Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

Institut für angewandte Informations-
technologie InIT

Information Engineering 1:

Praktikum 3: Indexing

Aufgabe 1: Indexieren

Bei der ersten Aufgabe geht es darum, mit IRLab die Indexierung zu untersuchen.
Sie benötigen dafür eine lauffähige Version von IRLab.

Vorgehen

- Starten Sie IRLab
- Wählen Sie die Dokumentenkollektion „Ozon“ als Source
- Klicken Sie auf den grünen Pfeil nach rechts, um den Index zu erzeugen
- Klicken Sie auf das Index-Symbol („blaues Buch“) und wählen Sie im Kontextmenü die Option „Show index“ um den Index anzeigen zu lassen

Fragen

1. Welche Buchstabennormalisierung wird durchgeführt?
2. Wie verändert sich der Index, wenn Sie Stemming einschalten?

    - Klicken Sie dazu auf den zweiten Slot im Kasten „Indexing Components“ und wählen Sie „German stemming“ aus

3. Wie sieht es aus, wenn Sie statt der deutschen Stemming-Komponente die englische verwenden?
4. Wie wirkt sich Stoppwort-Elimination aus?

    - Deaktivieren Sie zuerst die Stemming Komponente
    - Klicken Sie auf den ersten Slot im Kasten „Indexing Components“ und wählen Sie „German stopword filter“ aus

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

    - Wählen Sie den „Custom stopword filter“ aus
    - Passen Sie die Datei „<IRLab>/collections/stopwords_custom.txt“ entsprechend an und prüfen Sie ihre Resultate

Alles klein und keine äöü.
Werden zu aeo umgeschrieben.

Ohne Stemming 273
Terms, mit 218 Terms

Wenn die Englishe
verwendet wird, dann sind
es wieder 227 Terms.

Nur noch 182 Terms

Mit einer Liste (38) von den
meistverwendeten
Deutschen Wörtern aus
Wikipedia konnte ich den
Term auf 211 minimieren.

Aufgabe 2: Indexierung unter der Lupe

Die zweite Aufgabe nimmt die Indexierung etwas genauer unter die Lupe. Sie arbeiten wieder mit IRLab.

Vorgehen

- Sie arbeiten weiterhin mit der Dokumentenkollektion „Ozon“. Gehen Sie vor wie in Aufgabe 1.
- Erzeugen Sie den Index mit folgenden Parametern:
  - Stemming aus
  - „German stopword filter“ an

Fragen

Ohne Stemming hat man einen Termcount von 182. Wenn Stemming eingeschalten ist, beträgt dieser, neu, 171.
Da Wörter gekürzt angesehen werden (spieler und spiel sind gleich und spielen auch), führt dies zu einem kürzeren Term count.

1. Wie verändert sich die Anzahl der verschiedenen Terme, wenn man Stemming aktiviert? Haben Sie dafür eine Erklärung?
2. Sie haben den Index mit Stemming und Stoppwortfilter erzeugt. Trotzdem erscheint das Stoppwort „dies“. Wie kann das sein?

Es könnte sein, weil “dies” in den Varianten “dieser”, “dieses”, “diesen” etc. vorkommt und daher gestemmt und nicht mit “die” in Verbindung gebracht wird um zu Filtern.

3. Wie verändert sich die Anzahl der verschiedene Terme beim Hinzufügen von gleichartigen Dokumenten? Wie bei artfremden Dokumenten? Gehen Sie diese Frage wie folgt an:

Nun hat es einen Termcount von 191. Dies bedeutet neue Terme wurden gefunden welche noch nicht registriert/indexiert waren.

   - Kopieren Sie die Datei „gleichartig.html“ in den Ordner „<IR-Lab>/collections/unstructured/ozon/“ und erzeugen Sie den Index mit diesen Parametern: „German stemming“, „German stopword filter“. Notieren Sie ihre Beobachtungen.

Nun hat es einen Termcount von 203.

   - Entfernen Sie nun die Datei aus dem Ordner und kopieren Sie die Datei „artfremd.html“ hinein. Erzeugen Sie wiederum den Index und vergleichen Sie.

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

Für Jahreszahlen oder wichtige Daten wie Geschwindigkeit oder Messungen (wie zum Beispiel Weltrekorde).

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.

Ausdehnung, bestrahlung, empfehlung, furhung, lichtung

Aufgabe 3: Indexierung im grösseren Stil

Aufgabe 3 ist eine «offene» Aufgabe. Indexieren Sie eine substantielle Menge «eigener» Dokumente. Sie sollen dadurch ein Gespür für die in der Praxis auftretenden Probleme bei der Indexierung entwickeln.

IRLab bietet die Möglichkeit, eine «Custom»-Kollektion einzubinden. Sie finden die Kollektionen unter «<IRLab>/collections».

Um eine eigene Kollektion zu erstellen, ist wie folgt vorzugehen:

1. Teilen Sie ggf. die Dateien so auf, dass jedes Dokument in einer eigenen Datei zu liegen kommt.
2. Passen Sie das Textformat an. IRLab kann nur mit reinen Textdateien umgehen, Sie können nicht Worddateien, PDFs etc. direkt indexieren. Diese müssten Sie ggf. erst nach Text umwandeln. Jede Datei braucht einen minimalen Header:

```
<doc>
<recordId>1</recordId>
<text>
….
</text>
</doc>
```

Wobei das eigentlich Dokument zwischen <text> und </text> zu liegen kommt, und die Zahl zwischen <recordId> und </recordId> eine eindeutige Laufnummer ist – Sie können hier ggf. frei durchnummerieren, oder ggf. – falls in Ihren Dokumenten vorhanden – eine geeignete Id übernehmen.

3. Die so angepassten Dokumente kommen in den Ordner «<IRLab>/collections/custom/docs»
4. Listen Sie alle Dokumente in der Datei «doc_list_custom.txt» im Ordern «<IRLab>/collections/custom»

An allen anderen Dateien müssen Sie keine Änderungen vornehmen. Insbesondere bleibt der Ordner «<IRLab>/collections/custom/queries» leer, und die restlichen Dateien in «<IRLab/collections/custom» bleiben unverändert.

Sie können die Cranfield-Kollektion unter «<IRLab>/collections/cranfield» als Modell nutzen. Schauen Sie sich insbesondere die Datei «doc_list.txt» und den Ordner «docs» für Cranfield an.

Ein paar Ideen für Ihre Dokumentenkollektion:

- Indexieren Sie Ihre Emails (falls Ihr Mailclient einen geeigneten Export unterstützt)
- Indexieren Sie Ihre Worddokumente
- Bedienen Sie sich in einem freien Repository (Project Gutenberg, …)

Bitte beachten Sie: je weiter Ihre Ausgangsdokumente von reinem Text entfernt sind, desto mehr werden Sie mit dem «Cleaning» beschäftigt sein, was nicht der Kern der Aufgabe sein sollte.

Mit Ihrer neuen Kollektion machen Sie nun Indexierungsexperimente:

- Was lernen Sie über das Indexieren in der Praxis? Welcher «Noise» wird in Ihrem Index auftreten?
- Wie verhält sich der Index bei Stemming?
- Welche Beispiele von Understemming finden Sie?
- Welche Beispiele von Overstemming finden Sie?
- Welche weiteren Probleme beobachten Sie in der Praxis hinsichtlich Tokenisierung, Ein-/Ausschluss bestimmter Merkmale (Zahlen, Formatanweisungen, …)
- …