Einführung in Solr
Wahlpflichtmodul und Praktikum SS 2018

Philipp Schaer – 2018-04-18 – philipp.schaer@th-koeln.de

1 Installation einer eigenen Solr-Instanz auf den Laborrechnern

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

1.1 Kurz-Anleitung:

- Download Solr unter https://lucene.apache.org/solr/mirrors-solr-latest-redir.html
- Entpacken unter `c:\solr`
- Kommandozeile aufrufen und Wechsel in dieses Verzeichnis: `cd c:\solr`
- Solr starten: `bin\solr.cmd start`
- Leere Kollektion anlegen: `bin\solr.cmd create -c test`
- Daten importieren:

  ```
  java -jar -Dc=test -Dauto example\exampledocs\post.jar
  example\exampledocs\*.xml
  ```

- Ergebnis ansehen unter http://localhost:8983/solr/#/test

1.2 Umfangreiche Anleitung:

Solr 7.3.0 benötigt eine Java-Laufzeitumgebung (JRE Version 1.8 oder höher), die auf den Laborrechnern bereits vorhanden ist. Auf Ihren eigenen Rechnern müssen Sie für eine entsprechende Installation selbst sorgen. Für die Installation und den ersten Test von Solr folgen Sie den Schritten dieses Tutorials: https://lucene.apache.org/solr/quickstart.html.

Achtung: Es gibt ein paar Stolpersteine, die oft für Probleme sorgen. Die üblichen Probleme habe ich einmal aufgelistet!

- Achten Sie auf die genaue **Schreibweise von Befehlen auf der Kommandozeile**! Ein vergessenes Leerzeichen kann manchmal schon zu Problemen führen. **Vermeiden Sie Leerzeichen** in Datei oder Verzeichnisnamen! Diese müssen Sie ansonsten auf der Kommandozeile aufwendig ersetzen durch „\ “.
- **Installationsverzeichnis**: Entpacken Sie die Datei unter `c:\solr` (z.B. mit 7-Zip). Alternativ auch auf einen schnellen USB-3.0-Stick – Dann können Sie Solr auch zuhause ausführen. Laufwerk Z eignet sich auf Grund der Geschwindigkeit nicht für das Ausführen.
- Alle Befehle müssen Sie über die Eingabeaufforderung starten (Windows-Taste und dann `cmd` eingeben). Wechseln Sie in ihr Installationsverzeichnis, mittels `cd c:\solr` oder mit Hilfe des FreeCommanders (Verzeichnis auswählen, Extras, DOS-Fenster).

- Die meisten Aufrufe im Tutorial, die auf **cURL** basieren, können Sie auch einfach so, in Ihrem Browser aufrufen (also ohne den Präfix „curl“), da hier nur GET Befehle verwendet werden, die auch Ihr Browser umsetzen kann.

1.3 Umfang des Tutorials

Das Tutorial umfasst zahlreiche Einzelschritte, die zur Übung nicht alle durchlaufen werden sollen. Folgende Abschnitte können Sie auslassen:

- Faceting
- Exercise 3: Index Your Own Data
- Spatial Queries

Sie sollten etwas Zeit investieren und mit dem System „spielen“. Versuchen Sie die Syntax der jeweiligen Befehle zu verstehen und die einzelnen Komponenten zu ordnen. Sie sollen sich mit dieser Arbeitsumgebung vertraut machen und jederzeit in der Lage sein, diese wieder neu aufzusetzen, da dies im Laufe der Veranstaltung aus diversen Gründen notwendig sein kann (z.B. da Ihr USB-Stick defekt ist, Sie die Installation „kaputt konfiguriert“ haben und nicht mehr wissen, wie Sie sie lauffähig bekommen, etc.).

2 Indexierung der Baseline

Zunächst werden wir die komplette GIRT-Kollektion als Volltext indexieren, d.h. alle Felder werden als „Text“ verarbeitet, d.h. eine Standard-Indexierung wird durchgeführt. Hierbei wird nicht auf besondere Merkmale der Sprache Deutsch eingegangen, Stopworte werden nicht entfernt, Lemmatisierungen finden nicht statt, etc. Diese Volltextindexierung soll unsere „out-of-the-box“-Baseline sein. Also eine einfache Messlatte gegen die wir später andere Konfigurationen antreten lassen wollen.

- Laden Sie folgende Datei herunter und entpacken Sie den Inhalt nach `c:\solr\data`
  https://ixtrieve.fh-koeln.de/lehre/mp-wp-biw-17s/girt-solr.zip
- Erzeugen Sie eine neue Kollektion mit Namen girt_base:

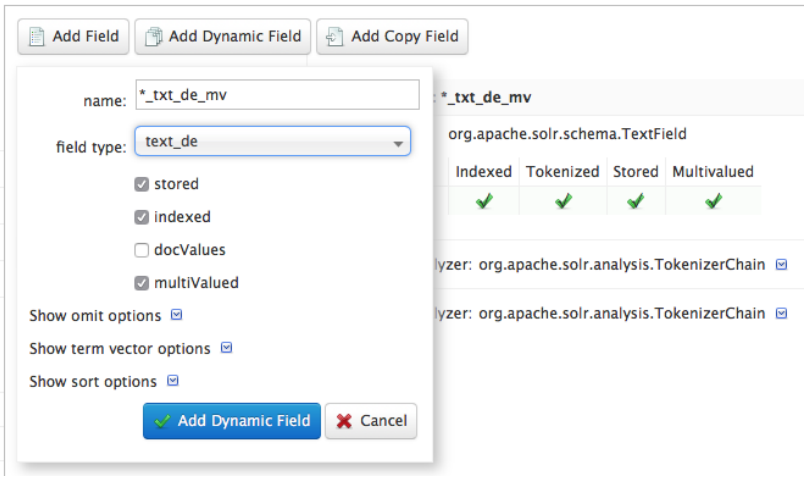
  ```
  bin\solr create -c girt_base
  ```
- Importieren Sie die Baseline mit folgendem Befehl

  ```
  java -jar -Dc=girt_base -Dauto example\exampledocs\post.jar
  data\*.xml
  ```
- Ergebnis ansehen unter http://localhost:8983/solr/#/girt_base

3 Aufgabe bis zum 9. Mai 2018

Importieren Sie die GIRT-Kollektion zusätzlich in zwei alternativen Varianten. Aktuell wird GIRT nur in einer sehr rudimentären Form indexiert. Es gibt keinerlei sprachliche Anpassungen bei der Indexierung. Da es sich aber um deutschen Text handelt, sollte dies auch entsprechend berücksichtigt werden.

- Fügen Sie ein neues Dynamic Field hinzu (analog zur Exercise 2 des Tutorials) mit dem Namen `*_txt_de_mv` hinzu und belegen Sie dieses mit dem Feldtyp `text_de` analog zu dem folgenden Screenshot. Achten Sie auch darauf, dass die Daten auf jeden Fall `multiValued` sein dürfen, die Felder also mehrfach in den Daten vorkommen dürfen.



- Ändern Sie in den GIRT-Dateien z.B. den Feldtyp für das Abstract von `abstract_txt` nach `abstract_txt_de_mv` mit Hilfe eines Texteditors und indexieren Sie die veränderte GIRT-Kollektion danach in eine neue Kollektion, z.B. mit dem Namen `girt_german`.
- Erstellen Sie zusätzlich zu der Grundinstallation `girt_base` insgesamt noch zwei alternative Varianten der GIRT-Kollektionen, indexieren Sie diese und führen Sie einen ersten Test durch, der die Unterschiede zwischen Ihren Varianten und der Baseline aufzeigt. Ändert sich z.B. die Zahl der Ergebnisse? Welche Suchen in welchen Feldern führen zu welchen Ergebnissen? Welche Änderungen haben keine/wenige/große Änderungen zur Folge? Usw. ...

Bei Fragen können Sie sich gerne im Moodle-Forum an mich wenden. Viel Erfolg.