Formalerschliessung von Zeitschriften (WS 20/21)
Aufgaben - 1

# 1 Einleitung

Das FID-Projekt (Fachinformationsdienst Linguistik) ist eine Kooperation von ACoLi mit dem Linguistik-Portal https://www.linguistik.de/ der Universititätsbibliothek Frankfurt. In einem Teilprojekt sollen Online-Publikationen, wie z.B. Artikel in Online-Zeitschriften für die Katalogsuche erschlossen werden. Dazu sollen automatisierte Verfahren entwickelt werden, die die Extraktion von Publikations-Metadaten wie Titel, Author, etc. aus Online-Zeitschriften durchführen. Die zu verarbeitenden Daten werden von der Bibliothek als PDF- bzw. Textdateien zur Verfügung gestellt.

**Datenbestand** Die Zeitschriftenartikel liegen in einer vorgegebenen Ordnerstruktur mit einem Wurzelverzeichnis und Unterordnern vor, die jeweils die Datensätze einer Online-Datenquelle (z.B. Zeitschrift) enthalten. Exemplarisch wird der Inhalt eines solchen Ordners in der Dokumentation im github Projekt (https://github.com/acoli-repo/SIA-cataloguing/documentation/input-data-format/items.json-doc.txt) beschrieben.

# 2 Aufgabe - Installation benötigter Software

Die Poppler-Bibiliothek wird für die XML-Konvertierung der Inputdaten, die im PDF-Format vorliegen benötigt.

**Linux (Ubuntu)** Die Installation für Linux ist ausführlich unter https://wiki.ubuntuusers.de/poppler-utils/ beschrieben

**Windows** Unter Windows sieht die Situation etwas schwieriger aus, da Poppler nur für Linux bereitgestellt wird. Ein Lösungsansatz ist die Installation von Windows-Subsystem-for-Linux (WSL). Eine Anleitung für die Installation von WSL für Windows 10 findet man z.B. hier https://towardsdatascience.com/poppler-on-windows-179af0e50150.
Ein zusätzliches Problem ist, dass von Java aus auf die Linux-Kommandozeile zugegriffen werden muss, was nicht einfach möglich ist. Das Tool pty4j kann dabei hilfreich sein https://github.com/JetBrains/pty4j.

# 3 Aufgabe - Umstellen auf neues JSON Eingabeformat

Ausgehend von dem Unit-Test **/testSourceHandlers/ testPDFExtractionWithSourceHandlerWithExtractionConfig.java** soll das Programm so erweitert werden, dass die Inputdaten jetzt mit Hilfe der Datei **items.jsonl** (siehe /documentation/input-data-format/items.json-doc.txt) eingelesen werden, die die Pfade der zu verarbeitenden Dokumente für eine Zeitschrift enthält. Die bisherige Konfiguration des Parsing Prozesses (in **/configs.json**) soll ebenfalls für die neue Eingaberoutine abgeändert werden.

**Änderungen in /configs.json**

1. URL Attribut streichen, da nicht mehr verwendet
2. Neues Attribut ID soll den Ordnernamen (PPN) speichern, damit PDFs einer Parsing-Konfiguration zugeordnet werden können
3. Für das Wurzelverzeichnis der Inputdaten, in denen sich die jeweiligen Unterordner mit Artikeln einer Zeitschrift befinden soll das neue Attribut **DocumentRootDirDocumentRootDir** : **path-to-root-directory-with-subfolders** angelegt werden

**Änderungen in der Klasse MetadataFromPDF.java**

1. Der einfachste Weg, um die Eingabe von bisher URLs auf jetzt lokale Files umzustellen ist es den Parameter URL im Konstruktor der Klasse MetadataFromPDF beizubehalten, aber jetzt eine File URL zu übergeben (file:///path-to-file)
2. In MetadataFromPDF.downloadURL() muss dann lediglich für den Fall das das Protokol der source URL = file, ein File Objekt zurückgegeben werden.

**JSON-Parser für das Einlesen der Dateipfade der Inputdaten** Ein JSON-Parser soll die Daten in **items.jsonl** (jsonl = JSON lines Format) zeilenweise auslesen. Ein PDF-Dateipfad kann aus dem Wurzelverzeichnis, dem Ordnernamen und dem Wert des JSON-Attributs *pdf_name* ermittelt werden.

# 4 Aufgabe - Testen der neuen JSON Eingaberoutine

- Als Testdaten sollen die Daten in /documentation/samples/input-examples/https-www-phon-ucl-ac-uk/047006471 verwendet werden
- Die folgende Konfiguration soll für die dort enthaltenen PDFs verwendet werden. Für die Parameter im Feld **sources** soll ? jeweils mit dem manuell ermittelten Wert (z.B. für titleFont) aus dem XML-Code einer konvertierten PDF-Eingabedatei ersetzt werden. (Anmerkung : Parameter dürfen nicht leer sein)

```
Konfiguration der Extraktionsparameter
in /configs.json:
{
"PathToTempDir" : "?",
"DocumentRootDir" : "?",
"sources" : [
    {
  "id" : "047006471",
      "type" : "pdf",
      "split" : false,
      "extractorConfig" : {
        "authorHeight" : ?,
        "authorFont" : ?,
        "titleHeight" : ?,
        "titleFont" : ?,
        "pageHeight" : ?,
        "pageFont" : ?
       }
     }
    ]
}
```