IR-Grundlagen - Worksheet 1

Dieses Worksheet behandelt den Inverted-Index sowie boolsches Retrieval und verschiedene merge-Algorithmen (AND, ANDNOT, OR, XOR). Ziel des Worksheets ist es am Ende einfache 2-Wort Queries durchführen zu können.

Zuerst beschäftigen wir uns mit der Indexierung. Dafür sind folgende Schritte notwendig:

1. Zuerst müssen die vorhandenen Dokumente indexiert werden. Sie müssen sich nicht darum kümmern die Dateien in Java einzulesen. Diese Arbeit wurde bereits übernommen. Die Main-Controller-Klasse besitzt eine ArrayList (`collection`) mit BooleanDocument-Objekten. Jedes Dokument der collection wird durch ein solches BooleanDokument-Objekt repräsentiert.
   Die InvertedIndex-Klasse bekommt `collection` vom Maincontroller übergeben. Sie besitzt außerdem folgende Variable:

   ```
   HashMap<String, ArrayList<Integer>> invertedIndex
   ```

   Diese Variable repräsentiert den Inverted-Index wobei der Key der Hashmap einen Term darstellt und der Value eine Liste der Dokumente in denen dieser Term auftritt.

   Nutzen sie die zur Verfügung stehenden `public` Methoden der BooleanDocument-Klasse um den Index zu befüllen. Erledigen sie diese Arbeit im Konstruktor der Inverted-Index-Klasse. legen sie bei bedarf weitere Methoden und Variablen an. Sie müssen bei dieser Aufgabe noch nicht auf linguistische Modifier achten da die Collection bereits präparierte Terme enthält.

2. Nun da jedes Dokument indexiert wurde benötigen sie eine Methode herauszufinden in welchen Dokumenten ein Term auftritt. Implementieren sie hierzu die Methode

   ```
   ArrayList<Integer> searchForSingleWord(String word)
   ```

   Diese Methode soll eine Liste von denjenigen Dokumenten-ID's (Integer-Werte) zurückgeben in denen sich der Term (`String word`) befindet.

Um Alle Schritte zu testen die wir bisher durchgeführt haben erstellen sie im MainController eine Methode die ein Wort vom User einliest und die Liste der Dokumente in denen dieses Wort vorkommt ausgibt. Denken sie über Case-folding nach.

Wir haben jetzt alle Vorraussetzungen für ein simples Retrievalsystem. Die nächsten beiden Aufgaben beschäftigen sich mit der Integration von boolschen Operatoren in die Suche

3. Als erstes müssen sie sich um das Einlesen der Query kümmern. Im MainController befinden sich die Variablen

   ```
   String firstTerm;
   String secondTerm;
   ```

   sowie die Methode

   ```
   getQueryTerms(firstTerms, secondTerms);
   ```

   Lesen sie in dieser Methode die Query vom User ein (Queries sollten die Form: “firstTerm AND secondTerm”), führen sie alle notwendigen Formattierungen durch (z.B.: Case-Folding) und befüllen sie die beiden Variablen.

4. Diese beiden Variablen werden an die Methode

   ```
   performANDMerge(String firstWord, String secondWord)
   ```

   in der InvertedIndex-Klasse übergeben. Implementieren sie diese so dass sie eine `public ArrayList<Integer>` Liste mit den Dokumenten-IDs der Dokumente die beide Terme beinhalten zurückgibt.

Zusätzliche Aufgaben:

5. Erweitern sie die InvertedIndex-Klasse um Methoden für folgende Operatoren

   - ANDNOT
   - OR
   - XOR

6. Wie müsste ein Vorgehen aussehen dass Suchanfragen mit mehreren (verschiedenen) Operatoren ermöglicht?