Praktikum 1

Einfaches Boolesches IR-System

Implementieren Sie ein einfaches IR-System bestehend aus Tokenizer, Index (Dictionary, Posting-Listen) und Anfragebearbeitung gemäß dem in der Vorlesung vorgestellten Booleschen Retrieval-Modell. Das System sollte folgende Eigenschaften haben:

- Tokenizer:
  - Der Tokenizer liest Textdokumente vom Filesystem ein und liefert einen Strom von Token.
  - Der Tokenizer kann davon ausgehen, dass Token durch Leerzeichen, Zeilenumbrüche und die folgenden Zeichen getrennt werden: . , ; :!?“-
  - Einfache Apostrophe werden wie normale Buchstaben behandelt.
  - Wandeln Sie alle Token in Kleinbuchstaben um.
  - Jedes Token, das der Tokenizer ausgibt, wird als Indexterm behandelt. Sie brauchen sich also nicht um die Normalisierung von Token zu kümmern.
- Index/Dictionary
  - Der Index kann komplett im Hauptspeicher gehalten werden. Sie brauchen sich also nicht um die Auslagerung des Index (oder von Teilen des Index) auf den Sekundärspeicher zu kümmern.
  - Kapseln Sie die Einträge des Dictionaries und der Posting-Listen in eigenen Klassen, so dass Sie neben dem Term bzw. der DocID später auch noch zusätzliche Informationen hinzufügen können.
  - Das Vokabular, d.h. die Term-Liste, kann als sortierte Liste implementiert werden (nicht unbedingt als B-Baum).
  - Implementieren Sie die unterschiedlichen Varianten des Intersect-Algorithmus für den Merge zweier oder mehrerer Posting-Listen.
- Anfragebearbeitung
  - Gehen Sie davon aus, dass eine Anfrage in disjunktiver Normalform an Ihr System übergeben wird, d.h. als OR-Verknüpfung von AND-Verknüpfungen. Überlegen Sie sich eine möglichst einfach zu parsende Syntax.
- Benutzerschnittstelle
  - Wie Sie die Benutzerschnittstelle realisieren, bleibt Ihnen überlassen. Eine einfache Möglichkeit besteht darin, das System von der Kommandozeile zu starten und nach dem Aufbau des Index auf die Eingabe der Anfrage durch den Benutzer zu warten.
- Eine kleine Dokumentsammlung zum Testen finden Sie auf der INR-Seite
- Testen Sie folgende Anfragen:
  - Hexe
  - Hexe AND Prinzessin
  - (Hexe AND Prinzessin) OR (Frosch AND König AND Tellerlein)
  - (Hexe AND Prinzessin) OR (NOT Hexe AND König)
- Zeitplan
  - Die Aufgabe sollte bis spätestens Mi, 18.4. bearbeitet sein.