Universitatea Tehnică „Gheorghe Asachi” Iași
Facultatea de Automatica și Calculatoare
Profil : Calculatoare și Tehnologia Informației
Specializare: Tehnologia Informației

REGASIREA INFORMATIILOR PE WEB

PROIECT - ETAPA I

Profesor coordonator: ARCHIP ALEXANDRU
Nume: NECHITA IOAN-RĂZVAN
Grupa: 1409B

PROIECT - RIW

Scurtă descriere a proiectului:

Proiectul este format din 4 pachete:

1) Main: Contine clasa principala al programului in care se apeleaza, in functie de optiunea utilizatorului, indexarea directa, indexarea indirecta sau cautarea booleana;

2) Stemmer: Contine clasa care implementeaza algoritmul de stemming (Porter);

3) Search: Contine clasa care implementeaza algoritmul de cautare booleana;

4) htmlProcessing: Contine fisierele sursa care implementeaza indexarea directa, indexarea indirecta cat si algoritmii de prelucrare si parsare al informatiilor necesare construirii acestora:

- ExceptionsAndStopwords.java: contine functiile de prelucrare al informatiilor (nume, continut, link-uri), incarca fisierele de exceptii si stopword-uri si functia de calculare al numarului de aparitii al cuvintelor din text;

- Parser.java: contine functia de parcurgere al directoarelor in mod iterativ si functiile de parsare necesare construirii indecsilor;

- DirectIndex.java + DirectIndexObject.java: calcularea indexului direct: Se izoleaza numele fiecarui fisier cu structura de date ce contine perechea <cuvant, nr. aparitie> si se construieste un obiect de tipul DirectIndexObject cu aceste date. Acestea se vor scrie intr-un fisier de tip JSON;

- IndirectIndex.java + IndexIndirectObject.java: pe baza indexului direct, se calculeaza indexul indirect. Se construieste un obiect de tipul IndexIndirectObject;

Nota autoevaluare proiect: 6