IT 3 - Suchmaschinentechnologie

Tutorial VR_SOLR

Klaus Lepsky / Jens Wille

5. Januar 2016 (v0.6)
Institut für Informationswissenschaft
Technische Hochschule Köln
Claudiusstraße 1, 50678 Köln
klaus.lepsky@th-koeln.de

Zusammenfassung

Dieses Skript enthält die Aufgabenstellung für den Programmteil *Solr* des Moduls IT 3 „Suchmaschinentechnologie“. Geübt werden sollen die Installation und die Benutzung der Suchmaschinen-Software *Solr*. Im Mittelpunkt stehen Aufbereitung, Import und Indexierung von unterschiedlichen Dokumentkollektionen und Dokumenttypen.

1 Einführung und Softwareinstallation

Solr ist eine *Java*-basierte Suchmaschinenlösung, die auf der ebenfalls *Java*-basierten Indexierungsmaschine *Lucene* aufsetzt. Für eine Übersicht der Eigenschaften und Möglichkeiten des Systems, vgl. https://lucene.apache.org/solr/features.html.

Solr benötigt eine *Java*-Laufzeitumgebung (Version 1.7 oder höher), die auf den Laborrechnern bereits vorhanden ist. Für die Installation und den ersten Test von *Solr* folgen Sie den Schritten dieses Tutorials: https://lucene.apache.org/solr/quickstart.html. Achtung: Das Tutorial bezieht sich auf eine nicht aktuelle Version und setzt eine *Unix*- bzw. *Mac OS*-Umgebung voraus; für die *Windows*-Laborumgebung sind daher folgende Anpassungen nötig:

- Installationsverzeichnis:
  `C:\solr-5.4.0` (alle folgenden Befehle beziehen sich auf dieses Verzeichnis)
- Startbefehl („Getting Started“):
  `bin\solr.cmd start -e cloud -noprompt`
- Indexierungsbefehl („Indexing a directory of ”rich” files“):
  ```
  java -classpath dist\solr-core-5.4.0.jar -Dauto=yes
  -Dc=gettingstarted -Ddata=files -Drecursive=yes
  org.apache.solr.util.SimplePostTool docs
  ```

- Indexierungsbefehl („Indexing Solr XML“):
  `java -classpath dist\solr-core-5.4.0.jar -Dauto=yes -Dc=gettingstarted -Ddata=files org.apache.solr.util.SimplePostTool example\exampledocs\*.xml`
- Indexierungsbefehl („Indexing JSON“):
  `java -classpath dist\solr-core-5.4.0.jar -Dauto=yes -Dc=gettingstarted -Ddata=files org.apache.solr.util.SimplePostTool example\exampledocs\books.json`
- Indexierungsbefehl („Indexing CSV (Comma/Column Separated Values)“):
  `java -classpath dist\solr-core-5.4.0.jar -Dauto=yes -Dc=gettingstarted -Ddata=files org.apache.solr.util.SimplePostTool example\exampledocs\books.csv`
- Löschbefehl („Deleting Data“):
  `java -classpath dist\solr-core-5.4.0.jar -Dauto=yes -Dc=gettingstarted -Ddata=args org.apache.solr.util.SimplePostTool "<delete><id>SP2514N</id></delete>"`
- Säuberungsbefehl („Cleanup“):
  `bin\solr.cmd stop -all`; `rd /s /q example\cloud`

Das Tutorial umfasst folgende Einzelschritte, die zur Übung alle durchlaufen werden sollen:

- Download
- Installation
- Systemstart
- Indexierung einer gemischten Testkollektion
- Indexierung von Testdaten im *Solr*-XML-Format
- Indexierung von Testdaten im JSON-Format
- Indexierung von Testdaten im CSV-Format
- Aktualisieren und Löschen von Daten
- Suchanfragen

2 Import bibliografischer Referenzdaten

In diesem Teil des Arbeitsprogramms wird eine eigene Kollektion bibliografischer Referenzdaten (100 englischsprachige Dokumente der Datenbank „Literatur zur Informationserschließung“) automatisch indexiert und in *Solr* importiert.

2.1 Vorbereitung der bibliografischen Referenzdaten

Laden Sie hier `https://ixtrieve.fh-koeln.de/lehre/solr-import/MatrNr.zip` die Archivdatei herunter (ersetzen Sie `MatrNr` in der URL durch Ihre Matrikel-Nr.).

Entpacken Sie das Archiv und öffnen Sie die darin enthaltene Datei `solr.dbm` mit dem bekannten und beliebten Tool *Midos*.

Erstellen Sie ein Ausgabeformat für eine *Lingo*-Indexierung der Kategorien „Titel“ und „Abstract“.

Führen Sie eine automatische Indexierung mit *Lingo*[1] mit folgenden Funktionsumfang durch: Indexierung der Kategorieninhalte von „Titel“ und „Abstract“ mit Grundformerzeugung und algorithmischer Mehrworterkennung (sequences). Verwenden Sie für den Indexierungslauf die in der Archivdatei enthaltene *solr*-Konfiguration (*solr.cfg* sowie die ebenfalls dort vorhandene sprachspezifische Konfiguration *en.lang*. Die *Lingo*-Indexierung lässt sich direkt im Datenverzeichnis starten mit dem Befehl:

- *Lingo*-Aufruf:
  `lingo -c solr.cfg -l en.lang export.txt`
  („export.txt“ durch den Namen der eigenen Export-Datei aus *Midos* ersetzen)

Importieren Sie die Indexierungsergebnisse in *Midos* über die Funktion „Daten mischen“; Grundformen sind in die neu anzulegende Kategorie *LEM*, *sequences* in die neu anzulegende Kategorie *SEQ* zu importieren.

Führen Sie einen Export der gesamten Dokumentkollektion durch. Bringen Sie die Daten dazu über ein selbst gestaltetes Ausgabeformat in das *Solr*-XML-Format gemäß folgender Vorgaben:

Tabelle 1: Feldzuordnung Solr – Midos

Solr-Schema	Midos-Feld
title_txt	TIT
author_txt	VER
abstract_txt	ABS
lemma_txt	LEM
sequence_txt	SEQ

Die Spezifikation des *Solr*-XML-Schemas findet sich hier: `https://wiki.apache.org/solr/UpdateXmlMessages`.

[1] Vgl. Kapitel 5 in: Winfried Gödert/Klaus Lepsky/Matthias Nagelschmidt: Informationserschließung und Automatisches Indexieren : ein Lehr- und Arbeitsbuch (X.media.press), Berlin [u.a.] 2012.

```
HEADER
<?xml version="1.0" encoding="UTF-8"?>
<add>

FOOTER
</add>

XMLBEGIN(02111)

<doc>
  <field name="id">*number*</field>
  <field name="title_txt">{TIT[@]}</field>
  <field name="author_txt">{VER[@]}</field>
  <field name="abstract_txt">{ABS[@]}</field>
  <field name="lemma_txt">{LEM[@]}</field>
  <field name="sequence_txt">{SEQ[@]}</field>
</doc>

XMLEND
```

2.2 Import der Daten

Importieren Sie die exportierten *Midos*-Daten in *Solr*. Orientieren Sie sich bei Ihrem Vorgehen an der Beschreibung im *Solr*-Tutorial unter „Indexing Solr XML“. Hinweis: Löschen Sie vorab die in den vorherigen Schritten importierten Testdaten.

2.3 Testrecherchen

Führen Sie Testrecherchen für die *Midos*-Kollektion und die *Solr*-Kollektion durch, bei denen auch Volltextsuche und kategorienspezifische Suche miteinander verglichen werden.

3 Übungsaufgabe

Führen Sie eine Suche nach allen in Abschnitt 2 importierten Dokumenten durch und geben Sie alle Felder der ersten zehn Dokumente im XML-Format aus. Speichern Sie das Ergebnis der Suche in einer Datei und senden Sie diese per E-Mail bis zum 28.02.2016, 19.00 Uhr, an klaus.lepsky@th-koeln.de. Die Aufgabe ist einzeln zu bearbeiten.

10 Punkte