

Übung „Netzbasierte Informationssysteme“ WS 2008/2009

Übungsblatt 7

Ausgabe am 6.1.2009

Abgabe bis spätestens 20.1.2009, 16.00 Uhr

Aufgabe Kurzbeschreibung

Sie werden eine Website im World Wide Web crawlen, ähnlich wie die Suchmaschinen ihre Indizes aufbauen, indem Sie einen Crawler zum Einsatz bringen, um die Website zu indexieren. Hierbei wird das Crawling auf Seiten über wissenschaftliche Veröffentlichungen beschränkt. Sie können unterschiedliche Dokumenttypen indexieren, z.B. HTML, PDF. Mit dem Index werden Sie eine Suche implementieren, in der nach Links zu einer Veröffentlichung oder von einer Veröffentlichung gefragt werden kann.

Aufgabe 1: Website crawlen (10 Punkte)

Um eine Website zu crawlen, brauchen Sie ein Crawler. Dafür eignet sich Nutch, eine Erweiterung von Lucene. Sie können Nutch Java von <http://www.apache.org/dyn/closer.cgi/lucene/nutch/> herunterladen. Momentan ist die Version 0.9 verfügbar.

Um einen Index zu erstellen, müssen Sie erst entschieden, was im Web Sie crawlen wollen. Um genügend wissenschaftliche Veröffentlichungen zu finden, nutzen Sie die CiteSeer Website (<http://citeseer.ist.psu.edu>). Wählen Sie hier eine wissenschaftliche Veröffentlichung als Startseite aus (z.B. durch eine Suche von der CiteSeer Hauptseite nach einem technologischen Begriff).

Konfigurieren Sie den Crawler, um von der Seite dieser Veröffentlichung ein Crawl durchzuführen. Wählen Sie passende Werte für Crawl Depth (Anzahl von Stufen) und topN (Maximum Seiten per Stufe) - es wird ein Depth von 4 mit topN von 100 vorgeschlagen. Lassen Sie die Website crawlen und nutzen Sie die Nutch Web Anwendung, um in dem resultierenden Index zu suchen.

Aufgabe 2: Website Linksuche implementieren (20 Punkte)

Jetzt sollen Sie durch Plugins die Nutch Suche erweitern, damit man nach Inlinks und Outlinks in den gecrawlten Seiten suchen kann. In der Nutch Web Anwendung soll es dann möglich sein, eine (gerankte) Liste von Seiten zu bekommen, die Outlinks zu einer genannten Seite haben (die URL der Seite kann benutzt werden) bzw. Inlinks von einer genannten Seite haben. Zum Beispiel, die Suche mit dem Term 'in:"<http://citeseer.ist.psu.edu/225114.html>"' ergibt eine Liste von Seiten, die Links auf der Seite zu der Veröffentlichung 'Practical Reasoning for Expressive Description Logics' beinhalten.

Abgabe per E-Mail:

Schicken Sie an Ihren Dozenten die folgenden Dateien:

Code: schicken Sie als Java und JAR Datei(en) die Nutch Plugins, die Sie neu geschrieben oder geändert haben, um die Suche nach In- und Outlinks zu ermöglichen

Index: schicken Sie als gezippte Datei den Index, der durch das Web Crawl erstellt wurde. Nennen Sie sowohl die Startseite für Ihr Crawl als auch die gewählten Depth und topN Werte.

Analyse: beschreiben Sie in einer Textdatei zwei Queries auf Seiten in Ihrem Index, einmal eine Suche nach Inlinks und einmal eine Suche nach Outlinks. Für beide, geben Sie die Anzahl der Hits.

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

Betreff: [CSW-NBI] Übung 7, <Name1>, <Matrikelnummer1>, <Name2>, <Matrikelnummer2>

Viel Erfolg!