Eksploracja danych w internecie (Web-mining)

Zadanie nr 4. Indeksowanie dokumentów za pomocą biblioteki Apache Lucene(TM).

© mgr inż. Maciej Łaski
mlaski@kis.p.lodz.pl

1. Wstęp

Apche Lucene(TM) jest wysoce wydajną biblioteką do przeszukiwania tekstu napisaną całkowicie w języku Java. Biblioteka jest dostępna na licencji Apache License i można ją ściągnąć za darmo ze strony:

http://lucene.apache.org/core/

W celu zapoznania się z biblioteką Lucene proszę przejrzeć tutorial ze strony:

http://www.lucenetutorial.com/lucene-in-5-minutes.html oraz przestudiować delikatnie zmodyfikowaną wersję tego pliku, który jest dostępny na stronie:

http://mlaski.kis.p.lodz.pl/dane/LuceneTutorial.java

2. Opis zadania

Należy pobrać aktualną bazę książek z projektu Gutenberg www.gutenberg.org. Mirror do pliku jest dostępny na stronie http://mlaski.kis.p.lodz.pl/dane/pgdvd042010.iso, ponieważ jest to ponad 7GB danych. Celem zadania jest poindeksowanie wszystkich książek wchodzących w skład projektu po tytułach oraz zaimplementowanie prostej wyszukiwarki (jak w tutorialu). Można ograniczyć się do książek w formacie tekstowym. Są one spakowane zip'em, każda w swoim katalogu i mają rozszerzenie *.txt. Inne formaty można pominąć (czyli wszystkie katalogi ETEXT* i inne, które nie spełniają wymienionych kryteriów). Należy je rozpakować i przetworzyć. Struktura książki jest usystematyzowana i odczytanie tytułu nie powinno być problemem.

Na wykonanie zadania są przeznaczone 2 godziny laboratoryjne.

3. Sposób zaliczenia

Zaliczenie zadania na podstawie rozmowy z prowadzącym o sposobach implementacji i optymalizacji kodu.

Istnieje możliwość wykorzystania biblioteki lucene dla innych języków:

Python: http://lucene.apache.org/pylucene/

.NET: https://lucenenet.apache.org/

C++: http://sourceforge.net/projects/clucene/

PHP: http://www.lucenetutorial.com/techniques/lucene-php.html