

Übung „Netzbasierte Informationssysteme“ WS 2008/2009

Übungsblatt 6

Ausgabe am 16.12.2008

Abgabe bis spätestens 6.1.2009, 16.00 Uhr

Aufgabe Kurzbeschreibung

Für die zweite Aufgabe werden Sie eine Websuche in Ihrer Website z.B. aus Übungsblatt 2 , Aufgabe 7 oder Webseite ihrer Wahl einbauen - die Webseite soll HTML Daten, XML Daten und mind. fünf Bilder enthalten. Dafür nutzen Sie das Tool "Lucene" (<http://lucene.apache.org/>). Sie können damit Ihre Website indexieren lassen: HTML Dokumente, XML Daten und Ihre Bilder. Danach führen Sie selber ein paar eigene Abfragen auf Ihrer Website durch und bekommen hoffentlich relevante Ergebnisse.

Aufgabe 22: Website indexieren (15 Punkte)

Um Ihre Website indexieren zu lassen, brauchen Sie ein Indexing Tool. Dafür geeignet ist Lucene von Apache. Sie können Lucene Java von <http://www.apache.org/dyn/closer.cgi/lucene/java/> herunterladen. Momentan ist Version 2.4.0 verfügbar.

Damit Sie einen nützlichen Index bekommen, sollten Sie aufpassen, dass genügend Text auf der Website zu finden ist (am besten wäre es, wenn Sie einige Artikel aus Wikipedia kopieren). Sie sollten sich auch Gedanken machen, wie Sie Ihren Index verbessern können, indem Sie z.B. HTML Markup und Metadata passend verwenden. Sie müssen auch berücksichtigen, dass Ihre XML Daten und Ihre Bilder irgendwie durchsucht werden können.

Deswegen sollen Sie die Code für die Indexierung in Lucene entsprechend erweitern/ändern. Als Ergebnis sollen sie dann drei Lucene Indexes erhalten, in dem man durch Eingabe eines Suchtextes Ihre HTML Daten, XML Daten und (fünf) Bilder finden kann.

Aufgabe 23: Websuche implementieren (10 Punkte)

Sie können die Demo Web Anwendung von Lucene so erweitern, dass über Tomcat eine Website mit Suchfunktion realisiert wird. Diese Suche soll dann folgende drei Optionen anbieten: HTML, XML, Bilder.

Die Suchfunktion soll auch durch HTML Markup und/oder Metadata die Suchergebnisse besser ranken können.

Durch Texteingabe sollen Ergebnisse gerankt zurückgeliefert werden. Dieses Ranking sollte zudem auf der Ergebnisseite beschrieben werden: Wenn beispielsweise HTML Markup benutzt wird, und man Text in <TITLE> Elementen besser rankt als Text in <P> Elementen, sollte diese Vorgehensweise auch im Ergebnis angegeben werden.

Aufgabe 24: Website nach Inhalt abfragen (5 Punkte)

Die Website kann jetzt durchsucht werden. In Lucene, kann man 6 weitere Typen von Abfragen unterstützen (neben der Standardtextsuche): Field, Wildcard, Fuzzy, Proximity, Range, Boolean.

Formulieren Sie für jeden der Abfragetypen eine Beispielabfrage. Wenden Sie die Beispielanfragen auf Ihrer Website an (d.h. am besten suchen Sie nach Inhalten, die auf Ihrer Website zu finden sind) und merken Sie sich die Ergebnisse. Mindestens eine Abfrage soll Ihre XML Daten durchsuchen, und mindestens eine Abfrage soll Ihre Bilder durchsuchen.

Abgabe

Abgabe per E-Mail:

Schicken Sie drei bzw. vier Dateien:

Als ZIP Ihre Lucene Demo Java Dateien (modifiziert für die Aufgabe, eine bessere HTML Suche, eine XML Suche und eine Bildsuche zu realisieren)

Als WAR Ihre Lucene Web Anwendung, daß Sie für die Website Suche geändert haben.

Als WAR Ihre Website, falls Sie hier etwas geändert haben (z.B. mehr Inhalt, mehr Markup und/oder Metadata)

Als Text, Ihre sieben Musterfragen und der jeweils erste Treffer von der Suche. Erklären Sie die Antworten (z.B. weil Text in HTML TITLE war, oder wegen IMG ALT usw.)

Abgabe Mail an: paschke@inf.fu-berlin.de (gepackt)

Betreff: [CSW-NBI] Übung 6, <Name1>, <Matrikelnummer1>, <Name2>, <Matrikelnummer2>

Hinweis:

<http://lucene.apache.org/>

Viel Erfolg!