    ]
76 }
```

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

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

TO DO…

- □ 1 (p. 1): Add abstract.