Eksploracja danych w internecie (Web-mining)

Zadanie nr 7. Bot internetowy

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

1. Opis zadania

Należy połączyć ze sobą dotychczas zrealizowane ćwiczenia i napisać program indeksujący strony internetowe. Do indeksowania można użyć biblioteki Lucene lub samemu zaimplementować indeks.

2. Sposób zaliczenia

Należy zaindeksować 10000 – 100000 stron www (stron i ich podstron łącznie). Prowadzący dostarczy plik z listą stron startowych, od których należy rozpocząć indeksowanie.

Ocena 3:

1. Program pobiera i indeksuje strony www
2. Program potrafi wyszukać witrynę zawierającą słowo, słowa lub dokładne dopasowanie do ciągu znaków (w tym spacji). Wyświetlone ma być 5 najbardziej trafnych zaindeksowanych stron.
3. Program posiada interfejs użytkownika i obsługę błędów.

Ocena 4:

1. Program indeksuje również adresy stron www, tak aby można było szukać po adresach i treści.

Ocena 5:

1. Program wyszukuje strony podobne. Obok wyszukanej strony ma pojawić się adres do strony o treści podobnej.

Na wykonanie zadania przeznaczone są 4 zajęcia laboratoryjne.

UWAGA:

Robot musi przestrzegać reguł zapisanych w pliku robots.txt

http://www.robotstxt.org