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

Proiect - RIW

Student: Narcis Ilie
Grupa: 1410B
Îndrumător: Alexandru Archip

2020

Etapa I

I. Descriere

Parcurgerea unui sistem de fișiere în mod recursiv pentru a identifica directoarele și fișierele conținute de directoare. După identificarea fișierelor se parcurge fiecare fișier și se citesc informațiile din ele urmând determinarea celor două forme de indexare:

- directă - pentru fiecare fișier se determina cuvintele existente

- indirectă - pentru fiecare cuvânt se determina fișierele care îl conțin.

Aceste două forme de indexare sunt stocate sub formă json atât în fișiere cât și într-o baza de date.

De asemenea este implementată și o metodă de căutare, căutare booleana, care parcurge fișierul cu indexarea indirectă și folosind metode sau, și, not, returnează fișierele ce conțin cuvintele respective în funcție de operația aleasă.

Pentru reducerea dimensiunii indecsilor se folosește un stemmer, care evita stocarea cuvintelor care reprezintă o conjugare sau o derivare a unui alt cuvânt existent. Stemmer-ul ales se numește porter.

II. De ce Porter?

1. În comparație cu alți algoritmi de stemming el are rezultatele cele mai bune și rata de eroare este mult mai mică.
2. Are suport pentru java.
3. Mai ușor de înțeles
4. Popular

2020