Analýza indexovania a vyhľadávania dokumentov

Vypracoval: Ľubomír Eľko
Dátum: 21.10.2009
Obsah: OpenSource-ové nástroje na indexovanie a vyhľadávanie dokumentov

Lucene Java - Demo

OpenSource podprojekt projektu **Apache Lucene**, ktorý poskytuje indexovanie a vyhľadávanie súborov. Založené na Jave. Lucene sa skladá z niekoľkých častí. Najdôležitejšou je asi indexátor (IndexWriter), ktorý pomocou vybraného analyzátora získa potrebné parametre o dokumente, ktorý chce používateľ zaindexovať a naplní týmito metadátami určené polia (Fields). Rozlišujeme 3 druhy analyzátorov:

- *SimpleAnalyzer* – používa iba Tokenizer, ktorý skonvertuje všetok vstup do malých písmen (angl. lower case).
- *StopAnalyzer* – používa rovnaký konvertor do malých písmen, no zároveň aj filtruje vstup podľa tzv. stopových slov, teda slov, ktoré nechceme aby sa indexovali (napr. the, that, a, b, use, atď.).
- *StandardAnalyzer* – používa konvertor do malých písmen aj filtrovanie stopových slov, navyše sa však snaží o prenesenie slova do základného tvaru (napr. vynecháva dokončenie apostrofom ['s]).

Po vytvorení indexu môžeme zadať vyhľadávací dotaz (angl. query), ktorý rovnaký analyzátor upraví do rovnakého tvaru ako bol upravený vstup do indexátora. Lucene vyhľadá (IndexSearcher) zodpovedajúce dokumenty a vráti nám tzv. hity (angl. Hits), čo je vlastne zoznam dokumentov zoradený podľa ich ratingu relevancie (angl. Score).

Lucene command-line demo pozostáva z dvoch aplikácií, ktoré demonštrujú rôzne funkcionality nástroja Lucene. Neindexuje názov súboru ale len obsah súborov typu HTML, TXT, plain/text a pod.), neindexuje PDF ani DOC. Veľkosť: cca 100 MB.
Po stiahnutí, rozzipovaní a pridaní dvoch JAR súborov do systémovej premennej Java CLASSPATH, je potrebné najprv zaindexovať súbory ľubovoľného priečinku príkazom:

java org.apache.lucene.demo.IndexFiles {cesta_k_priecinku}

Následne sa nám v aktuálnom priečinku vytvorí priečinok *index*, ktorý je pred každým ďalším novým indexovaním potrebné zmazať. Teraz už máme zaindexované súbory a môžeme ľahko vyhľadávať dokumenty podľa kľúčových slov príkazom:

java org.apache.lucene.demo.SearchFiles

Týmto príkazom sa nám spustí demo, ktoré nás vyzve k zadaniu dotazu na vyhľadanie. Potrebné je teda zadať kľúčové slovo.

Indexovanie DOC a PDF súborov pomocou Lucene sa dá docieliť tak, že na spracovanie dokumentov **DOC** použijeme knižnicu **Apache POI** (http://poi.apache.org/) a na **PDF** súbory knižnicu **PDFBox.org** (http://incubator.apache.org/pdfbox/). Pomocou týchto knižníc (ktoré majú podporu integrácie Lucene Search Engine) bude môcť aplikácia získať metainformácie z jednotlivých dokumentov a tiež

transformovať ich obsah do typu String, s ktorým je možné v aplikácii pracovať. Lucene následne vytvorí index, do ktorého budú načítané všetky nájdené dokumenty.

Zdroje:

1. Apache Lucene - Building and Installing the Basic Demo
   http://lucene.apache.org/java/2_9_0/demo.html
2. Apache POI - Java API To Access Microsoft Format Files
   http://poi.apache.org/
3. Apache PDFBox - Java PDF Library
   http://incubator.apache.org/pdfbox/

Lucene Apache Tika

OpenSource podprojekt projektu **Apache Lucene**. **Tika** je toolkit na detekciu a extrakciu metadát a štruktúrovaný text z rôznych dokumentov, ktorý používa existujúce knižnice na parsovanie. Podporuje formáty ako PDF, DOC, XLS, PPT, HTML, XML, JAR, RTF, TAR, ZIP, MP3, MIDI, WAV atď.

Zdroje:

1. Apache TIKA
   http://lucene.apache.org/tika/index.html
2. Podporované formáty + knižnice
   http://lucene.apache.org/tika/formats.html

Zilverline Search Engine

Zilverline by sa dal označiť ako Reverse Search Engine. Zilverline je indexátor a vyhľadávač, ktorý ponúka webový prístup k osobným dátam alebo dátam intranetu. Je dosť podobný Google Desktop, no pracuje na základe Lucene. Zilverline podporuje formáty PDF, Word, Excel, Powerpoint, RTF, TXT, java, CHM, rovnako ako aj zip, rar, a mnoho ďalších archívov. Zilverline je postavený na Jave, Lucene a Spring. K jeho spusteniu potrebujete Servlet Engine, ako je napr. Tomcat. Základom je vytváranie kolekcií (sada súborov a adresárov v adresári), ktoré sa indexujú a vyhľadáva sa v nich.

Zdroje:

Zilverline Search Engine: http://www.zilverline.org

BDDBot

BDDBot je webový robot, vyhľadávací nástroj a webový server napísaný v Jave. Je to príklad k výučbe ako tvoriť veľmi jednoduchý vyhľadávací nástroj. Dokáže indexovať len súbory typu HTML a plain/text. BDDBot prichádza s vlastným zabudovaným webovým serverom. Jeho indexy sú veľmi malé (cca 10% obsahu textu). Napriek tomu je veľmi malý (len niečo okolo 200 KB). Tento projekt skončil a už dlhšiu dobu sa nevyvíja ďalej.

Zdroje:

BDDBot: http://www.twmacinta.com/bddbot/