

Eksploracja Danych w Internecie

Zadanie nr 6:

Indeksowanie treści dokumentów za pomocą biblioteki Apache Lucene(TM).

mgr inż. Maciej Łaski
dr inż. Piotr Duch

1. Wstęp

Apache 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/>

2. Opis zadania

Należy rozbudować program z zadania poprzedniego. Celem jest dodanie indeksowania dla całych dokumentów dzieląc je odpowiednio na tytuł i treść (należy pominąć nagłówek dodawany przez projekt Gutenberg). Należy wykonać interfejs wyszukiwarki książek po tytułach i treści dokumentu. W tym celu należy zapoznać się ze składnią języka zapytań projektu Lucene:

<http://www.lucenetutorial.com/lucene-query-syntax.html>

3. Bibliografia

<http://www.lucenetutorial.com/lucene-in-5-minutes.html>

4. Ocena

Oceniana jest poprawność wykonania zadania, znajomość kodu i zrozumienie teoretyczne podstaw problemu.