Übung: Indexaufbau, Boolean Search etc.

October 29, 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.

Boolean Suche

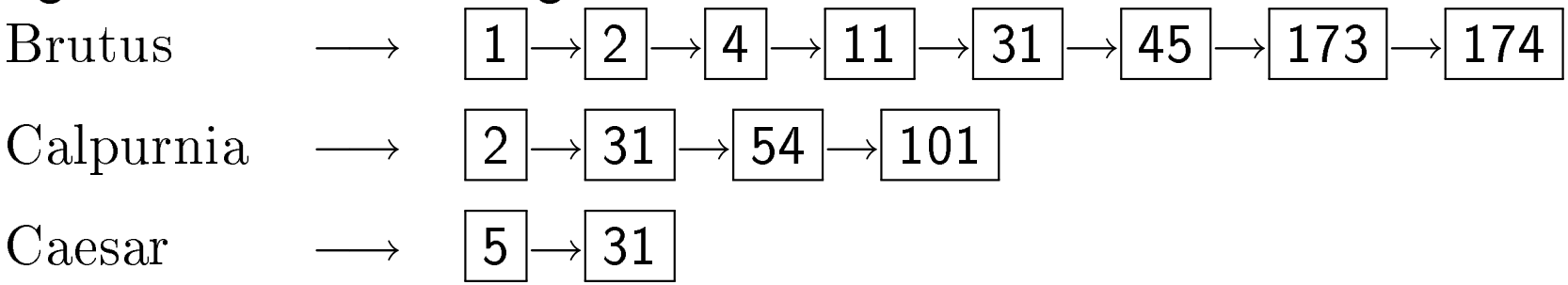
- Was liefern folgende booleschen Anfragen für die vier Dokumente in der vorherigen Aufgabe zurück:
  - neu AND Krebs
  - Krebsbehandlung
  - Krebs AND NOT Medikament
- Wieweit hängen die Ergebnisse von der Normalisierung der Token (linguistische, Bearbeitung wie Lemmatisierung etc.) ab?
- Wieweit hängen die Ergebnisse von der Tokenisierung ab?

Token vs. Terme

- Wie könnten extrahierte Informationen aus org.wikitionary.de helfen eine bessere Suche zu realisieren?
- Beispiel: Seite zu Haus

Reihenfolge bei AND-Merge

Gegeben sind die Posting Listen:



- Vergleichen die Anzahl der Postinglisteneinträge die angesehen werden, für beide mögliche Reihenfolgen der Auswerung:
  - (Brutus AND Calpurnia) AND Caesar
  - Brutus AND (Calpurnia AND Caesar)
- Hierbei müssen Sie auch die Einträge mitzählen, die in der "gemergten" Postingliste nachgeschlagen werden. Wieviele sind es in den beiden Fällen.

Algorithmus für Posting-List-Union (OR)

- Schreiben Sie einen Algorithmus für das Zusammenführen zweier Postinglisten zur Auswertung einer OR-Query - analog dem AND-Algorithmus der Vorlesung