Übung Information Retrieval: xIndexaufbau - Erste Schritte mit Lucene

May 10, 2009

Prinzip der Indizierung

- Gegeben sind folgende Dokumente
  - Doc1: Neues aus der Medizin: Durchbruch in der Krebsbehandlung.
  - Doc2: Ein neues Medikament für die Behandlung von Krebs.
  - Doc3: Hoffnung für Krebspatienten. Vorstellung eines neuen Medikaments.
  - Doc4: Krebsbehandlung: BelüDrug stellt neue Krebstherapie vor.
- Konstruieren Sie (händisch) einen Index durch folgende Zwischenschritte: Tokenisierung, Normalisierung (Lemmatisierung), Sortieren und Gruppieren.

Prinzip der Indizierung

- Was liefern folgende booleschen Anfragen für die vier Dokumente in der vorherigen Aufgabe zurück:
  - neu AND Krebs
  - Krebsbehandlung
  - Krebs AND NOT Medikament
- Wieweit hängen die Ergebnisse von der Lemmatisierung ab?
- Wieweit hängen die Ergebnisse von der Tokenisierung ab?

Indizierung mit Lucene

- Indizieren Sie die Dokumente mit Lucene
- Hinweise:
  - Schreiben Sie den Code selber
  - Erzeugen Sie Strings für Dokument-Inhalte direkt im Code
  - Die zu importierenden Pakete, finden Sie in den Lucene Java-Docs
  - Beachten Sie das Exception-Handling

Suche mit Lucene

- Schreiben Sie eine Kommandozeilen Suche für den Index der vorherigen Übung
- Hinweise:
  - Die Suchanfrage soll als Komandozeilenparameter übergeben werden
  - Benutzen sie den gleichen Analyser wie bei der Indizierung