School of Engineering

Information Retrieval

Aufgabenstellung

- Demo-Applikation in Java mit Lucene
- Indexierung aller Schuldateien von 4 Jahren ZHAW
- Indexierung von PDF Inhalten
- Vergleich der Suchresultate von Lucene mit denjenigen des Betriebssystems (Windows)

Indexierung verschiedener Dateitypen

Die Lucene Library kann von Haus aus alle textbasierten Files indexieren. Darunter fallen Files wie .txt, .java, .c. Will man jedoch weitere Dateitypen unterstützen, müssen zusätzliche Libraries eingebunden werden. Anbei zwei kurze Beispiele.

PDF

- mehrere Libraries notwendig aus „org.apache.pdfbox“
- Erstellen eines neuen Dokumentes mit:
  ```
  LucenePDFDocument converter = new LucenePDFDocument();
  ```

Office Dokumente

- mehrere Libraries notwendig aus „org.poi“
- neues Dokument erstellen (hier XML-basiertes Word Dokument):
  ```
  XWPFWordExtractor wordxmlextractor = null;
  try {
  OPCPackage pkgDoc = POIXMLDocument.openPackage(file.toString());
  wordxmlextractor = new XWPFWordExtractor(pkgDoc); }
  ```

Vergleich Windows-Search-Engine mit Lucene