4. Übung zur Vorlesung "Internet-Suchmaschinen" im Wintersemester 2007/2008

Dr. Andreas Hotho, Prof. Dr. Gerd Stumme, M.Sc. Wi.-Inf. Beate Krause

13. Dezember 2007

1 Levenstein-Metrik

Berechnen Sie die Levenstein-Metrik *lev*("carsten", "christina").
Geben Sie das Berechnungsschema an, und nennen Sie die einzelnen Umformungsschritte einer kürzesten Umformung in der Art: test → tost → toast

2 Soundex

1. Vergleichen Sie die folgenden Wörter mit Soundex!
   - through, thru, trough
   - Mr, Mayer, Meier
   - Smith, Schmidt, Schmitz
   - data, date, dito
2. Sehen Sie Probleme und Verbesserungsmöglichkeiten?
3. Schlagen Sie vor, wie Sie die Probleme in Soundex beheben können.
4. Welche neuen Nachteile bringen Ihre Verbesserungen mit sich?

3 XML/XPath

1. Betrachten Sie das XML-Dokument auf Seite 27 des Kapitels "Strukturelle Anfragen" (`<library ... >`).

   Geben Sie XPath-Ausdrücke, die folgende Teile des Dokuments auswählen:

   a) alle Bücher

b) alle Bücher von William Smart

2. Funktionieren die folgenden Ausdrücke auf dem Dokument auf Seite 19 genau so?

   a) alle Personen

   b) alle Personen, die Robert heißen. *Tip:* Den Textinhalt von Kindelementen kann man prüfen, indem man einen relativen Pfad statt `@attribut` in die eckige Klammer schreibt!

4 Eigenschaften von Texten/Power Laws

1. Begründen Sie die Aussage von Luhn auf Seite 6 des Kapitels: Warum sind besonders häufige und besonders seltene Wörter nicht sehr nützlich?

2. Auch die Grade von Webseiten sind nach einem Potenzgesetz (*power law*) verteilt. Die folgende Tabelle gibt einige Ingrade einer Menge von Webseiten zu einem Zeitpunkt im Jahr 1999 wieder:

Grad	Anzahl Seiten
1	63100000
10	501200
100	4000
1000	32
5000	1

   Bestimmen Sie den Koeffizienten c im Potenzgesetz!

5 Praxisübung; Abgabe am 09.01.2008

Implementieren Sie die Phrasensuche wie auf Übungsblatt 3, Aufgabe 1, skizziert. Wenn nötig, erweitern Sie die entsprechenden Datenstrukturen.