F2 Der Index

Im Posten I2 hast du gelernt, dass jeder Informationsdienst einen Teil des Internets für die Suche zugänglich macht. Wird dem Dienst eine Suchanfrage gestellt, so schaut der Suchdienst in seinem Index nach, ob er ein (oder mehrere) dazu passendes Dokument verzeichnet hat.

An diesem Posten wollen wir uns ansehen, wie sich ein Suchdienst seinen Index aufbaut, d.h. wie er sich merkt, was im für ihn wichtigen Teilbereich des Internets für Web-Seiten vorhanden sind.

Zuerst muss der Suchdienst möglichst viele Seiten aufstöbern, die zum untersuchten Teilbereich im Internet gehören. Dies geschieht dadurch, dass der Informationsdienst kleine Roboter (eigentlich sind es kleine Programme) ins Internet aussendet. Die Roboter, auch Bots genannt, „sehen“ sich die Seiten im Internet an. Wie ein Surfer gehen sie von Link zu Link und somit von einer Seite zur anderen. Bei jeder Seite, die der Suchdienst in seinem Index haben möchte, wird wird das folgende Verfahren angewendet.

Beispiel: Wir betrachten die folgende (fiktive) Internet-Seite www.mondlandung.ch, die dem Index des Suchdienstes hinzugefügt werden soll:

> Am 20. Juli 1969 landeten die Amerikaner erstmals auf dem Mond. Neil Armstrong berührte den Mond als Erster und verkündete: „Ein kleiner Schritt für einen Menschen, aber ein Riesenschritt für die Menschheit.“

Zuerst wird festgestellt, um welche Sprache es sich handelt. Da die Seite in Deutsch ist, wird nun eine **Buchstabenumwandlung** gemacht. ä, ö, ü wird zu ae, oe, ue.

> Am 20. Juli 1969 landeten die Amerikaner erstmals auf dem Mond. Neil Armstrong beruehrte den Mond als Erster und verkuendete: „Ein kleiner Schritt fuer einen Menschen, aber ein Riesenschritt für die Menschheit.“

Anschliessend wird die **Wortextraktion** durchgeführt, d.h. es werden die Satzzeichen weggelassen.

> Am 20 Juli 1969 landeten die Amerikaner erstmals auf dem Mond Neil Armstrong beruehrte den Mond als Erster und verkuendete Ein kleiner Schritt fuer einen Menschen aber ein Riesenschritt für die Menschheit

Nun werden alle **Stoppwörter eliminiert**. Stoppwörter sind Begriffe, die nichts oder nur sehr wenig zur Beschreibung des Inhalts beitragen. Bsp. am, die, auf,...

> 20 Juli 1969 landeten Amerikaner erstmals Mond Neil Armstrong beruehrte Mond Erster verkuendete kleiner Schritt Menschen Riesenschritt Menschheit

In der nächsten Phase werden die **Wörter zerlegt und normalisiert**. So wird etwa aus „Riesenschritt“ „riesig“ und „Schritt“. Bei der Normalisierung wird lediglich der Wortstamm, ohne Endungen und verschiedene Schreibweisen, übernommen. Z.B. statt „beruehrte“ „beruehr“. Zudem werden alle Gross- in Kleinbuchstaben umgewandelt.

> 20 juli 1969 land amerika erst mal mond neil armstrong beruehr mond erst verkuend klein schritt mensch riesig schritt menschheit

Die so zurechtgemachte Seite wird nun analysiert und dem Index beigefügt. Dazu werden die einzelnen Begriffe im Dokument gezählt und in einer Tabelle aufgelistet.

Begriff	Häufigkeit	Positionen	...	...
mond	2	8, 12		
erst	2	6, 13		
schritt	2	16, 19		
1969	1	3		
...	...	...		

Der Suchdienst weiss nun, welche Begriffe sich auf der Seite www.mondlandung.ch befinden. Wird nun eine Suchanfrage wie beispielsweise „mondlandung erste“ gestellt, so erkennt der Suchdienst, nachdem er die Suchanfrage ebenfalls in „mond“ „landung“ „erst“ zerlegt hat, dass die Seite www.mondlandung.ch dieser Anfrage entspricht und liefert sie als Treffer zurück. Wo in der Rangliste die Seite dann erscheint, hängt von den Rangierungsprinzipien ab. Vgl Posten F1.

Das gezeigte Verfahren ist nur ein Muster für die Erstellung eines Indexes. Nicht alle Suchdienste wenden alle Teile an, oder sie verwenden noch andere Mechanismen. Im Experimentierteil weiter unten wirst du bei einigen Suchdiensten versuchen herauszufinden, ob der Index nach dem gezeigten Muster erstellt wird.

Überlege dir, wie du bei einem Suchdienst herausfinden kannst, ob er Wortzerlegung, Wortnormalisierung, Stoppworteliminierung und Umlauterkennung durchführt.

Untersuche dann bei den beiden Suchdienste www.google.de (12.12.02) und www.fireball.de (12.12.02) ob sie...

... Wortzerlegung durchführen. Suchanfrage „toskanareise“

... Wortnormalisierung durchführen. Suchanfrage „landen“

... Stoppwörter eliminieren. Suchanfrage „sein und haben“ resp. „sein haben“

... Umlaute/Accent erkennen. Verwende als Suchanfrage „nestlé“ resp. „nestle“

Das Schweizerische Parlament (www.parlament.ch 26.04.06) verwendet für seine Seiten einen Suchdienst mit dem Namen RotondoSpider. Untersuche diesen auf Wortzerlegung, Wortnormalisierung, Stoppworteliminierung und Umlauterkennung. Dokumentiere deine Suchanfragen und erkläre deine Interpretation.