ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Δομές Δεδομένων - Εργασία 3

Τμήμα Πληροφορικής

Φθινοπωρινό Εξάμηνο 2021-2022

Διδάσκων: Ε. Μαρκάκης

Προθεσμία παράδοσης: Κυριακή, 30/1/2022, 23:59

Δέντρα Δυαδικής Αναζήτησης

Σκοπός της εργασίας 3 είναι η εξοικείωση με τις δομές που χρησιμοποιούνται για την υλοποίηση πίνακα συμβόλων, όπως τα δέντρα δυαδικής αναζήτησης (σχετικές ενότητες διαφανειών: 12-14). Το ζητούμενο της εργασίας είναι να κατασκευάσετε ένα Μετρητή Λέξεων. Η βασική λειτουργία του μετρητή είναι ότι θα μπορεί να "φορτώνει" ένα αρχείο αγγλικού κειμένου και θα μετρά πόσες φορές εμφανίζεται η κάθε λέξη. Για παράδειγμα, ας δούμε το παρακάτω κέιμενο.

- Hello, how are you?
- Very well, thank you. I study for the exams. How about you?
- Fine, thank you. How many exams will you have?
- Too many…

Η λέξη «you» εμφανίζεται 5 φορές, η λέξη «how» 3 φορές, οι λέξεις «thank», «many», «exams» από 2 φορές, ενώ από 1 φορά εμφανίζονται οι λέξεις hello, are, very, well, I, study, about, fine, will, have. Ο Μετρητής Λέξεων θα πρέπει να μην λαμβάνει υπόψη τα σημεία στίξης, π.χ., τα , . ? ; ! - :, καθώς και παρενθέσεις, αγκύλες, σύμβολα πράξεων, ή οποιονδήποτε άλλο χαρακτήρα που δεν είναι γράμμα του αγγλικού αλφαβήτου. Από το κείμενο παραπάνω, όταν διαβάσει π.χ. το string "well," θα πρέπει να βγάλει το κόμμα για να μείνει μόνο η λέξη well. Αν υπήρχε φράση σε παρενθέσεις, π.χ. "(Maria asked me)", θα πρέπει να θεωρήσει ότι υπάρχουν μόνο οι λέξεις Maria, asked και me. Επιτρέπεται ΜΟΝΟ η μονή απόστροφος μέσα σε λέξη, π.χ. το don't θα το θεωρούμε σαν 1 λέξη. Επίσης θα πρέπει να αγνοούνται πλήρως όλα τα strings που περιέχουν αριθμούς π.χ. 17:25 ή 1980's. Τέλος θα πρέπει να δίνεται η δυνατότητα στο χρήστη της βιβλιοθήκης να ορίζει ειδικές λέξεις (stop words) που θέλει να αγνοούνται πλήρως, π.χ. σε σχετικές εφαρμογές αγνοούμε συνήθως τα άρθρα, όπως τα "a", "an" και "the". Το πρόγραμμα θα είναι case in-sensitive. Για παράδειγμα, οι λέξεις Hello και hello είναι ίδιες.

Μέρος Α [10 μονάδες]. Για να ξεκινήσουμε, θα πρέπει πρώτα να ορίσετε τις εξής 2 κλάσεις.

Η κλάση WordFreq. Για κάθε διαφορετική λέξη που διαβάζετε, θα πρέπει να δημιουργείτε ένα αντικείμενο με κλειδί αυτή τη λέξη. Συγκεκριμένα, σε κάθε λέξη θα αντιστοιχεί ένα αντικείμενο τύπου WordFreq. Η κλάση αυτή περιέχει (τουλάχιστον) 2 πεδία: την ίδια τη λέξη (private String) και τον αριθμό εμφανίσεων (private int). Μπορείτε εδώ να υπερφορτώσετε κατάλληλα την μέθοδο toString για να σας χρησιμεύσει για εκτύπωση αποτελεσμάτων. Η WordFreq πρέπει να περιέχει και μια μέθοδο key() που θα επιστρέφει το κλειδί, δηλαδή τη λέξη.

Η κλάση TreeNode. Η κλάση αυτή θα είναι ιδιωτική μέσα στον πίνακα συμβόλων (δείτε το Μέρος Β) και αντικείμενα αυτής της κλάσης αντιστοιχούν στους κόμβους του δέντρου δυαδικής αναζήτησης που θα χρησιμοποιήσετε. Κάθε κόμβος του δέντρου είναι ένα αντικείμενο TreeNode, και πρέπει να περιέχει ένα αντικείμενο τύπου WordFreq, όπως περιγράφεται παραπάνω. Επιπλέον, πρέπει να περιέχει τους δείκτες προς το αριστερό και δεξιό υποδέντρο, και ένα πεδίο που θα δηλώνει πόσους κόμβους έχει το υποδέντρο που ξεκινά από αυτόν τον κόμβο. Επομένως στην κλάση TreeNode πρέπει να υπάρχουν τουλάχιστον τα εξής πεδία (και ενδεχομένως ό,τι άλλο θέλετε εσείς να προσθέσετε):

```
private class TreeNode {
   WordFreq item
   TreeNode left   // pointer to left subtree
   TreeNode right  // pointer to right subtree
   int subtreeSize //number of nodes in subtree starting at this node
   ...
}
```

Η πρόσβαση στο κλειδί του κόμβου θα πρέπει να γίνεται μέσω της μεθόδου key() του item. Αν h είναι ένα αντικείμενο τύπου TreeNode, το κλειδί του κόμβου θα το παίρνετε από την κλήση `h.item.key()`, όπως και στις διαφάνειες του μαθήματος.

Μέρος Β [80 μονάδες]. Η κλάση του πίνακα συμβόλων/ΔΔΑ. Η δομή σας για την υλοποίηση του πίνακα συμβόλων θα είναι ένα ΔΔΑ (Binary Search Tree). Η κλάση που θα ορίσετε θα λέγεται BST.java και θα ακολουθεί το παρακάτω υπόδειγμα:

```
public class BST implements WordCounter {
private class TreeNode {
...
};
private TreeNode head; //root of the tree
private List stopWords; // list of stopwords
...
}
```

Το interface WordCounter περιέχει τις εξής μεθόδους (ακολουθούν επεξηγήσεις)

```
public interface WordCounter {
void insert(String w);
WordFreq search(String w);
void remove(String w)
void load(String filename);
int getTotalWords();
int getDistinctWords();
int getFrequency(String w);
WordFreq getMaximumFrequency();
```

```
double getMeanFrequency();
void addStopWord(String w);
void removeStopWord(String w);
void printTreeAlphabetically(PrintStream stream);
void printTreeByFrequency(PrintStream stream);
}
```

Συνοπτική περιγραφή των απαιτούμενων μεθόδων:

- `void insert(String w)`: ψάχνει να βρει αν υπάρχει ήδη στο δέντρο κόμβος με κλειδί w. Αν ναι, τότε του αυξάνει τη συχνότητα κατά 1. Αν όχι, τότε εισάγει ένα νέο κόμβο στο δέντρο (χρησιμοποιώντας την απλή εισαγωγή ως φύλλο), με αυτό το κλειδί και με συχνότητα ίση με 1.
- `WordFreq search(String w)`: Ψάχνει στο δέντρο για την ύπαρξη της λέξης w (επιστρέφει null αν δεν υπάρχει). Η μέθοδος search θα δουλεύει αρχικά όπως και η αντίστοιχη μέθοδος που έχουμε δει στο μάθημα με την εξής όμως σημαντική διαφοροποίηση: όταν βρίσκει τη λέξη w μέσα στο δέντρο, αν η συχνότητα της w είναι μεγαλύτερη της μέσης συχνότητας (από την `getMeanFrequency()`), τότε με χρήση περιστροφών θα φέρνει τη λέξη αυτη στη ρίζα του δέντρου. Ένας τρόπος να γίνει αυτό (αλλά όχι και ο μοναδικός), είναι να καλείτε πρώτα τη `remove` για να βγάλει τον κόμβο αυτό από το ΔΔΑ και στη συνέχεια να κάνετε εισαγωγή στη ρίζα για τον συγκεκριμένο κόμβο. Το σκεπτικό είναι ότι λέξεις με μεγάλη συχνότητα είναι λέξεις για τις οποίες μπορεί να γίνουν πολλές αναζητήσεις και επομένως θέλουμε να τις έχουμε όσο το δυνατόν πιο ψηλά στο δέντρο.
- `void remove(String w)`: αφαιρεί τον κόμβο με κλειδί w (αν υπάρχει τέτοιος κόμβος). Μπορείτε να χρησιμοποιήσετε τη μέθοδο αφαίρεσης που έχουμε δει στο μάθημα.
- `void load(String filename)`: ξεκινώντας από το τρέχον ΔΔΑ (που μπορεί να είναι και κενό), διαβάζει το αρχείο με όνομα filename, με κείμενο Αγγλικής γλώσσας, και ενημερώνει το δέντρο με τις λέξεις που διάβασε ώστε να φτιαχτεί το τελικό ΔΔΑ. Θα πρέπει να τηρούνται οι προϋποθέσεις που αναφέρθηκαν στην Σελίδα 1 σχετικά με τα σημεία στίξης, την ύπαρξη αριθμών, την αγνόηση των λέξεων που έχουν δοθεί ως stop words, κτλ. Δεν απαιτείται να ελέγξετε αν το κείμενο είναι όντως Αγγλικά, ούτε και αν οι λέξεις είναι έγκυρες (π.χ. αν κατά λάθος υπάρχει η λέξη matematics αντί για το σωστό mathematics, θα την θεωρήσει σαν μια κανονική νέα λέξη).
- `int getTotalWords()`: επιστρέφει τον συνολικό αριθμό λέξεων του κειμένου που έχει φορτωθεί στο ΔΔΑ, λαμβάνοντας υπόψη τη συχνότητα κάθε λέξης (έχοντας δηλαδή αγνοήσει ήδη stopwords, αριθμούς, κτλ). Μπορεί να γίνει με απλή διάσχιση δέντρου (όποια διάσχιση θέλετε).
- `int getDistinctWords()`: επιστρέφει τον αριθμό διαφορετικών λέξεων του δέντρου. **Πρέπει να τρέχει σε O(1).**
- `int getFrequency(String w)`: επιστρέφει τον αριθμό εμφανίσεων της λέξης w (αν η λέξη δεν υπάρχει στο δέντρο, επιστρέφει 0).
- `WordFreq getMaximumFrequency()`: επιστρέφει ένα αντικείμενο WordFreq που περιέχει τη λέξη με τις περισσότερες εμφανίσεις (δεν χρειάζεται να κάνετε ταξινόμηση για να λύσετε το πρόβλημα αυτό). Σε ισοβαθμίες, μπορείτε να επιλέξετε αυθαίρετα ποια λέξη θα επιστρέψετε.
- `double getMeanFrequency()`: υπολογίζει και επιστρέφει την μέση συχνότητα. Ο μέσος όρος παράγεται από τις συχνότητες όλων των διαφορετικών λέξεων μέσα στο κείμενο.
- `void addStopWord(String w)`: προσθέτει στη λίστα stopWords τη λέξη w. **Πρέπει να τρέχει σε O(1).**
- `void removeStopWord(String w)`: αφαιρεί τη λέξη w από τη λίστα stopWords.
- `printTreeAlphabetically(PrintStream stream)`: εκτυπώνει τις λέξεις του δέντρου, μαζί με τον αριθμό εμφανίσεων κάθε λέξης, με αλφαβητική σειρά. Πρέπει να υλοποιηθεί με κάποια μέθοδο διάσχισης του δέντρου.
- `printTreeByFrequency(PrintStream stream)`: εκτυπώνει τις λέξεις και τον αριθμό εμφανίσεων ταξινομημένες σε αύξουσα σειρά ως προς τον αριθμό εμφανίσεων.

Σχόλια και πρόσθετες οδηγίες υλοποίησης:

- Μπορείτε να βασιστείτε στον κώδικα από τις διαφάνειες και από τα εργαστήρια για ΔΔΑ.

- Χρησιμοποιήστε τις μεθόδους της βασικής βιβλιοθήκης της Java για την μετατροπή κεφαλαίων σε μικρούς χαρακτήρες, την αφαίρεση των σημείων στίξης και όποια άλλη επεξεργασία κάνετε για τις λέξεις. Όπως και στις Εργασίες 1 και 2, υπάρχουν διάφορες μέθοδοι που μπορείτε να χρησιμοποιήσετε για να διαβάσετε και να επεξεργαστείτε κείμενο.
- Δεν θα θεωρήσετε κάποια stop word ως δεδομένη. Η λίστα με τα stopwords θα δημιουργείται μέσα από κλήσεις της addStopWord. Για τη λίστα stopwords υλοποιήστε όποιο τύπο λίστας θέλετε (μονής ή διπλής σύνδεσης), χωρίς την χρήση έτοιμων υλοποιήσεων της Java.
- Το κλειδί των αντικειμένων που αποθηκεύουμε στο δέντρο είναι τύπου String. Επομένως σύγκριση κλειδιών εδώ σημαίνει σύγκριση μεταξύ strings. Επίσης, κάθε κλειδί θα εμφανίζεται το πολύ σε έναν κόμβο του δέντρου.
- Αρκετές μέθοδοι χρειάζονται διάσχιση του δέντρου. Θα πρέπει λοιπόν να υλοποιήσετε μέσα στον πίνακα συμβόλων και κάποια ή κάποιες μεθόδους διάσχισης.
- Προσοχή να γίνεται σωστή ενημέρωση του πεδίου subtreeSize, της κλάσης TreeNode.
- Η υλοποίηση της printTreeByFrequency μπορεί να γίνει είτε με συνδυασμό κάποιας διάσχισης και μετέπειτα κάποιας μεθόδου ταξινόμησης (χρησιμοποιώντας Quicksort, Mergersort ή Heapsort), είτε με άλλους τρόπους. Εναλλακτικά, π.χ. μπορείτε (χωρίς να είναι απαραίτητο, με τον πρώτο τρόπο λύνεται πιο απλά), όταν καλείται η printreeByFrequency να δημιουργείτε επί τόπου ένα προσωρινό ΔΔΑ με τον κατάλληλο τύπο κλειδιού και έναν κατάλληλο Comparator (σκεφτείτε τι πρέπει να συγκρίνετε) για να διατρέξετε το δέντρο με βάση τον αριθμό εμφανίσεων της κάθε λέξης.
- **Πολυπλοκότητα μεθόδων:** Δεν υπάρχει αυστηρή απαίτηση να πετύχετε την βέλτιστη πολυπλοκότητα όλων των μεθόδων, παρά μόνο για τις μεθόδους που αναφέρεται κάτι ρητά στην επεξήγηση τους στην προηγούμενη σελίδα. Μπορείτε να έχετε ως γνώμονα ό,τι έχουμε πει και στο μάθημα. Αν όμως κάνετε κάτι υπερβολικά χρονοβόρο, π.χ. αν υλοποιήσετε την printAlphabetically σε χρόνο $O(N^2)$, τότε δεν θα πάρετε όλες τις μονάδες που αντιστοιχούν στη μέθοδο αυτή. Γενικά θα πρέπει να αποφύγετε το $O(N^2)$ για όλες τις μεθόδους εκτός της load.
- **Δεν επιτρέπεται να χρησιμοποιήσετε έτοιμες υλοποιήσεις δομών δεδομένων για λίστες, ουρές, δέντρα, κτλ, από την βιβλιοθήκη της Java** (π.χ. Vector, ArrayList κτλ).

Προαιρετικά: Όποιος θέλει, μπορεί αντί για ΔΔΑ να χρησιμοποιήσει κάποια από τις δομές του Κεφαλαίου 13, που έχουν ως αποτέλεσμα το δέντρο να είναι πιο ισοζυγισμένο, π.χ. τυχαιοποιημένα ΔΔΑ ή δέντρα κόκκινου-μαύρου. Επίσης, μπορείτε αν θέλετε να έχετε και μια μέθοδο main η οποία να κάνει κάποιο ενδεικτικό τρέξιμο, π.χ. να προσθέτει κάποια stopwords, μετά να καλεί τη μέθοδο load, και μετά να τυπώνει τη μέση συχνότητα, ή ακόμα και να εμφανίζει κάποιο μενού διαχείρισης (αυτό είναι καλό να το κάνετε ούτως ή άλλως για να ελέγξετε τις μεθόδους σας).

Μέρος Γ - Αναφορά παράδοσης [10 μονάδες]. Ετοιμάστε μία σύντομη αναφορά σε pdf αρχείο (μην παραδώσετε Word ή txt αρχεία!) με όνομα project3-report.pdf, στην οποία θα αναφερθείτε στα εξής:

a. Εξηγήστε συνοπτικά πώς υλοποιήσατε κάθε μέθοδο του Μέρους Β (αρκούν το πολύ 4-5 γραμμές για κάθε μέθοδο).
b. Σχολιάστε την πολυπλοκότητα των μεθόδων αυτών.

Το συνολικό μέγεθος της αναφοράς θα πρέπει να είναι τουλάχιστον 2 σελίδες. Μην ξεχνάτε τα ονοματεπώνυμα και τους ΑΜ σας στην αναφορά.

Οδηγίες Παράδοσης

Η εργασία σας θα πρέπει να μην έχει συντακτικά λάθη και να μπορεί να μεταγλωττίζεται. Εργασίες που δεν μεταγλωττίζονται χάνουν το **50%** της συνολικής αξίας.

Η εργασία θα αποτελείται από:

1. Τον πηγαίο κώδικα (source code). Τοποθετήστε σε ένα φάκελο με όνομα **src** τα αρχεία java που έχετε φτιάξει. Χρησιμοποιήστε τα ονόματα των κλάσεων όπως

δίνονται. Επιπλέον, φροντίστε να συμπεριλάβετε όποια άλλα αρχεία πηγαίου κώδικα φτιάξατε και απαιτούνται για να μεταγλωττίζεται η εργασία σας. Φροντίστε επίσης να προσθέσετε επεξηγηματικά σχόλια όπου κρίνετε απαραίτητο στον κώδικά σας.

2. Την αναφορά παράδοσης.

Όλα τα παραπάνω αρχεία θα πρέπει να μπουν σε ένα αρχείο zip. Το όνομα που θα δώσετε στο αρχείο αυτό θα είναι ο αριθμός μητρώου σας πχ. 3030056_3030066.zip ή 3030056.zip (αν δεν είστε σε ομάδα). Στη συνέχεια, θα υποβάλλετε το zip αρχείο σας στην περιοχή του μαθήματος «Εργασίες» στο e-class. Δεν χρειάζεται υποβολή και από τους 2 φοιτητές μιας ομάδας.