Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

Institut für angewandte Informations-
technologie InIT

Information Engineering 1:

Praktikum 3: Indexing (ML)

Aufgabe 1: Indexieren

1. Welche Buchstabennormalisierung wird durchgeführt?

Gross-/Kleinschreibung, diakritische Zeichen umwandeln. Ligaturen werden nicht behandelt.

2. Wie verändert sich der Index, wenn Sie Stemming einschalten?

Die Terme werden auf "Wortstämme" reduziert. Die Gesamtzahl der Terme nimmt ab, da potentiell mehrere Terme auf denselben Stamm "fallen". Einzelne dieser gestemmten Terme haben nun eine höhere Anzahl Vorkommen.

3. Wie sieht es aus, wenn Sie statt der deutschen Stemming-Komponente die englische verwenden?

Es werden englische Endungen abgetrennt. Bei deutschen Dokumenten führt dies naturgemäss zu unbefriedigenden Ergebnissen (Stemming ist sprachabhängig!)

4. Wie wirkt sich Stoppwort-Elimination aus?

Es werden die allgemein in der deutschen Sprache häufigsten Wörter gestrichen (fixe Liste)

5. Erzeugen Sie eine möglichst kurze benutzerdefinierte Stoppwortliste, so dass weniger als 220 Terme übrigbleiben.

Es sind viele individuelle Lösungen möglich. Bauen Sie die Liste langsam aus, und überprüfen Sie immer wieder die verbleibende Anzahl Merkmale

Aufgabe 2: Indexierung unter der Lupe

1. Wie verändert sich die Anzahl der verschiedenen Terme, wenn man Stemming aktiviert? Haben Sie dafür eine Erklärung?

Die Anzahl der Terme sinkt. Potentiell fallen mehrere Wortformen auf denselben Stamm (dies ist ja der gewünschte Effekt)

2. Sie haben den Index mit Stemming und Stoppwortfilter erzeugt. Trotzdem erscheint das Stoppwort „dies“. Wie kann das sein?

Es handelt sich gar nicht um das Wort "dies", sondern um die Stammform "dies", die aus "dieses" erzeugt wurde.

3. Wie verändert sich die Anzahl der verschiedene Terme beim Hinzufügen von gleichartigen Dokumenten? Wie bei artfremden Dokumenten?.

Das Vokabular der "gleichartigen" Dokumente ist dem vorhandenen Dokumentenset ähnlicher. Die Anzahl verschiedener Terme/Merkmale im Index wächst langsamer als nach dem Hinzufügen der "artfremden" Dokumente

4. Im Index kommen nicht nur Wörter vor, sondern auch Zahlen. Warum ist es unter Umständen sinnvoll, Zahlen in den Index aufzunehmen?

Die Suche nach Zahlen ist oft sinnvoll oder nötig. Beispiele sind zahlreich: "Expo 02", "Boeing 767" usw. Aber Achtung: Zahlen sind nicht gleich verteilt wie Wörter (potentiell Auswirkung auf Gewichtung). Auch sind Zahlen in Isolation stark mehrdeutig (z.B. 36 – Hausnummer? Alter? Jahreszahl? Etc.)

5. Finden Sie alle Substantiv-Endungen (z.B. „-ung“) heraus, welche durch das Stemming abgetrennt werden. Wenn Sie nicht mehr weiter wissen, durchstöbern Sie die Dokumentenkollektion „Stemming“ und experimentieren damit.

Kann wie erwähnt durch Experimentieren ermittelt werden.

Aufgabe 3: Indexierung im grösseren Stil

Individuell. Keine Musterlösung.