Responsabili:

- Alexandru GHEORGHIU
- Radu STOENESCU

Data publicarii: 21-10-2012

Ultima modificare: 04-11-2012

Ultima actualizare a testerului: 01-11-2012, 19:38

Termenul de predare: 06-11-2012, 23:55

Datorita faptului ca interfata vmchecker a fost pusa in functiune atat de tarziu si pentru ca site-ul de cursuri este picat de ceva vreme, deadline-ul a fost prelungit cu 2 zile.

Temele trebuie uploadate pe vmchecker, chiar daca au fost trimise pe site-ul de curs.

Cuprins

[ascunde]

Document retrieval

Document retrieval este o metoda de a identifica o multime de documente in care se gasesc (sau nu se gasesc) anumite cuvinte. Functionalitatea este asemanatoare cu cea a unui motor de cautare cu diferenta majora ca efectuam o cautare la nivel de cuvinte individuale si nu la nivel de grup de cuvinte (propozitii). Fiind mai simpla decat un motor de cautare, implementarea procedurii va avea o eficienta sporita si va fi utila in situatii in care avem un numar foarte mare de documente.

Cerinta

Tema presupune sa implementati document retrieval pornind de la un dictionar. Dictionarul este dat ca fisier text, iar cuvintele sunt in ordine lexicografica. El trebuie retinut intern intr-o structura de date care sa asigure acces rapid la orice element.

Un dictionar este o multime de asocieri de tip cheie-valoare. In cazul de fata, cheile sunt reprezentate de cuvinte, iar valorile sunt multimi de indici asociati documentelor. Se presupune ca dimensiunea dictionarului este foarte mare, deci pentru a gasi rapid un anumit cuvant in dictionar, trebuie utilizata o structura de date corespunzatoare. Va recomandam una din urmatoarele: trie, skip list, hash table.

Input-ul este reprezentat de o serie de interogari de tipul *w*, (*w1* and *w2*), (*w1* or *w2*), (*w1* or (*w2* and *w3*)) unde *w*, *w1*, *w2*, *w3* sunt cuvinte din dictionar. Pe cazul general, o interogare este o expresie *E* ce poate fi fie *w*, un cuvant din dictionar, fie (*E1* and *E2*), fie (*E1* or *E2*), unde *E1* si *E2* sunt subexpresii, definite in acelasi mod. O interogare simpla in care se da un singur cuvant va returna lista de documente in care se gaseste cuvantul respectiv. Interogarea (*E1* and *E2*) va returna lista de documente obtinute prin **intersectia** rezultatelor evaluarii expresiilor *E1* si *E2*. Interogarea (*E1* or *E2*) va returna lista de documente obtinute prin **reuniunea** rezultatelor evaluarii expresiilor *E1* si *E2*.

Cerintele temei sunt urmatoarele:

- Citirea dictionarului
- Implementarea unei structuri de date care sa retina dictionarul
- Parsarea interogarilor date ca input
- Procesarea interogarilor

Exemplu

Sa presupunem ca avem urmatorul dictionar:

apple 1 2 7 10 200

cherry 7 8 9 205 300

pineapple 2 3 4 5 6 7

strawberry 1 3 4 5 7 234 300

Consideram urmatoarele interogari precum si outputul intors:

apple -> 1 2 7 10 200

(apple and strawberry) -> 1 7

((pineapple and cherry) and strawberry) -> 7

(apple or cherry) -> 1 2 7 8 9 10 200 205 300

(apple and (pineapple or cherry)) -> 2 7

banana ->

(apple and banana) ->

(apple or banana) -> 1 2 7 10 200

Precizari implementare

Veti porni de la acest schelet de cod.

Programul primeste ca argument in linia de comanda numele fisierului dictionar si apoi citeste de la intrarea standard (stdin) interogari. Pentru fiecare interogare va afisa, la iesirea standard (stdout), o serie de numere separate prin spatiu, reprezentand indici de documente. Programul se opreste la comanda "exit".

Se presupune ca interogarile data la intrare sunt corecte si nu trebuie verificate. In schimb, pot exista situatii in care se dau cuvinte care nu exista in dictionar (vezi exemplu).

In Eclipse, pentru a da argumente in linia de comanda intrati in Run -> Run Configurations -> Arguments.

In dezvoltarea aplicatiei nu aveti voie sa folositi decat structuri de date implementate de voi. Cu alte cuvinte, nu puteti folosi colectii din java.util (ArrayList, HashMap etc).

Javadoc

Fisierele sursa trebuie comentate corespunzator:

- toate clasele si metodele trebuie sa fie insotite de blocuri de comentarii specifice limbajului Java. **Doar tagurile @param si @return sunt obligatorii pentru aceasta tema**;
- operatiile neintuitive din interiorul metodelor vor trebui comentate.

Se cere crearea documentatiei in format html, folosind utilitarul javadoc. Acesta se poate folosi prin intermediul IDE-ului sau direct din linia de comanda.

Comanda de creare a documentatiei: javadoc <lista fisiere sursa>

Documentatia completa poate fi gasita aici.

Incarcarea temelor

Trimiterea temelor se face folosind interfata vmchecker. Temele se vor incarca pe site sub forma unei arhive **zip**.

Arhiva va avea urmatoarea structura:

- **folderul src** - contine toate fisierele sursa;
- **folderul doc** - contine fisierele si folderele generate de utilitarul javadoc pentru fisierele sursa;
- **fisierul text README** - contine descrierea programului vostru

Notare

Notarea se va efectua dupa urmatoarea schema:

- 80 puncte - acordate de testerul automat;
- 20 puncte - respectarea conceptelor de POO studiate, existenta javadoc-urilor, calitatea codului, a comentariilor si a readme-ului - apreciate de asistent.

Pot exista situatii exceptionale in care nu se tine cont de aceasta schema de notare (de exemplu, daca tema este implementata doar pentru a trece testele si nu respecta cerintele temei, sau daca tema este copiata).

Temele vor fi comparate pentru a se depista cazuri de teme copiate. In cazul in care se considera ca doi studenti au temele copiate se va anula intreaga situatie de pe parcurs a acestora, conducand la repetarea materiei.

Pentru notare se va folosi un tester automat, toate testele fiind publice. Testerul este aici.

Daca testerul acorda 0 puncte, atunci acesta va fi punctajul pentru intreaga tema, indiferent de indeplinirea celorlalte cerinte! (Temele care nu compileaza nu vor fi luate in considerare).

Temele trimise inainte de aparitia tester-ului nu fac exceptie de la aceasta regula.

Testarea temelor se va face utilizand interfata vmchecker.

Pentru fiecare zi de intarziere se vor scadea cate 5/100 puncte. Dupa 14 zile de la trecerea termenului nu se vor mai putea incarca arhive pe site.