UNIVERSITATEA TEHNICĂ „Gheorghe Asachi” din IAȘI
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE
DOMENIUL: Calculatoare și tehnologia informației
SPECIALIZAREA: Tehnologia informației

Proiect la disciplina
Regasirea informațiilor pe Web

ETAPA 1

Profesor coordonator: Archip Alexandru
Nume: Piu Claudiu-Catalin
Grupa: 1409B

Scurta descriere a proiectului

Proiectul este format din 4 pachete:

1. Pachetul clase

În acest pachet se află clasele ajutătoare folosite fie pentru a parsa documentele HTML fie pentru a stabili formatul necesar prelucrării datelor din index direct și index indirect

- IndexDirectTemplate.java este un template pentru stocarea indexului direct
- IndexDirectTemplate.java este un template pentru stocarea indexului direct
- Metadata.java este clasa pentru lucru cu metadatele dintr-un fișier HTML

2. Pachetul procesareHtml

În acest pachet se găsesc clasele cu ajutorul cărora se parsează documentelor HTML și se calculează indexul direct și indexul indirect:

- Data.java este clasa ajutătoare pentru prelucrarea documentelor HTML
  - în constructor se încarcă fișierele de excepții și stopwords
  - funcții pentru returnarea metadatelor, link-urilor, titlului și a cuvintelor dintr-un document HTML
- In clasa Parser.java are loc scrierea informațiilor din și despre documentele HTML în fișiere. Tot aici se parcurge un director care returnează toate fișierele din director.
- IndexDirect.java este clasa în care pe baza informațiilor oferite de Data.java creează indexul direct și stocarea lui în fisierul indexDirect.json.
- IndirectIndex.java se folosește de indexul direct pentru a crea indexul indirect și pentru a crea fisierul indexIndirect.json.

3. Pachetul stemmer

În acest pachet se află clasa Stemmer.java în care se regăsește algoritmul de stemmatizare.

4. Pachetul search

Acest pachet conține:

- CautareBooleana.java este clasa pentru căutarea booleană:
  - se încarcă indexul indirect
  - se primește query de utilizator
  - se prelucrează query separându-l în operanzi și operatori
  - se fac verificări asupra operanzilor în vederea identificării lor în documente

- se returnează un set de documente în funcție de operatorii induși(„+” -and, „-” - not, space – or)

5. Pachetul main

Acesta conține clasa Main.java care este main-ul programului:

- se instanțiază clasa Data
- se crează indexul direct / indexul indirect(în funcție de opțiunea utilizatorului)
- se realizează căutarea booleana

Nota autoevaluare proiect: 6