**Martin-Luther-Universität Halle (Saale)**

****

Dokumentation der Studienleistung der Lehrveranstaltung Websuche- und Information Retrieval

Name: Erik Kinder

E-Mail: erik.kinder@student.uni-halle.de

Matrikelnummer:

Name: Jan Heinrich Reimer

E-Mail: jan.reimer@student.uni-halle.de

Matrikelnummer: 216204166

Name: Lukas Neef

E-Mail: lukas.neef@student.uni-halle.de

Matrikelnummer: 215246264

**Inhaltsverzeichnis**

[**1. Einleitung** 2](#_Toc536521453)

[**2. Datenquellen** 2](#_Toc536521454)

[**3. Datenstruktur** 3](#_Toc536521455)

[**3.1. Struktur von Libretto** 3](#_Toc536521456)

[**3.2. Hilfsdatenstruktur** 4](#_Toc536521457)

[**4. Technische Implementierung** 4](#_Toc536521458)

[**4.1. Crawler** 4](#_Toc536521459)

[**4.2. Parser** 4](#_Toc536521460)

[**5. Suchmaschine** 5](#_Toc536521461)

[**6. Indizieren** 5](#_Toc536521462)

[**7. Evaluation** 5](#_Toc536521463)

[**8. Zusammenfassung und Ausblick** 5](#_Toc536521464)

# **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  Beispiele für Autoren und Komponisten |
| Opera Folio | Noch nicht benutzt | 512 Libretto | Italienische Libretto von 18. bis 20. Jhd.  Französische Libretto vom 17. bis 20. Jhd.  Englische Libretto vom 18. Jhd.  Deutsche Libretto vom 18. bis 19. Jhd. |
| Opera Glass | Noch nicht benutzt | 170 Libretto | 65 Komponisten  389 Übersetzungen |
| Aria | Noch nicht benutzt | 177 Libretto | Creator = Rob Glaubitz |
| 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
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 Walküre",  
"subtitle": "Erster Tag des Bühnenfestspiels Der Ring des Nibelungen.",  
"language": "de",  
"authors": [  
 {  
 "name": "WAGNER",  
 "fullName": "Wilhelm Richard WAGNER",  
 "scopes": [  
 "TEXT",  
 "MUSIC"  
 ]  
 }  
],

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 Topics entworfen, welche die Suchmaschine auf ihre Funktionalität untersuchen. Ein etwaiges Topic besteht aus einer zentralen Suchanfrage, aus der weitere Unteranfragen resultieren können. Um wichtige Aspekte der Suchmaschine abzudecken, werden die Topics „Autor“, „Struktur“ und „Plot“ verwendet. Die Resultate der Suchanfragen werden anhand einer Skala von 0 bis 1 erfasst:

* 0: nicht relevant, beziehungsweise kein hilfreiches Resultat
* 0,2: relevant, erfüllt jedoch nur einen sehr geringen Umfang der Suchanfrage
* 0,4: relevant, erfüllt jedoch nur einen geringen Umfang der Suchanfrage
* 0,6: relevant, erfüllt die Suchanfrage in befriedigendem Umfang
* 0,8: sehr relevant, die Suchanfrage erfüllt die gestellten Kriterien
* 1: sehr relevant und erfüllt die Suchanfrage in vollem Umfang

Mit der Hilfe eines unabhängigen Dritten wurden alle Anfragen angewendet, um die Suchmaschine zu testen. Der Prüfende hatte hierfür Zugriff auf die Kommandozeilenanwendung und konnte Rückfragen stellen, sofern ihm bestimmte Aspekte unklar gewesen sind. Die für ihn zutreffenden Einträge der Skala wurden während der Durchführung zusammen mit einem Kommentar erfasst, welcher Aufschluss über die Einordnung der vorgegebenen Skala gibt. Die Anfragen und Ergebnisse der Evaluation der letzten Version der Suchmaschine sind im Anhang „tabelle\_evaluation“ hinterlegt.

Im Mittel ergeben die Resultate einen Wert von 0,344, was eine (sehr) geringe Erfüllung der Suchanfragen bedeutet. Hierbei sind die Mittelwerte der jeweiligen Topics in Autor: 0,51, Sektion: 0 und Plot: 0,4 aufgeteilt. Die Topics „Autor“ und „Plot“ liefern also beide bessere Suchergebnisse, was im derzeitigen Stand der Umsetzung zu begründen ist. Die Suchmaschine ist in der Lage nach spezifischen Informationen zu suchen, berücksichtigt jedoch noch keine Operatoren oder Suchanfragen, welche eine Umwandlung der Inhalte anhand einer bestimmten Fragestellung enthält. Dies führt auch dazu, dass die Intention einer Suchanfrage nicht erkannt wird und die eigentlich richtig eingeordneten Suchergebnisse mit falschen Parametern ausgegeben werden.

# **8. Zusammenfassung und Ausblick**

<https://aws.amazon.com/de/elasticsearch-service/what-is-elasticsearch/>

<https://de.wikipedia.org/wiki/Elasticsearch>