Aufgabe 1: Erzeugen und Anzeigen eines Indexes

Bei der ersten Aufgabe geht es darum, mit Soekia die Funktionsweise eines Indexes kennen zu lernen. Sie benötigen dazu eine lauffähige Version von Soekia und die Beispiel-Kollektion «Ozon».

Vorgehen

- Starten Sie Soekia.
- Öffnen Sie die Dokumenten-Kollektion «Ozon»
- Wählen Sie einen Speicherort für den Index. Sie können dazu das Verzeichnis der Dokumenten-Kollektion verwenden.
- Erzeugen Sie den Index.
- Lassen Sie sich den Index anzeigen.

Fragen

1. Welche Buchstaben-Normalisierungen führt Soekia durch?
2. Wie wirkt sich der Parameter «Sprache» auf den Index aus?
3. Wie wirkt sich die «Wortstamm-Reduktion» auf den Index aus?
4. Was bewirkt der Parameter «Sprache», wenn die «Wortstamm-Reduktion» aktiviert ist?
5. Was macht die «Stoppwort-Elimination»? (Schalten Sie vorher die «Wortstamm-Reduktion» wieder aus.)
6. Erzeugen Sie eine möglichst kurze benutzerdefinierte Stoppwort-Liste, so dass der Index weniger als 220 Terme umfasst.

Aufgabe 2: Index unter der Lupe

Die zweite Aufgabe nimmt den Index etwas genauer unter die Lupe. Sie arbeiten wieder mit Soekia. Zusätzlich benötigen Sie die Beispiel-Kollektion «Stemming» und die HTML-Dateien «gleichartig.html» und «artfremd.html».

Vorgehen

- Sie arbeiten weiterhin mit der Dokumenten-Kollektion «Ozon». Gehen Sie wie in Aufgabe 1 vor.
- Erzeugen Sie den Index mit folgenden Parametern: Sprache = Deutsch, Wortstamm-Reduktion = deaktiviert und Stoppwort-Elimination = fixe Liste.

Fragen

1. Wie verändert sich die Länge des Indexes, wenn man die Wortstamm-Reduktion aktiviert? Haben Sie dafür eine Erklärung?
2. Sie haben den Index mit aktivierter Wortstamm-Reduktion und Stoppwort-Elimination erzeugt. Trotzdem erscheint das Stoppwort «die» im Index. Wie kann das sein?
3. Wie verändert sich die Grösse des Indexes beim Hinzufügen von gleichartigen Dokumenten? Wie bei artfremden Dokumenten?

   Gehen Sie diese Frage wie folgt an:

   - Kopieren Sie die Datei «gleichartig.html» in den Ordner «Ozon» und erzeugen Sie den Index mit den Parametern: Sprache = Deutsch, Wortstamm-Reduktion = pseudo-linguistisch und Stoppwort-Elimination = fixe Liste. Notieren Sie ihre Beobachtungen.
   - 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?
5. Finden Sie alle Substantiv-Endungen (z. B. -ung) heraus, die Soekia abtrennt. Wenn Sie nicht mehr weiter wissen, nehmen Sie die Dokumenten-Kollektion «Stemming» zur Hilfe.

Aufgabe 3: Ausbeute und Präzision

Die dritte Aufgabe bringt Ihnen die Begriffe «Ausbeute» und «Präzision» näher. Sie arbeiten weiterhin mit der Dokumenten-Kollektion «Ozon».

Definition

$$\text{Ausbeute} = \frac{\text{Anzahl gefundener, relevanter Dokumente}}{\text{Anzahl relevanter Dokumente}}$$

$$\text{Präzision} = \frac{\text{Anzahl gefundener, relevanter Dokumente}}{\text{Anzahl gefundener Dokumente}}$$

Vorgehen

- Wählen Sie die Dokumenten-Kollektion «Ozon» und erzeugen Sie den Index mit folgenden Parametern: Sprache = Deutsch, Wortstamm-Reduktion = deaktiviert und Stoppwort-Elimination = fixe Liste.
- Eine Schülerin sucht für den Geografie-Unterricht Informationen zum Ozonloch. Sie benutzt dazu Soekia und die Dokumenten-Kollektion «Ozon».

Fragen

1. Durchsuchen Sie die Dokumenten-Kollektion nach potenziell relevanten Dokumenten. Welche Dokumente sind für die Schülerin relevant?
2. Die Schülerin stellt die Suchanfrage «Ozonloch». Welche Dokumente erhält sie? Wie gross ist die Ausbeute? Wie gross ist die Präzision? (Wenn Sie bei der vorherigen Frage keine Lösung gefunden haben, nehmen Sie an, dass vier Dokumente der Kollektion relevant seien.)
3. Die Schülerin hat erfahren, dass Wortstamm-Reduktion nützlich sei. Deshalb aktiviert sie das pseudo-linguistische Stemming und stellt nochmals dieselbe Suchanfrage. Wie gross ist nun die Ausbeute? Wie gross die Präzision?
4. Jemand hat der Schülerin gesagt, dass man auch andere Schreibweisen ausprobieren soll. Sie gibt deshalb die Suchanfrage «Ozonloch Ozon Loch» ein. Wie gross ist nun die Ausbeute? Wie gross die Präzision?
5. Angenommen Sie kopieren alle HTML-Dateien aus dem Ordner «Stemming» in den Ordner «Ozon» und erstellen den Index neu. Wie verändern sich die Ausbeute- und Präzisionswerte der vorherigen Fragen? Begründen Sie kurz ihre Antwort.
6. Sie kennen nun die Dokumenten-Kollektion gut. Zu welcher Suchanfrage raten Sie der Schülerin, um eine möglichst hohe Ausbeute und Präzision zu erreichen?

Aufgabe 4: Relevanz-Rangierung

Die vierte Aufgabe beschäftigt sich mit der Rangierung der gefundenen Dokumente gemäss ihrer Relevanz.

Rangierungsprinzipien von Soekia

1. Je mehr Suchbegriffe in einem Dokument vorkommen, desto relevanter ist das Dokument.
2. Je häufiger ein Suchbegriff in einem Dokument vorkommt, desto relevanter ist das Dokument.
3. Dokumente, die seltene Suchbegriffe enthalten, sind relevanter als Dokumente, die häufige Suchbegriffe enthalten.

Fragen

1. Entscheiden Sie für die folgenden drei Anfragen, welches der beiden Dokumente das relevantere ist. Welches Rangierungsprinzip gibt den Ausschlag?

Anfrage

Ozonloch Antarktis

Dokument 1	**Dokument 2**
Treibhauseffekt, Ozonloch, Überfischung der Weltmeere, Pinguinsterben in der Antarktis – die Menschheit zerstört ihren Planeten fortlaufend.	Im Winter entsteht über den Polkappen jeweils ein Ozonloch. Das Ozonloch über der Antarktis ist seit 1985 bekannt.

Anfrage

Ausdehnung Ozonloch Antarktis

Dokument 1	**Dokument 2**
Das Ozonloch über der Antarktis erreichte im September 2001 eine Ausdehnung von 29 Millionen Quadratkilometern.	Die Ausdehnung des Packeises über der Antarktis beträgt im Winter über 20 Millionen Quadratkilometer.

Anfrage

Geschichte Ozonloch Antarktis

Dokument 1	**Dokument 2**
Geschichte der Antarktis Am 16. Januar 1820 erreichten zum ersten Mal Menschen die Antarktis. Leiter der russischen Expedition war der baltische Deutsche Fabian Bellinghausen.	Ozonloch über der Antarktis Seit 1978 wird die Ozonschicht über der Südhalbkugel regelmässig gemessen. Anzeichen für ein Ozonloch wurden aber erst 1985 ernst genommen

Zur Beantwortung der folgenden Fragen benötigen Sie Soekia und die Beispiel-Kollektion «Ranking». Starten Sie Soekia und erstellen Sie für die Kollektion «Ranking» den Index mit folgenden Parametern: Sprache = Deutsch, Wortstamm-Reduktion = deaktiviert und Stoppwort-Elimination = deaktiviert. Die Rangierungprinzipien müssen alle auf «normal» eingestellt sein.

2. Stellen Sie die Suchanfrage «Computer Sicherheit». Begründen Sie die Rangierung der gefundenen Dokumente. Erklären Sie konkret, warum Dokument A vor B, A vor C und C vor B erscheint.
3. Stellen Sie die Suchanfrage «Internet Mail Spam». Warum ist das Dokument B weiter oben als Dokument D?
4. Welches Rangierungsprinzip muss auf «irrelevant» gesetzt werden, damit Dokument D als erstes erscheint?
5. Welches Rangierungsprinzip muss zusätzlich auf «irrelevant» gesetzt werden, damit Dokument A nicht mehr zu unterst auf der Rangliste steht?

Wettbewerb: Stemming (Wortstamm-Reduktion)

Beim Stemming-Wettbewerb ist Ihre Aufgabe, möglichst viele Wörter zu finden, die auf genau einen Stamm reduziert werden. Zur Inspiration können Sie sich die Beispiel-Kollektion «Stemming» anschauen.

Regeln

1. Es dürfen nur Wörter, die im Deutschen tatsächlich existieren, verwendet werden. Die Lehrperson kontrolliert mittels Wörterbuch (z. B. DUDEN Rechtschreibung) die Einhaltung dieser Regel.
2. Abkürzungen, die ohne Punkte geschrieben werden, dürfen nicht verwendet werden (z. B. ADAC, DVD, UNO).
3. Wörter, die sich nur in Gross-/Klein-Schreibung unterscheiden, zählen nur ein Mal (z. B. gehen, das Gehen).

Vorgehen

- Erstellen Sie einen neuen Ordner.
- Öffnen Sie ein Programm zum Erstellen von HTML-Seiten (z. B. Microsoft Word).
- Schreiben Sie Ihre Wörter in das Dokument und speichern Sie das Dokument im HTML-Format in den vorher erstellten Ordner.
- Starten Sie Soekia.
- Wählen Sie den erstellten Ordner als Dokumenten-Kollektion.
- Wählen Sie denselben Ordner als Index-Speicherort.
- Stellen Sie die Sprache auf «Deutsch» ein und aktivieren Sie das pseudo-linguistische Stemming (Wortstamm-Reduktion).
- Erstellen Sie den Index und kontrollieren Sie, ob wirklich nur ein Term enthalten ist.
- Ändern Sie das Dokument so lange, bis alle Wörter auf einen Term abgebildet werden und Sie keine weiteren Wörter mehr finden.

Wettbewerb: Web-Spamming

Szenario

Eine Autovermietungsfirma bietet im Engadin an verschiedenen Standorten Autos zur Vermietung an. Nachdem die Buchungen im letzten Jahr zurückgegangen sind, entschliesst sich die Firma, ihre Dienstleistungen auch über das Internet anzubieten. Ihre Aufgabe ist nun, die Homepage so zu gestalten, dass möglichst viele Engadin-Besucher/innen das Angebot finden. Konkret soll Ihre Homepage für typische Anfragen im Bereich «Autovermietung» möglichst weit oben in der Trefferliste erscheinen.

Regeln

1. Erlaubt ist, was gefällt. Gestalten Sie die HTML-Seite nach Belieben. Beachten Sie aber, dass sich ein/e Benutzer/in auch auf der Seite zurecht finden sollte.
2. Die Index-Parameter von Soekia sind wie folgt eingestellt: Sprache = Deutsch, Wortstamm-Reduktion = pseudo-linguistisch und Stoppwort-Elimination = deaktiviert.
3. Die Lehrperson hat eine Liste mit drei Suchanfragen. Alle Rangierungsprinzipien sind auf «normal» eingestellt. Gewonnen hat, wer bei den drei Suchanfragen die beste durchschnittliche Platzierung erreicht.

Vorgehen

- Öffnen Sie ein Programm zum Erstellen von HTML-Seiten (z. B. Microsoft Word).
- Erstellen Sie das Dokument.
- Speichern Sie das Dokument im HTML-Format.
- Geben Sie die Datei der Lehrperson, welche den Wettbewerb durchführt.

Um vorher zu testen, wie gut Ihre Homepage ist, können Sie Ihr HTML-Dokument in den Ordner «Autovermietung» kopieren und mit Soekia den Index für die Dokumenten-Kollektion erzeugen. Danach können Sie mit eigenen Suchanfragen überprüfen, ob Ihre Seite besser ist, als die vorgegebenen.

Anhang für die Lehrperson: Lösungen

Lösung zu Aufgabe 1: Erzeugen und Anzeigen eines Indexes

1. *Welche Buchstaben-Normalisierungen führt Soekia durch?*

   Soekia wandelt alle Grossbuchstaben in Kleinbuchstaben um. Zusätzlich werden ä, ö und ü durch a, o und u ersetzt, sowie ß durch ss.

2. *Wie wirkt sich der Parameter «Sprache» auf den Index aus?*

   Wenn die Sprache auf «English» eingestellt ist, wird die Umwandlung der Umlaute nicht durchgeführt.

3. *Wie wirkt sich die «Wortstamm-Reduktion» auf den Index aus?*

   Die Wortstamm-Reduktion schneidet von Wörtern die Endungen ab, so dass verwandte Wörter auf denselben Stamm abgebildet werden. Dieses Verfahren wird auch «Stemming» genannt. Soekia trennt 40 häufige Endungen ab, darunter Substantivendungen wie -heit, -keit und -ung, Adjektivendungen wie -bar, -er, -sten und -lich sowie Verbalendungen -end, -et, -st. Zusätzlich werden die Vorsilben ge-, ver- und un- abgetrennt.

4. *Was bewirkt der Parameter «Sprache», wenn die «Wortstamm-Reduktion» aktiviert ist?*

   Ist die Sprache «English» gewählt, werden englische Endungen abgetrennt. Bei deutschen Dokumenten zeigt das englische Stemming naturgemäss schlechte Ergebnisse.

5. *Was macht die «Stoppwort-Elimination»? (Schalten Sie vorher die «Wortstamm-Reduktion» wieder aus.)*

   Ist für die Stoppwort-Elimination die fixe Liste eingestellt, streicht Soekia die 50 häufigsten Wörter der gewählten Sprache aus dem Index. Wählt man eine benutzerdefinierte Liste, kann man angeben, welche Wörter nicht aufgenommen werden sollen.

6. *Erzeugen Sie eine möglichst kurze benutzerdefinierte Stoppwort-Liste, so dass der Index weniger als 200 Terme umfasst.*

   Ohne Stoppwort-Elimination enthält der Index 234 Terme. Das bedeutet, dass die benutzerdefinierte Stoppwort-Liste mindestens 15 Wörter umfassen muss. Es sind viele Lösungen möglich. Eine mögliche Liste wäre: «das, den, der, des, die, ein, es, im, in, ist, uber, und, von, vor, zu»

   Weil die Stoppwort-Elimination nach der Buchstaben-Normalisierung ausgeführt wird, müssen die Stoppwörter klein und ohne Umlaute geschrieben werden.

Lösung zu Aufgabe 2: Index unter der Lupe

1. *Wie verändert sich die Länge des Indexes, wenn man die Wortstamm-Reduktion aktiviert? Haben Sie dafür eine Erklärung?*

   Die Anzahl Terme sinkt von 198 auf 181. Nach der Wortstamm-Reduktion (Stemming) werden mehrere Wörter auf einen Term abgebildet. Vorher waren «Ozonloch» und «Ozonloches» eigene Terme. Mit Stemming werden beide auf «Ozonloch» abgebildet.

2. *Sie haben den Index mit aktivierter Wortstamm-Reduktion und Stoppwort-Elimination erzeugt. Trotzdem erscheint das Stoppwort «die» im Index. Wie kann das sein?*

   Soekia führt zuerst die Stoppwort-Elimination durch. Dabei wird das Wort «die» entfernt. Danach werden die Wörter auf den Wortstamm reduziert. Das Wort «dieses» wird auf «die» reduziert. Deshalb taucht «die» im Index auf, obwohl es ein Stoppwort ist.

3. *Wie verändert sich die Grösse des Indexes beim Hinzufügen von gleichartigen Dokumenten? Wie bei artfremden Dokumenten?*

   Beide Dokumente enthalten 60 Wörter. Fügt man das Dokument «gleichartig.html» der Dokumenten-Kollektion hinzu, wächst der Index von 181 auf 201 Terme. Beim Dokument «artfremd.html» wächst der Index von 181 auf 214 Terme. Das Dokument «gleichartig.html» enthält viele Terme, die bereits im Index sind. Deshalb wächst er nicht so stark wie beim Dokument «artfremd.html».

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

   Wenn Benutzer/innen von Suchmaschinen nach Zahlen suchen wollen, müssen diese auch im Index vorkommen. Typische Beispiele sind die Suche nach Jahreszahlen «Ausdehnung Ozonloch 2001», nach Büchern «Geschichten aus 1001 Nacht» oder nach technischen Begriffen «3,5 Zoll Diskette».

5. *Finden Sie alle Substantiv-Endungen (z. B. -ung) heraus, die Soekia abtrennt. Wenn Sie nicht mehr weiter wissen, nehmen Sie die Dokumenten-Kollektion «Stemming» zur Hilfe.*

   Folgende Substantiv-Endungen werden von Soekia abgetrennt:

-heit	Vergangen-heit	-lung	Hand-lung
-in	Sekretär-in	-lungen	Hand-lungen
-innen	Sekretär-innen	-s	Internet-s
-ion	Institut-ion	-ung	Versicher-ung
-ionen	Institut-ionen	-ungen	Versicher-ungen
-keit	Gemütlich-keit		

   Von den Substantiven werden auch Verbal-Endungen wie -end, -est und -et abgeschnitten, diese waren aber nicht gefragt. An der fehlenden Wortart-Analyse erkennt man die Grenzen des pseudo-linguistischen Stemming.

Lösung zu Aufgabe 3: Ausbeute und Präzision

1. *Durchsuchen Sie die Dokumenten-Kollektion nach potenziell relevanten Dokumenten. Welche Dokumente sind für die Schülerin relevant?*

   Vier Dokumente sind für die Fragestellung der Schülerin relevant: «Ausdehnung.html», «Ozon.html», «Ozonloch.html» und «Ozonschicht.html».

2. *Die Schülerin stellt die Suchanfrage «Ozonloch». Welche Dokumente erhält sie? Wie gross ist die Ausbeute? Wie gross ist die Präzision?*

   Die Anfrage liefert «Ozonloch.html» als einziges Suchresultat. Die Ausbeute ist 1/4 = 25%, die Präzision ist 1/1 = 100%.

3. *Die Schülerin hat erfahren, dass Wortstamm-Reduktion nützlich sei. Deshalb aktiviert sie das pseudo-linguistische Stemming und stellt nochmals dieselbe Suchanfrage. Wie gross ist nun die Ausbeute? Wie gross die Präzision?*

   Die Anfrage liefert die Dokumente «Ozonloch.html», «Ausdehnung.html» und «Antarktis.html». Relevant sind die ersten beiden. Die Ausbeute ist 2/4 = 50%, die Präzision ist 2/3 = 66,67%.

4. *Jemand hat der Schülerin gesagt, dass man auch andere Schreibweisen ausprobieren soll. Sie gibt deshalb die Suchanfrage «Ozonloch Ozon Loch» ein. Wie gross ist nun die Ausbeute? Wie gross die Präzision?*

   Die Anfrage liefert alle zehn Dokumente zurück. Die Ausbeute ist 4/4 = 100%, die Präzision ist 4/10 = 40%.

5. *Angenommen Sie kopieren alle HTML-Dateien aus dem Ordner «Stemming» in den Ordner «Ozon» und erstellen den Index neu. Wie verändern sich die Ausbeute- und Präzisionswerte der vorherigen Fragen? Begründen Sie kurz ihre Antwort.*

   In den Definitionen der Ausbeute und der Präzision kommt die Anzahl Dokumente der Kollektion gar nicht vor. Es zählen nur die relevanten und die gefundenen Dokumente. Die Ausbeute- und Präzisionswerte bleiben also gleich.

6. *Sie kennen nun die Dokumenten-Kollektion gut. Zu welcher Suchanfrage raten Sie der Schülerin, um eine möglichst hohe Ausbeute und Präzision zu erreichen?*

   Bei dieser Aufgabe sind mehrere Lösungen möglich. Wir entschieden uns für die Anfrage «Ausdehnung Ozonschicht», welche die Dokumente «Ozonschicht.html», «Ausdehnung.html», «Ozon.html» und «Ozonloch.html» liefert. Alle vier sind relevant. Die Ausbeute ist 4/4 = 100%, die Präzision ist 4/4 = 100%

   Die Anfrage «Ozonloch Ozonschicht» liefert ein ähnlich gutes Resultat: «Ozonschicht.html», «Ozonloch.html», «Ausdehnung.html», «Ozon.html» und «Antarktis.html». Vier davon sind relevant. Die Ausbeute ist 4/4 = 100%, die Präzision ist 4/5 = 80%.

Lösung zu Aufgabe 4: Relevanz-Rangierung

1. *Entscheiden Sie für die folgenden drei Anfragen, welches der beiden Dokumente das relevantere ist. Welches Rangierungsprinzip gibt den Ausschlag?*

Anfrage

Ozonloch Antarktis

Dokument 1	**Dokument 2**
Treibhauseffekt, Ozonloch, Überfischung der Weltmeere, Pinguinsterben in der Antarktis – die Menschheit zerstört ihren Planeten fortlaufend.	Im Winter entsteht über den Polkappen jeweils ein Ozonloch. Das Ozonloch über der Antarktis ist seit 1985 bekannt.

Für die erste Anfrage ist das Dokument 2 relevanter. Ausschlaggebend ist das Rangierungsprinzip 2.

Anfrage

Ausdehnung Ozonloch Antarktis

Dokument 1	**Dokument 2**
Das Ozonloch über der Antarktis erreichte im September 2001 eine Ausdehnung von 29 Millionen Quadratkilometern.	Die Ausdehnung des Packeises über der Antarktis beträgt im Winter über 20 Millionen Quadratkilometer.

Für die zweite Anfrage ist das Dokument 1 relevanter. Ausschlaggebend ist das Rangierungsprinzip 1.

Anfrage

Geschichte Ozonloch Antarktis

Dokument 1	**Dokument 2**
Geschichte der Antarktis Am 16. Januar 1820 erreichten zum ersten Mal Menschen die Antarktis. Leiter der russischen Expedition war der baltische Deutsche Fabian Bellinghausen.	Ozonloch über der Antarktis Seit 1978 wird die Ozonschicht über der Südhalbkugel regelmässig gemessen. Anzeichen für ein Ozonloch wurden aber erst 1985 ernst genommen

Für die dritte Anfrage ist das Dokument 2 relevanter. Ausschlaggebend ist das Rangierungsprinzip 3. Der Begriff «Ozonloch» ist spezifischer als der Begriff «Geschichte». «Ozonloch» gehört eindeutig in den Themenbereich Geografie, Ökologie und Politik. «Geschichte» kann in «Geschichte der Mathematik», «Geschichte der Raumfahrt», «Die unendliche Geschichte» und vielen anderen Dokumenten vorkommen.

2. *Stellen Sie die Suchanfrage «Computer Sicherheit». Begründen Sie die Rangierung der gefundenen Dokumente. Erklären Sie konkret, warum Dokument A vor B, A vor C und C vor B erscheint.*

   A vor B: Dokument A enthält mehr Suchbegriffe als B (Rangierungsprinzip 1) und der Suchbegriff «Computer» kommt häufiger vor (Rangierungsprinzip 2).

   A vor C: Der Suchbegriff «Computer» kommt in Dokument A häufiger vor als in C (Rangierungsprinzip 2).

   C vor B: Dokument C enthält mehr Suchbegriffe als B (Rangierungsprinzip 1).

3. *Stellen Sie die Suchanfrage «Internet Mail Spam». Warum ist das Dokument B weiter oben als Dokument D?*

   Zwar enthalten beide Dokumente gleich viele Suchbegriffe und Dokument D den Begriff «Internet» sogar häufiger als Dokument B, aber der Begriff «Spam» in B ist viel spezifischer als Mail. «Spam» kommt in der ganzen Kollektion nur zwei Mal vor, «Mail» hingegen sechs Mal. Rangierungprinzip 3 gibt also den Ausschlag.

4. *Welches Rangierungsprinzip muss auf «irrelevant» gesetzt werden, damit Dokument D als erstes erscheint?*

   Rangierungsprinzip 3 muss auf «irrelevant» gesetzt werden.

5. *Welches Rangierungsprinzip muss zusätzlich auf «irrelevant» gesetzt werden, damit Dokument A nicht mehr zu unterst auf der Rangliste steht?*

   Dokument A enthält nur einen der Suchbegriffe, diesen dafür vier Mal. Deshalb muss das erste Rangierungsprinzip auf «irrelevant» gesetzt werden.

Lösung zu Wettbewerb: Stemming (Wortstamm-Reduktion)

Beim Stemming-Wettbewerb ist Ihre Aufgabe, möglichst viele Wörter zu finden, die auf genau einen Stamm reduziert werden. Zur Inspiration können Sie sich die Beispiel-Kollektion «Stemming» anschauen.

Als Lehrperson überwachen Sie die Einhaltung der in der Aufgabenstellung definierten Regeln. Dazu benötigen Sie ein Wörterbuch, z. B. DUDEN Rechtschreibung. Um zu einem gegebenen Wort alle Beugeformen (Flexion) zu finden, kann zum Beispiel der Internet-Dienst http://www.canoo.net/ benutzt werden.

Die beste Lösung, die wir bisher gefunden haben, ist eine Liste von 84 Wörtern, die alle auf den Stamm «bild» reduziert werden. Weitere viel versprechende Wortstämme sind «ein» (eine, einst, einig, Verein, …), «fall» (fallen, gefallen, Verfall, …), «lang» (lange, gelangen, Verlangen, …) und «sich» (gesichert, unsicher, Versicherung, …). Wenn Sie eine längere Liste erhalten, würden wir uns freuen, wenn Sie uns diese zukommen lassen. Die Kontakt-Adresse von Swisseduc finden Sie auf http://www.swisseduc.ch.

Bild	bildende	bildlichstem	verbilde
Bilde	bildendem	bildlichsten	verbildend
Bilder	bildenden	bildlichster	verbildende
Bildern	bildendes	bildlichstes	verbildendem
Bildes	bildest	gebildet	verbildenden
Bilds	bildet	gebildete	verbildendes
Bildung	bildete	gebildetem	verbildest
Bildungen	bildeten	gebildeten	verbildet
Gebilde	bildetest	gebildeter	verbildete
Gebilden	bildlich	gebildetere	verbildeten
Gebildes	bildliche	gebildeterem	verbildetest
Unbilden	bildlichem	gebildeteren	verbildetet
Unbildung	bildlichen	gebildeterer	verbildliche
Verbilden	bildlicher	gebildeteres	verbildlichen
Verbildens	bildlichere	gebildetes	verbildlichend
Verbildlichung	bildlicherem	gebildetste	verbildlichende
Verbildlichungen	bildlicheren	gebildetstem	verbildlichendem
Verbildung	bildlicherer	gebildetsten	verbildlichenden
Verbildungen	bildlicheres	gebildetster	verbildlichendes
bilden	bildliches	gebildetstes	verbildlichest
bildend	bildlichste	verbild	verbildlichst

Lösung zu Wettbewerb: Web-Spamming

Szenario

Eine Autovermietungsfirma bietet im Engadin an verschiedenen Standorten Autos zur Vermietung an. Nachdem die Buchungen im letzten Jahr zurückgegangen sind, entschliesst sich die Firma, ihre Dienstleistungen auch über das Internet anzubieten. Ihre Aufgabe ist nun, die Homepage so zu gestalten, dass möglichst viele Engadin-Besucher/innen das Angebot finden. Konkret soll Ihre Homepage für typische Anfragen im Bereich «Autovermietung» möglichst weit oben in der Trefferliste erscheinen.

Vorgehen

- Kopieren Sie die HTML-Dateien der Schüler/innen in den Ordner «Autovermietung». Sie müssen darauf achten, dass alle einen unterschiedlichen Dateinamen tragen. Die Dateinamen haben keinen Einfluss auf die Rangliste.
- Starten Sie Soekia und erstellen Sie für den Ordner «Autovermietung» den Index mit folgenden Parametern: Sprache = Deutsch, Wortstamm-Reduktion = pseudo-linguistisch und Stoppwort-Elimination = deaktiviert.
- Stellen Sie die folgenden drei Anfragen:

  autovermietung graubünden

  Die Urheberin dieser Anfrage sucht einfach eine Autovermietung in Graubünden. Anfragen dieser Art sind sehr häufig. Der Suchbegriff «Graubünden» ist weniger spezifisch als «Engadin». Vielleicht weiss die Touristin gar nicht, dass das Tal, in welchem sie ihren Urlaub verbringt, Engadin heisst.

  luxuswagen mit chauffeur mieten in st. moritz

  Dieser Benutzer formuliert seine Anfrage «natürlich-sprachlich». Er möchte für die Dauer seines Aufenthaltes in St. Moritz einen Wagen der Luxusklasse mit Chauffeur mieten. Der Tourist kennt «St. Moritz» als Luxusferienort, weiss aber nicht, dass St. Moritz im Engadin liegt und das Engadin in Graubünden.

  car rental engadine switzerland

  Diese Benutzerin kommt aus den Vereinigten Staaten und hat über ein schweizerisches Tourismus-Portal vom «engadine» gehört. Die Aufgabenstellung fordert zwar nicht, dass man auch eine englische Seite gestalten muss. Wer trotzdem daran gedacht hat, erhält hier leicht ein paar zusätzliche Punkte.

- Die Auswertung geschieht wie folgt: Für den ersten Platz in der Rangierung erhält man 10 Punkte, für den zweiten 9 und so weiter. Ab dem elften Platz

gibt es keine Punkte mehr. (Wenn man bei «Google» nicht auf die erste Seite, d. h. unter die ersten zehn Resultate, kommt, wird man von den meisten Benutzern gar nicht beachtet.)
Die Punkte für alle drei Anfragen werden zusammengezählt. Gewonnen hat, wer am meisten Punkte hat. Es kann auch vorkommen, dass mehrere Schüler/innen dieselbe Punktzahl erreichen. In diesem Fall gibt es halt mehrere Sieger/innen.

Web-Spamming Tipps

- Das erste Rangierungsprinzip verlangt, dass man möglichst viele Suchbegriffe auf einer Webseite hat. Die Gestalterin der Homepage muss deshalb möglichst alle Begriffe, die irgendwie mit «Autovermietung» zusammenhängen, auf der Homepage erwähnen. Zu beachten sind Synonyme (Auto, Wagen, Fahrzeug), Getrennt- und Zusammenschreibungen (Autovermietung, Auto Vermietung), Verb- und Substantivformen (mieten, Vermietung) und eventuell fremdsprachliche Begriffe (car rental).
- Das zweite Rangierungsprinzip verlangt, dass man die Suchbegriffe möglichst oft auf der Webseite hat. Ein beliebter Trick ist deshalb, Begriffe hundertfach in weisser Farbe auf weissem Grund zu wiederholen. Bei Soekia funktioniert dieser Trick, bei «Google» werden solche Seiten aus dem Index verbannt.
- Das dritte Rangierungsprinzip verlangt, dass man spezifische Suchbegriffe auf der Webseite verwendet. Nebst «Graubünden» muss man also auch «Engadin» und «St. Moritz» auf der Webseite erwähnen.