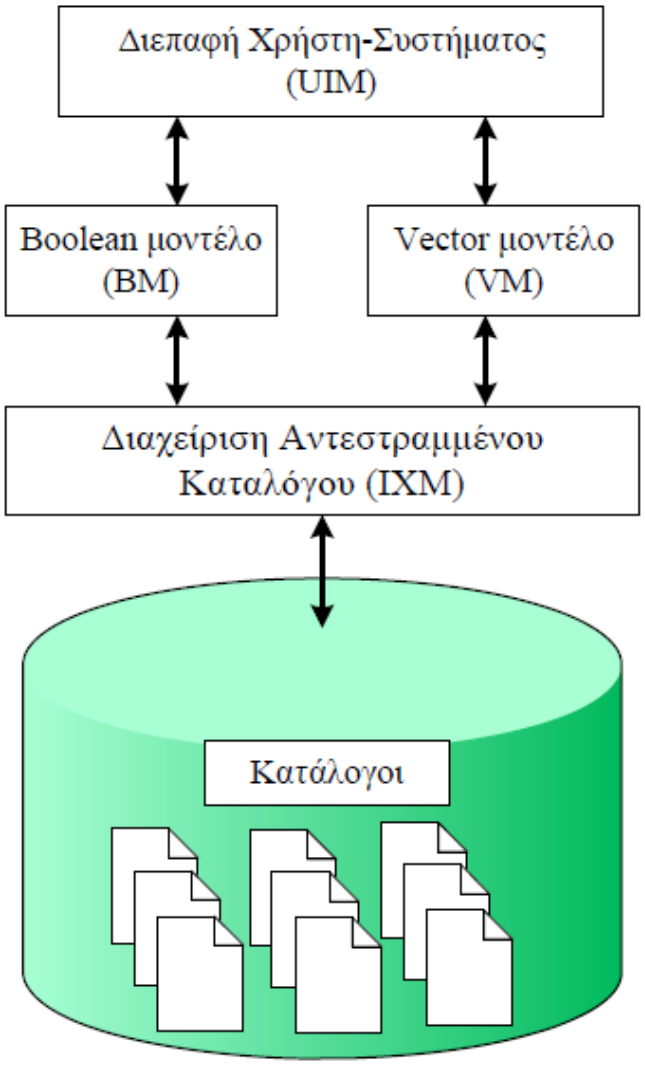
Εργασία στην Ανάκτηση Πληροφορίας informAtion REtrieval System

Γκανιάς Ευριπίδης *1866*
Γιαννουλούδης Στέργιος *1877*

Πρόλογος

Η αρχιτεκτονική του συστήματος φαίνεται από το παρακάτω σχήμα. Η κατασκευή του ξεκίνησε από τα χαμηλότερα επίπεδα στα υψηλότερα. Παρακάτω φαίνονται με τη σειρά υλοποίησης τα επίπεδα και οι κλάσεις τους.



Σχήμα 1. Αρχιτεκτονική συστήματος ARES.

Περιεχόμενα

Διαχείρηση Ανεστραμμένου Καταλόγου (IXM)

Το επίπεδο αυτό περιέχει κάποιες δομές για τη διευκόλυνση της αναπαράστασης των ανεστραμμένων καταλόγων, καθώς και τις βασικές λειτουργίες και κάποιες στατιστικές πληροφορίες για τους καταλόγους αυτούς. Σε αυτό το επίπεδο έχουν υλοποιηθεί δομές όπως: ανεστραμμένος κατάλογος, χειριστής και διαχειριστής ανεστραμμένου καταλόγου, λίστα εγγράφων και πληροφορίες εγγράφου, οι οποίες αναλύονται παρακάτω.

Πληροφορίες Εγγράφου (Document Info)

Συγκρατεί πληροφορίες για έναν όρο που βρίσκεται σε ένα συγκεκριμένο έγγραφο. Οι πληροφορίες αυτές είναι το όνομα του εγγράφου, η συχνότητα του όρου μέσα στο έγγραφο και το βάρος του γι' αυτό το έγγραφο.

Ο σκοπός της δομής είναι η άμεση ανάκτηση του βάρους του όρου για το έγγραφο όταν χρειάζεται να υπολογιστούν συγκεκριμένα μεγέθη, πράγμα που συνεπάγεται τη γρηγορότερη εκτέλεση ερωτημάτων.

Λίστα Εγγράφων Καταλόγου (Document List)

Μία λίστα με τα ονόματα όλων των εγγράφων που περιέχουν μια λέξη του ανεστραμμένου καταλόγου. Κρατάει πληροφορία για τη συχνότητα της λέξης συνολικά σε όλα τα έγγραφα.

Σκοπός είναι η ευκολότερη και γρηγορότερη αναζήτηση των τιμών που χρειάζονται για τον υπολογισμό συγκεκριμένων μεγεθών.

Επεξεργαστής Ειδικών Εγγράφων (Parser)

Είναι μια δομή που χειρίζεται τα ειδικά έγγραφα. Ειδικά έγγραφα είναι:

- Αυτά που περιέχουν τα έγγραφα μια συλλογής
- Αυτά που περιέχουν τα ερωτήματα για μια συλλογή
- Αυτά που περιέχουν τα επιθυμητά αποτελέσματα των ερωτημάτων για κάποια συλλογή

Στην πρώτη περίπτωση, διαβάζει το έγγραφο της εισόδου και δημιουργεί τα αρχεία στον φάκελο “DOCS” μέσα στον φάκελο της συλλογής. Στη δεύτερη, παίρνει ως είσοδο το όνομα του αρχείου και επιστρέφει μια λίστα με τα ερωτήματα που περιέχει. Τέλος, στην τελευταία περίπτωση, δέχεται ως όρισμα το όνομα του αρχείου και επιστρέφει (για κάθε ερώτημα) μια λίστα με τα ονόματα των σχετικών εγγράφων ενός ερωτήματος.

Ανεστραμμένος Κατάλογος (Index)

Μια δομή που χρησιμοποιεί τις ιδιότητες ενός χάρτη για την αποθήκευση πληροφοριών για τις λέξεις μιας συλλογής. Οι πληροφορίες που αποθηκεύει είναι:

- Το όνομα του καταλόγου
- Μια Λίστα Εγγράφων Καταλόγου (Document List)για κάθε όρο που βρίσκεται σε κάποιο από τα κείμενα που έχουν εισαχθεί
- Τα ονόματα των εγγράφων που έχουν εισαχθεί στον ανεστραμμένο κατάλογο
- Την τιμή της συχνότητας του όρου που εμφανίζεται πιο συχνά σε ένα έγγραφο, για κάθε έγγραφο

Σκοπός των δύο τελευταίων δομών είναι η ευκολότερη και γρηγορότερη αναζήτηση των τιμών που χρειάζονται για τον υπολογισμό συγκεκριμένων μεγεθών. Ενημερώνονται κάθε φορά που εισάγεται η διαγράφεται ένα έγγραφο από τη συλλογή.

Η δομή αυτή περιέχει και κάποιες επιπλέον λειτουργίες:

- **Εισαγωγή εγγράφου σε έναν όρο (put)**

  Αν ο όρος δεν υπάρχει, τον δημιουργεί και βάζει το έγγραφο στη λίστα του. Αν το έγγραφο υπάρχει ήδη στη λίστα του όρου, απλά αυξάνει τη συχνότητά του. Τέλος, ενημερώνει τη λίστα των εγγράφων και των μέγιστων συχνοτήτων.

- **Διαγραφή ενός όρου (remove)**

  Διαγράφει έναν όρο από το χάρτη και ενημερώνει τη λίστα των συχνότερων όρων. Αν κάποιο από τα έγγραφα του όρου που διαγράφουμε δεν βρίσκεται στη λίστα των εγγράφων κάποιου άλλου όρου, τότε διαγράφουμε και το έγγραφο από τη λίστα των εγγράφων.

- **Διαγραφή ενός εγγράφου (remove doc)**

  Ενημερώνει τις δομές του καταλόγου και αφαιρεί το έγγραφο από τις λίστες των λέξεων που περιέχει. Αν κάποιος όρος δεν έχει άλλο έγγραφο στη λίστα του, τον αφαιρεί από τον κατάλογο. Ενημερώνει τα βάρη των όρων.

- **Αφαίρεση συχνών όρων**

  Ορίζουμε σαν συχνό όρο, τον όρο που εμφανίζεται σε περισσότερα από το 80% των εγγράφων μίας συλλογής. Για να το διαπιστώσουμε αυτό συγκρίνουμε το μέγεθος της λίστας εγγράφων του όρου με τον συνολικό αριθμό των εγγράφων της συλλογής. Αν είναι συχνός, αφαιρείται από τη συλλογή. Αυτή η λειτουργία καλείται μετά την μαζική εισαγωγή όρων και εγγράφων στον ανεστραμμένο κατάλογο ώστε να μαζέψει πρώτα όλα τα δεδομένα του καταλόγου και έπειτα να αφαιρέσει τους συχνούς όρους. Αν η εισαγωγή των δεδομένων γίνει μία προς μία, τότε μόλις η συχνότητα ξεπεράσει το 80% ο όρος θα διαγραφεί, αλλά θα προστεθεί ξανά όταν ο όρος βρεθεί σε ένα επόμενο έγγραφο.

- **Υπολογισμός βάρους όρου στο έγγραφο**

  Παίρνουμε τις τιμές των μεγεθών που χρειαζόμαστε από τις αντίστοιχες δομές του καταλόγου, υπολογίζουμε το βάρος του όρου χρησιμοποιώντας τις κατάλληλες συναρτήσεις και το ενημερώνουμε.

Διαχειριστής Ανεστραμμένων Καταλόγων(Index Manager)

Είναι υπεύθυνος για τη δημιουργία, διαγραφή, άνοιγμα, αποθήκευση και κλείσιμο ενός ανεστραμμένου καταλόγου. Περιέχει πληροφορία για το source φάκελο των ανεστραμμένων καταλόγων. Χρησιμοποιείται για διαχείριση του καταλόγου με απόκρυψη της δομής αυτού.

Χειριστής Ανεστραμμένου Καταλόγου (Index Handle)

Δίνει τη δυνατότητα στον χρήστη να χειριστεί τον ανεστραμμένο κατάλογο. Πιο συγκεκριμένα, του δίνει τη δυνατότητα εισαγωγής και διαγραφής ενός ή πολλαπλών εγγράφων. Επίσης, μπορεί να δώσει πληροφορίες στο χρήστη για το αν ο κατάλογος είναι ανοιχτός ή κλειστός, αν υπάρχουν ή όχι αλλαγές σε αυτόν, ακόμα και το όνομα της συλλογής από την οποία δημιουργήθηκε ο κατάλογος. Τέλος, επιτρέπει την εισαγωγή και ανάγνωση του ίδιου του καταλόγου μέσω κατάλληλων συναρτήσεων.

Boolean Μοντέλο (BM)

Αυτό το επίπεδο διαχειρίζεται τα λογικά ερωτήματα. Περιέχει πέντε βασικές λειτουργίες, οι οποίες αναλύονται παρακάτω.

Εκτέλεση ερωτήματος

Μετατρέπει το ερώτημα με την κατάλληλη συνάρτηση από μορφή Infix σε postfix. Δημιουργεί ένα πίνακα όπου εισάγει τους όρους του postfix ερωτήματος. Δημιουργεί μία δομή στοίβας την οποία χρειάζεται για την εκτέλεση των υπο-ερωτημάτων. Διαβάζει τα στοιχεία του πίνακα ένα-προς-ένα και παίρνει περιπτώσεις:

1. Αν είναι κάποια λέξη, ανατρέχει στον ανεστραμμένο κατάλογο και παίρνει μια λίστα με τα έγγραφα που την περιέχουν, την οποία και βάζει στη στοίβα.

2. Αν είναι ο όρος “NOT”, εκτελεί την αντίστοιχη συνάρτηση για τον πρώτο όρο της στοίβας, τον οποίο αφαιρεί από αυτή και προσθέτει το αποτέλεσμα.

3. Αν είναι ο όρος “AND” ή “OR”, εκτελεί την αντίστοιχη συνάρτηση για τους δύο πρώτους όρους της στοίβας, τους οποίους αφαιρεί από αυτή και προσθέτει το αποτέλεσμα.

Τέλος, επιστρέφει τη λίστα με τα συνολικά αποτελέσματα του ερωτήματος.

Μετατροπή ερωτήματος σε μορφή κατάλληλη για ευκολότερη επεξεργασία

Δέχεται ένα ερώτημα σε μορφή Infix και το μετατρέπει σε μορφή Postfix. Επιστρέφει το τελικό αποτέλεσμα σε μορφή συμβολοσειράς.

Εφαρμογή του τελεστή της ένωσης

Δέχεται δύο λίστες από ονόματα εγγράφων, εφαρμόζει σε αυτές τον τελεστή “OR” και επιστρέφει το τελικό αποτέλεσμα σε μορφή λίστας.

Εφαρμογή του τελεστή της τομής

Δέχεται δύο λίστες από ονόματα εγγράφων, εφαρμόζει σε αυτές τον τελεστή “AND” και επιστρέφει το τελικό αποτέλεσμα σε μορφή λίστας.

Εφαρμογή του τελεστή της άρνησης

Δέχεται μία λίστα από ονόματα εγγράφων, εφαρμόζει σε αυτή τον τελεστή “NOT” χρησιμοποιώντας τη λίστα των εγγράφων της συλλογής που είναι αποθηκευμένη στον κατάλογο και επιστρέφει το τελικό αποτέλεσμα σε μορφή λίστας.

Vector Μοντέλο (VM)

Αυτό το επίπεδο διαχειρίζεται τα μη λογικά ερωτήματα. Αποτελείται από δύο βοηθητικές δομές οι οποίες δημιουργήθηκαν για τη διευκόλυνση ομαδοποίησης κάποιων δεδομένων και μια κεντρική δομή που διαχειρίζεται τα ερωτήματα που τίθενται στο μοντέλο.

Βαθμολογία Εγγράφου (Document Rank)

Συγκρατεί πληροφορίες για την βαθμολογία ενός εγγράφου για κάποιο ερώτημα.

Μετρικές (Metrics)

Σκοπός της κλάσης είναι να υπολογίσει τα μεγέθη Recall και Precision για κάθε ερώτημα μίας συλλογής. Για το λόγο αυτό, έχει σαν πεδία τα μεγέθη που χρειάζονται για των υπολογισμό των παραπάνω τα οποία υπολογίζει και αρχικοποιεί.

Διαχειριστής Διανυσματικού Μοντέλου (Vector Model Manager)

Είναι υπεύθυνη για την εκτέλεση ερωτημάτων. Παρέχει μια συνάρτηση για τον υπολογισμό του βάρους ενός όρου στο query, καθώς και μια συνάρτηση για να διαβάσει το βάρος του όρου που είναι αποθηκευμένο στον κατάλογο. Επίσης, παρέχει μεθόδους για τον υπολογισμό της ομοιότητας εγγράφου-ερωτήματος. Συγκεκριμένα υλοποιεί σε ξεχωριστές συναρτήσεις τις μεθόδους Ευκλείδειας Απόστασης, Εσωτερικού γινομένου, Συνημίτονου Γωνίας, Dice και Jaccard. Σε αυτές τις συναρτήσεις υλοποιούνται οι μαθηματικές πράξεις για τον υπολογισμό της ομοιότητας εγγράφου ερωτήματος. Να σημειωθεί εδώ ότι δεν δημιουργούμε το διάνυσμα βαρών για κάθε έγγραφο.

Ο τρόπος με τον οποίο υπολογίζονται τα αποτελέσματα είναι ο εξής: Για κάθε όρο του ερωτήματος, υπολογίζουμε το βάρος του για το ερώτημα και το βάρος του για το έγγραφο (για το οποίο θέλουμε να βρούμε τον βαθμό ομοιότητας). Οι μαθηματικές πράξεις γίνονται για τα αντίστοιχα βάρη του κάθε όρου του ερωτήματος. Επίσης ελέγχεται αν το βάρος του όρου για το έγγραφο είναι μηδέν. Σε αυτή την περίπτωση επιστρέφεται μηδενική τιμή, αλλιώς επιστρέφεται μία αριθμητική τιμή που δηλώνει την ομοιότητα εγγράφου-ερωτήματος.

Διεπαφή Χρήστη-Συστήματος (UIM)

Στο επίπεδο αυτό, δίνεται η δυνατότητα στο χρήστη να χρησιμοποιήσει τις δυνατότητες του συστήματος. Μέσα από ένα παραθυρικό περιβάλλον, ο χρήστης μπορεί να διαχειριστεί συλλογές, να αναζητήσει για έγγραφα χρησιμοποιώντας το Boolean ή το Vector μοντέλο, να χρησιμοποιήσει αρχεία ερωτημάτων και να πάρει πολλαπλά αποτελέσματα, να επιλέξει ποια συνάρτηση θέλει να χρησιμοποιηθεί για την εκτέλεση ερωτημάτων, κ.ά. Το περιβάλλον αποτελείται από το menu bar, την επιφάνεια εργασίας και το status bar.

Από το menu bar, ο χρήστης μπορεί να διαχειριστεί τους καταλόγους και να ορίσει πρόσθετες ρυθμίσεις. Του δίνεται η δυνατότητα να δημιουργήσει έναν καινούργιο ανεστραμμένο κατάλογο επιλέγοντας το αρχείο από το οποίο θα δημιουργηθεί. Αφού ο κατάλογος έχει δημιουργηθεί, μπορεί να προσθέσει ή να αφαιρέσει αρχεία από αυτόν. Το σύστημα καθοδηγεί το χρήστη για το τι μπορεί να κάνει την τρέχουσα στιγμή και τι όχι.

Το status bar, ενημερώνει κάθε στιγμή το χρήστη για την κατάσταση του συστήματος. Όταν ο χρήστης τρέχει ερωτήματα, τον ενημερώνει για το χρόνο εκτέλεσής τους και σε περίπτωση πολλαπλών ερωτημάτων από αρχείο, τον ενημερώνει για το συνολικό χρόνο.

Η επιφάνεια εργασίας αποτελείται από τρία κομμάτια:

- Το **πάνελ του ερωτήματος**, όπου ο χρήστης μπορεί να θέσει ένα ή πολλαπλά ερώτημα στο σύστημα και να θέσει κάποιες παραμέτρους γι αυτά.
- Το **πάνελ των αποτελεσμάτων**, όπου εκεί εμφανίζονται τα αποτελέσματα της αναζήτησης.
- Το **πάνελ της προβολής του περιεχομένου**, όπου εκεί φαίνεται το περιεχόμενο του εγγράφου που επέλεξε ο χρήστης.

Επεκτάσεις

Υλοποιήθηκαν δύο επεκτάσεις.

Η πρώτη είναι η αφαίρεση των όρων που απαντώνται σε ποσοστό εγγράφων μεγαλύτερο του 80%. Έχει υλοποιηθεί σε μία ξεχωριστή συνάρτηση της κλάσης Index. Αυτή η αφαίρεση των όρων γίνεται κάθε φορά που αποθηκεύεται ο κατάλογος εφόσον έχει προηγουμένως τροποποιηθεί. Περισσότερες πληροφορίες γι αυτή την επέκταση θα βρείτε στην ενότητα «*Αφαίρεση συχνών όρων*».

Η δεύτερη είναι η υλοποίηση παραθυρικού περιβάλλοντος η οποία περιγράφεται στην ενότητα «*Διεπαφή Χρήστη-Συστήματος* (*UIM*)».

Περιγραφή του τρόπου σκέψης/υλοποίησης κάποιων λειτουργιών

Κατά τη δημιουργία του καταλόγου εισάγονται όλες οι λέξεις και ενημερώνεται η τιμή της μέγιστης συχνότητας του πιο συχνά εμφανιζόμενου όρου σε κάθε έγγραφο, καθώς και οι τιμές των βαρών κάθε όρου για κάθε έγγραφο. Επίσης μετά τη δημιουργία, ο κατάλογος αποθηκεύεται και κλείνει. Στη φάση της αποθήκευσης γίνεται έλεγχος και αφαιρούνται από αυτόν οι λέξεις που απαντώνται σε ποσοστό εγγράφων μεγαλύτερο του 80%.

Κατά την εισαγωγή και διαγραφή εγγράφου του καταλόγου ενημερώνονται εκ νέου τα βάρη και οι μέγιστες συχνότητες.

Η εφαρμογή δημιουργεί και χρησιμοποιεί δομές, οι οποίες περιγράφονται παραπάνω, που σκοπό έχουν την ταχύτερη εκτέλεση υπολογισμών κάποιων μεγεθών και κατ' επέκταση των ερωτημάτων. Από τα έγγραφα που επιστρέφει το κάθε ερώτημα, έχουν αποκλειστεί όσα είχανε μηδενική βαθμολογία.

Ιδιαιτερότητες και σημεία που χρίζουν προσοχής!

- Όλα τα αρχεία που χειρίζεται η εφαρμογή πρέπει να βρίσκονται μέσα σε φάκελο με όνομα “Collections”. Μέσα σε αυτό τον φάκελο δημιουργείται από την εφαρμογή ο φάκελος “INDEXES” στον οποίο αποθηκεύονται οι ανεστραμμένοι κατάλογοι που δημιουργεί η εφαρμογή μας.

- Η δημιουργία ενός καταλόγου γίνεται από το αρχείο “*_docs.txt”. Δεν υποστηρίζεται δημιουργία καταλόγου χωρίς αυτό το αρχείο. Έπειτα από τη δημιουργία του καταλόγου, μπορούν να προστεθούν κι άλλα αρχεία ή να αφαιρεθούν κάποια άλλα. Για την σωστή λειτουργία του συστήματος, θα ήταν προτιμότερο τα επιπλέον αρχεία να μεταφερθούν στο φάκελο “DOCS” που δημιουργεί το σύστημα για την κάθε συλλογή.

- Γίνεται έλεγχος ώστε να μην μπορούν να προστεθούν αρχεία με κατάληξη διαφορετική από “.txt”. Επίσης το αρχείο με τα ερωτήματα αναφοράς πρέπει να είναι της μορφής “*_queries.txt”.

- Οι συλλογές πρέπει να έχουν τη μορφή φακέλου με το όνομά τους και να υπάρχουν στο φάκελο “Collections”, για παράδειγμα “Collections/MED/”. Σε περίπτωση που θέλετε να εκτελέσετε πολλαπλά ερωτήματα, μέσα στο φάκελο της συλλογής πρέπει να υπάρχει το αρχείο αποτελεσμάτων “*_relevant.txt”.

- Η δημιουργία των καταλόγων για τις συλλογές αναφοράς διήρκεσε περίπου επτά λεπτά.

Οδηγίες Χρήσης

Στο αρχείο ZIP του project περιέχονται τρία αρχεία: “ARES_SRC”, “ARES” και “Documentation.pdf”.

Ο πρώτος φάκελος είναι ο πηγαίος κώδικας της JAVA εφαρμογής, ο οποίος έχει υλοποιηθεί στο NetBeans.

Ο δεύτερος φάκελος περιέχει το εκτελέσιμο JAR αρχείο και τον φάκελο “Collections” ο οποίος περιέχει τους φακέλους “MED”, “CRAN” και “INDEXES”. Οι δύο πρώτοι φάκελοι είναι ένα παράδειγμα του πως πρέπει να είναι οι φάκελοι των συλλογών. Περιέχουν τα αρχεία δημιουργίας καταλόγων, εκτέλεσης πολλαπλών ερωτημάτων και αποτελεσμάτων καθώς και τον φάκελο “DOCS” όπου υπάρχουν τα έγγραφα της συλλογής. Στον τρίτο φάκελο περιέχονται οι δύο κατάλογοι “MED.idx” και “CRAN.idx” ενδεικτικά, γιατί η δημιουργία καταλόγων είναι χρονοβόρα διαδικασία με το συγκεκριμένο σύστημα (δόθηκε βάση στη γρήγορη υλοποίηση των ερωτημάτων αντί της γρήγορης δημιουργίας των καταλόγων).

Το αρχείο είναι η τεκμηρίωση (το παρών έγγραφο) όπου φαίνεται ο τρόπος σκέψης κατά την υλοποίηση του συστήματος.

Για να εκτελεστεί σωστά το πρόγραμμα, πρέπει ο φάκελος “ARES” να μην τροποποιηθεί. Η εξαγωγή του εκτελέσιμου αρχείου από το φάκελο αυτό ή η διαγραφή κάποιου άλλου αρχείου μπορεί να οδηγήσει στη μη σωστή λειτουργία του συστήματος. Παρόλα αυτά ο φάκελος μπορεί να μεταφερθεί στο προορισμό που επιθυμεί ο χρήστης.

Κάποιες λειτουργίες του συστήματος αργούν να εκτελεστούν και κατά την αναμονή δεν υπάρχει κάποια ένδειξη ότι το σύστημα δουλεύει. Όταν εκτελεστεί η εργασία, θα εμφανιστεί κατάλληλο μήνυμα στο status bar που θα επιβεβαιώσει ότι η λειτουργία εκτελέστηκε επιτυχώς ή όχι.