Αναφορά εφαρμογής Lucene Searcher

Αλκιβιάδης Πετρόχειλος (p3130171)

Δυσκολίες

Το πρώτο μέρος της εργασίας υλοποιήθηκε σε Python. Για να τηρηθεί ομοιογένεια με το πρώτο μέρος, έγινε χρήση της PyLucene. Ωστόσο η συγκεκριμένη βιβλιοθήκη δεν έχει διαθέσιμα δυαδικά αρχεία, παρά μόνο πηγαίο κώδικα.

Γίνανε διάφοροι πειραματισμοί προκειμένου να παραχθούν δυαδικά αρχεία της PyLucene σε Windows 7, πράγμα που απαιτούσε πολλές ανορθόδοξες δοκιμές με makeFile αρχεία.

Αφού απέτυχε η δημιουργία δυαδικών αρχείων σε Windows 7, έγινε μεταφορά του πρότζεκτ σε Linux Ubuntu. Η δημιουργία δυαδικών αρχείων κατέστη εφικτή, ωστόσο δεν υπήρχε επαρκής βιβλιογραφία προκειμένου να εκτελεστεί ορθά η PyLucene.

Τελικά έγινε αλλαγή της γλώσσας προγραμματισμού του δεύτερου μέρους της εργασίας σε Java, με την οποία δεν προέκυψαν παρόμοια προβλήματα.

Υλοποίηση

Η εφαρμογή (με τίτλο Lucene Searcher) αποτελείται από δύο ενωμένα μέρη. Το πρώτο είναι το λογικό κομμάτι της εφαρμογής (κλάση SearchEngine.java), ενώ το δεύτερο είναι το γραφικό κομμάτι της (κλάση GUI.java).

Η εφαρμογή ανοίγει όλα τα αρχεία της προκαθορισμένης διαδρομής (βλέπε μέρος «Προκαθορισμένα αρχεία και διαδρομές») που τελειώνουν σε ".xml" κατάληξη, καθώς και το αρχείο με τα ερωτήματα του χρήστη.

Για κάθε ένα αρχείο .xml της βάσης εντοπίζει τα πεδία "title" (τίτλος), "objective" (λεπτομέρειες), "identifier" (κατηγορία) και "rcn" (μοναδικό αναγνωριστικό) και τα βάζει σε ένα ευρετήριο πολλαπλών πεδίων.

Μόλις ολοκληρωθεί η ευρετηρίαση των αρχείων της βάσης (αρχεία .xml), η εφαρμογή είναι έτοιμη για να ξεκινήσει την αναζήτηση με βάση τα ερωτήματα που υπάρχουν στο αρχείο "Queries.xml".

Η εφαρμογή έχει προκαθορισμένα βάρη για κάθε ένα από τα σχετικά πεδία ("title", "objective", "identifier"), τα οποία μπορούν να αλλάξουν από τη γραφική διεπαφή της εφαρμογής.

Είτε με τη χρήση των προκαθορισμένων βαρών και διαδρομών για το "Queries.xml", είτε με την προσαρμογή τους μέσω της διεπαφής, η εφαρμογή εκκινεί την αναζήτηση με το πάτημα του κουμπιού "SEARCH" της γραφικής διεπαφής.

Σε αυτό το στάδιο η εφαρμογή διαβάζει το αρχείο των ερωτημάτων προς τη βάση και για κάθε ένα από αυτά αποθηκεύει τα πεδία προς αναζήτηση. Αφού έχει διαβάσει όλα τα απαραίτητα πεδία του αρχείου ερωτημάτων, αναζητάει τα 20 πιο ταιριαστά αρχεία της βάσης για τα πεδία "title" και "objective". Αυτό σημαίνει ότι κάνει δύο ξεχωριστές αναζητήσεις με τα 20 καλύτερα αποτελέσματα για κάθε ένα πεδίο ξεχωριστά. Τα σχετικά αποτελέσματα αποθηκεύονται σε δύο λίστες.

Στη συνέχεια οι παραπάνω δύο λίστες ενσωματώνονται σε μια τελική λίστα (χωρίς διπλότυπα). Για κάθε ένα στοιχείο της τελικής λίστας γίνεται έλεγχος για το αν η κατηγορία του ερωτήματος (call identifier) ταυτίζεται με την κατηγορία του στοιχείου της τελικής λίστας (boolean μεταβλητή identifierHit). Σε περίπτωση που οι κατηγορίες είναι ίδιες, τίθεται σε ενεργή κατάσταση ο υπολογισμός ενός μπόνους συνάφειας (προκαθορισμένο βάρος 30) που προστίθεται στο τελικό σκορ. Αν οι κατηγορίες δεν είναι ίδιες, τότε δεν υπολογίζεται κανένα μπόνους συνάφειας. Το μπόνους μπορεί να αλλάξει εύκολα από τη γραφική διεπαφή.

Αφού ολοκληρωθεί η φάση της δημιουργίας της τελικής λίστας και ο έλεγχος συνάφειας κατηγορίας, η εφαρμογή υπολογίζει την τελική βαθμολογία του κάθε στοιχείου της τελικής λίστας.

Ο τύπος που χρησιμοποιείται είναι ο ακόλουθος:

→ Score = w(title) * score(title) + w(objective) * score(objective) + w(call) * identifierHit

• Το w()αντιπροσωπεύει το βάρος του στοιχείου εντός των παρενθέσεων,

• Το score() να αντιπροσωπεύει το βαθμό ομοιότητας του στοιχείου εντός των παρενθέσεων,

• Το identifierHit είναι μια δυαδική μεταβλητή που αντιπροσωπεύει το αν οι κατηγορία του αντικειμένου ταυτίζεται με την κατηγορία του ερωτήματος.

Αφού ολοκληρωθεί η βαθμολόγηση όλων των στοιχείων της τελικής λίστας, γίνεται ταξινόμηση τους με βάση το τελικό τους σκορ.

Τα 20 αποτελέσματα με το υψηλότερο σκορ γράφονται στο αρχείο με τίτλο ”@RESULTS.txt”. Το εν λόγω αρχείο περιέχει στοιχεία όπως:

• βάρη της αναζήτησης

• αποτελέσματα με βάση το τελικό σκορ

• τίτλος, περιγραφή και σκορ κάθε τελικού αποτελέσματος

Γραφική διεπαφή

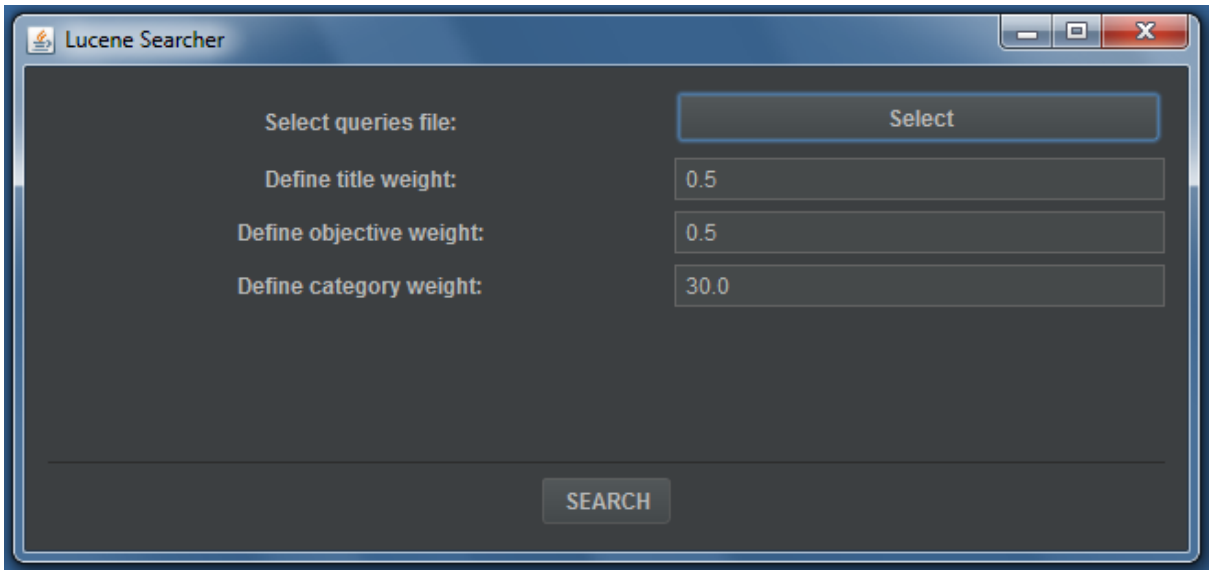
Κατά την εκτέλεσή της η εφαρμογή εμφανίζει μια γραφική διεπαφή από την οποία ο χρήστης μπορεί να επιλέξει τα βάρη που θα έχουν οι όροι της αναζήτησης, το βάρος της συνάφειας κατηγορίας, καθώς και το αρχείο με τα ερωτήματα προς τη βάση (αν θέλει να αλλάξει τα προκαθορισμένα).

Όλα τα διαθέσιμα πεδία έχουν ήδη προκαθορισμένες τιμές μέσα στην εφαρμογή και δεν είναι απαραίτητος ο ορισμός τους από τη γραφική διεπαφή (βλέπε «Παραδείγματα εκτέλεσης»).

Παραδείγματα εκτέλεσης

1) Παράδειγμα εκτέλεσης με προκαθορισμένες τιμές:

Εκτελούμε την εφαρμογή πατώντας το αρχείο SearchEngine.jar και εμφανίζεται η αρχική παρακάτω γραφική διεπαφή.



(Αρχική οθόνη)

Περιμένουμε λίγα δευτερόλεπτα για να ολοκληρωθεί το indexing και πατάμε το κουμπί “SEARCH”.

Η εφαρμογή θα πραγματοποιήσει την αναζήτηση με τα προκαθορισμένα βάρη, καθώς και με βάση το προκαθορισμένο αρχείο “Queries.xml”.

Τα αποτελέσματα θα αποθηκευτούν στο αρχείο “@RESULTS.txt”.

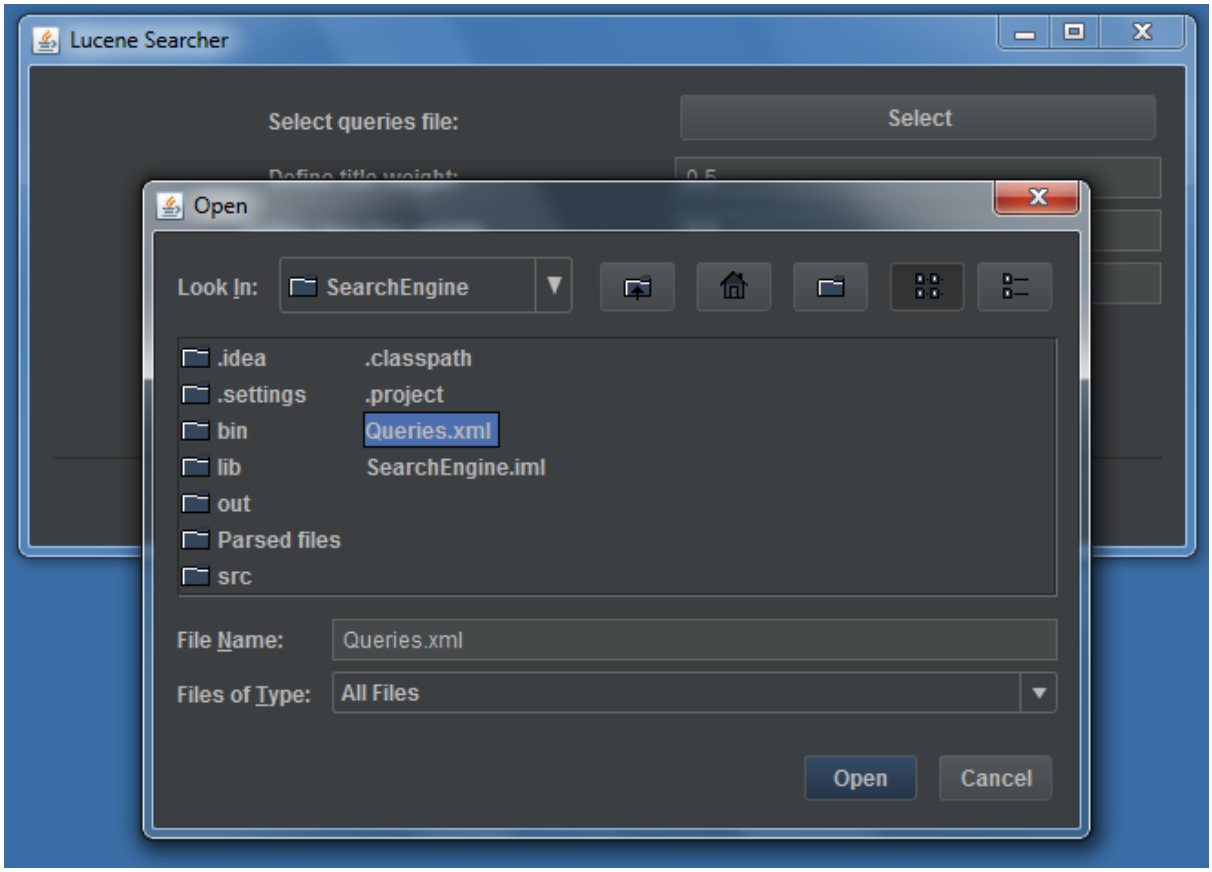
@RESULTS.txt Queries.xml

```
Query weights:
Title weight --> 0.5
Objective weight --> 0.5
Category weight --> 30.0

**********
QUERY 1
**********
-----------------------------------------
(1)
*SCORE --> 197.86252
*TITLE --> Support towards the Europe PMC initiative-Contribution for 2014-2016
*ABSTRACT --> "The proposed action will provide continued support to the European Research Council (ERC) in the
implementation of its Open Access strategy for projects funded in the Life Sciences domain. It follows on from the
project ""Support towards the Europe PMC initiative-Contribution for 2013""(ERC-EuropePMC-SUP-2013) which has allowed
the ERC to offer the benefits of Europe PMC to its funded researchers for the first time in 2013. The ERC Open Access
strategy, and how the present project will assist the ERC in its implementation, is explained below"

-----------------------------------------
(2)
*SCORE --> 170.91873
*TITLE --> Support to the Europe PMC initiative – Co-funding grant for the 2016-2021 period
*ABSTRACT --> The proposed coordination and support action will provide continued support to the European Research
Council (ERC) in the implementation of its Open Access strategy for projects funded in the Life Sciences domain. It
follows on from the projects “Support towards the Europe PMC initiative – Contribution for 2014-2016”
(ERC-EuropePMC-1-2014) and “Support towards the Europe PMC initiative – Contribution for 2013”
(ERC-EuropePMC-SUP-2013) which allowed the ERC to offer the benefits of Europe PubMed Central (Europe PMC
http://europepmc.org/) to its funded researchers for the first time in 2013.

The proposed project concerns the giving of a grant to EMBL-EBI for the purpose of ensuring the further development
and maintenance of Europe PMC, beyond the work done as part of ERC-EuropePMC-SUP-2013 and ERC-EuropePMC-1-2014, for
the period 1st April 2016 – 31st March 2021 (60 months). Support towards this project is requested from the ERC on a
co-funding basis.

-----------------------------------------
(3)
*SCORE --> 76.52574
*TITLE --> Support to the Vice-Presidents of the ERC Scientific Council 2014
*ABSTRACT --> The proposed Action will provide the necessary support to the Vice-Presidents of the European Research
```

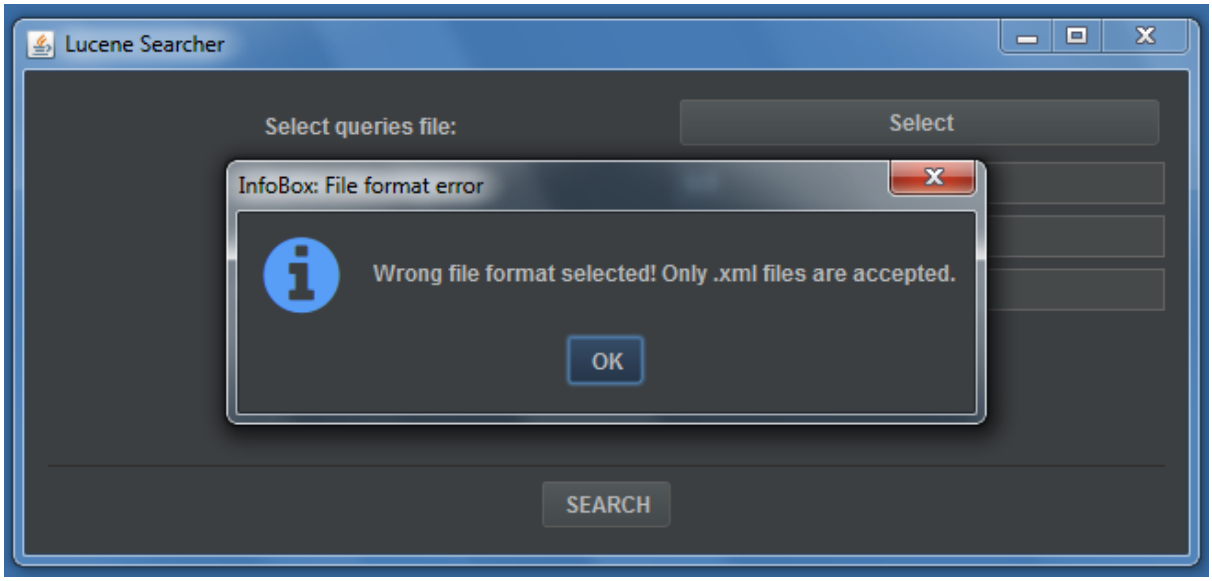
(Ενδεικτικά αποτελέσματα)

2) Παράδειγμα εκτέλεσης με μη προκαθορισμένες τιμές & μηνύματα εφαρμογής:

Αφού εμφανιστεί η αρχική οθόνη της εφαρμογής, πατάμε το κουμπί “Select”. Στο καινούργιο παράθυρο που θα εμφανιστεί, επιλέγουμε το .xml αρχείο που περιέχει τα ερωτήματα προς τη βάση. Σε περίπτωση επιλογής αρχείου με μη έγκυρη κατάληξη θα εμφανιστεί κατάλληλο μήνυμα.

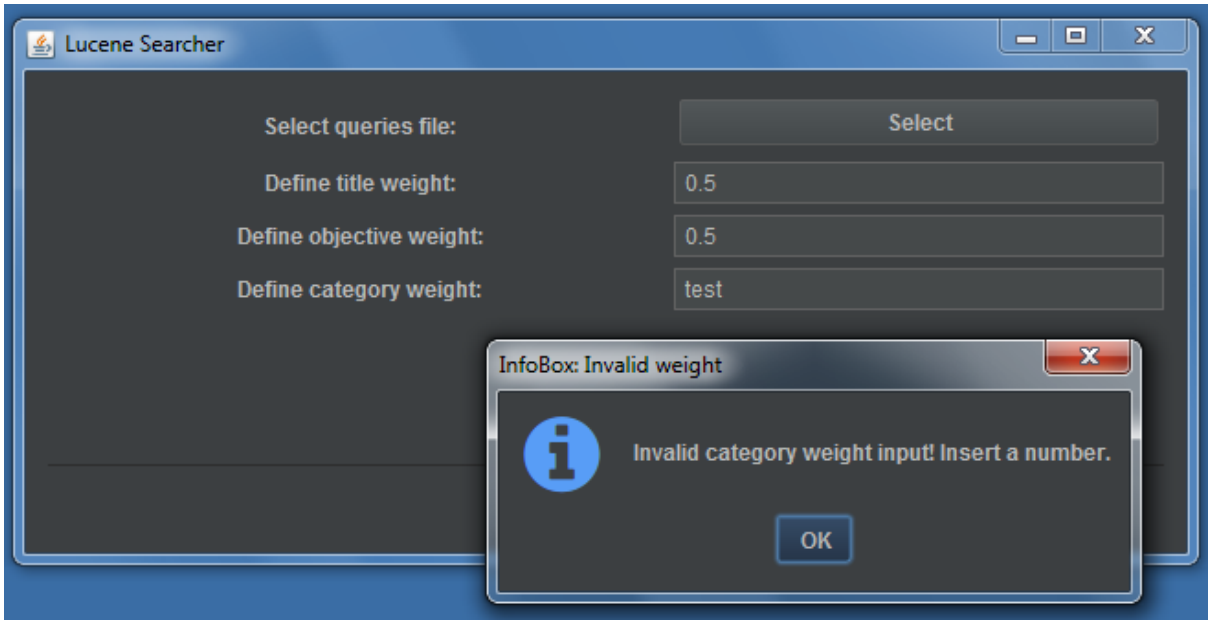


(Προαιρετική επιλογή αρχείου ερωτημάτων)



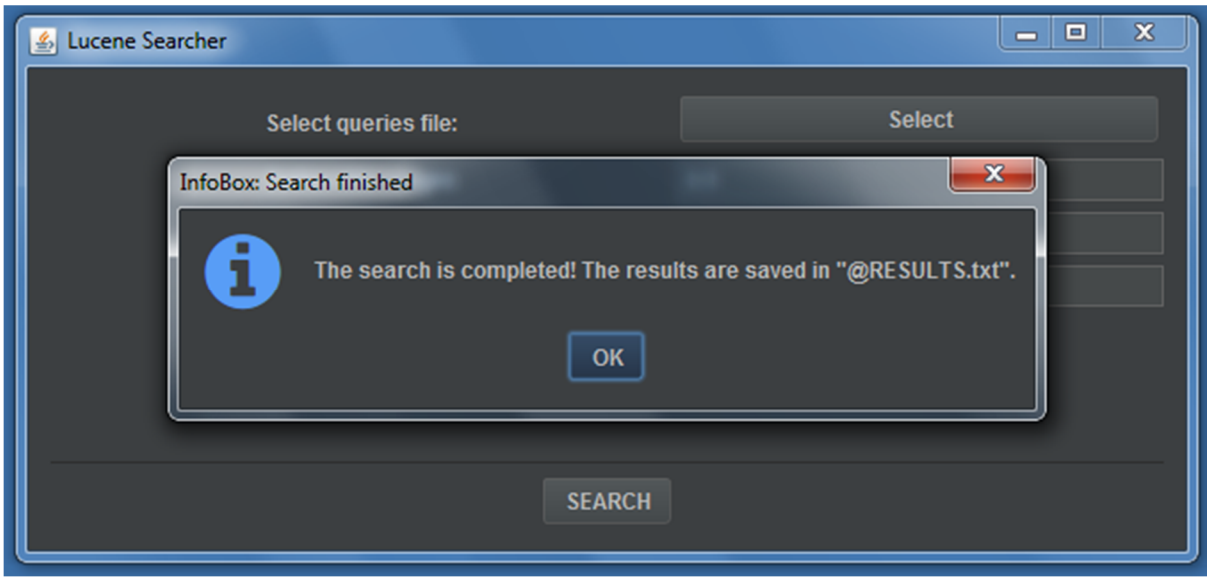
(Μήνυμα επιλογής μη έγκυρου αρχείου ερωτημάτων)

Στη συνέχεια μπορούμε να τροποποιήσουμε τα βάρη που μας ενδιαφέρουν, απλώς γράφοντας στο αντίστοιχο πεδίο τον επιθυμητό αριθμό. Σε περίπτωση μη αποδεκτού βάρους εμφανίζεται σχετικό μήνυμα.



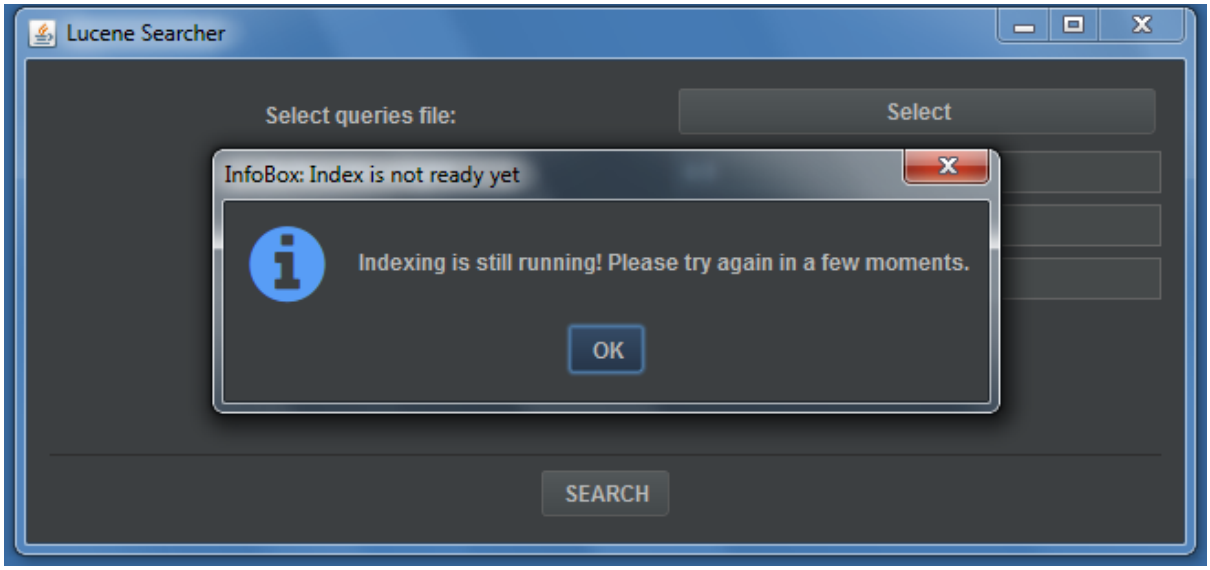
(Μήνυμα κατά την εισαγωγή μη αποδεκτού βάρους)

Αφού ολοκληρώσουμε τα παραπάνω βήματα, πατάμε το κουμπί “SEARCH”. Η αναζήτηση θα πραγματοποιηθεί και τα αποτελέσματα θα εκτυπωθούν στο αρχείο ”@RESULTS.txt”. Μόλις ολοκληρωθεί η αναζήτηση, θα εμφανιστεί το κατάλληλο μήνυμα στο χρήστη.



(Μήνυμα επιτυχούς ολοκλήρωσης της αναζήτησης)

Σε περίπτωση που η ευρετηρίαση της βάσης δεν έχει ολοκληρωθεί, θα εμφανιστεί σχετικό μήνυμα. Πρέπει απλώς να περιμένουμε μερικά δευτερόλεπτα και να ξαναπατήσουμε το κουμπί “SEARCH” για να πραγματοποιηθεί η αναζήτηση.



(Μήνυμα σε περίπτωση που δεν έχει ολοκληρωθεί το indexing)

Παρατηρήσεις

• Το γραφικό θέμα που χρησιμοποιήθηκε για τη γραφική διεπαφή της εφαρμογής είναι το “Darcula” (https://github.com/bulenkov/Darcula).

• Μέρος της γραφικής διεπαφής της εφαρμογής δημιουργήθηκε με το πρόσθετο JFormDesigner (https://www.formdev.com/).

• Η κονσόλα (command line) προσφέρει επιπρόσθετες λεπτομέρειες κατά την εκτέλεση της εφαρμογής, όπως τη λίστα και τις βαθμολογίες με όλα τα βαθμολογημένα αρχεία (και όχι μόνο τα 20 καλύτερα), το μέγεθος της τελικής λίστας κ.α. (μη απαραίτητα για τον χρήστη).

• Μερικά στοιχεία της τελικής λίστας δεν ανήκουν και στις δύο αρχικές λίστες. Σε αυτή την περίπτωση το μερικό σκορ του πεδίου της λίστας στην οποία δεν ανήκει το στοιχείο, παραβλέπεται καθώς δεν υπάρχει.

Προκαθορισμένα αρχεία και διαδρομές

Η εφαρμογή εντοπίζει σε ποιον φάκελο βρίσκεται το αρχείο .jar. Στον ίδιο φάκελο με το αρχείο SearchEngine.jar θα πρέπει να υπάρχει και ο φάκελος “Parsed files” όπως έχει προκύψει από την εκτέλεση του πρώτου μέρους της εργασίας, ο οποίος περιέχει τα .xml αρχεία προς ευρετηρίαση.

Δεν είναι δυνατή η αλλαγή της τοποθεσίας της βάσης με τα αρχεία προς ευρετηρίαση από τη γραφική διεπαφή της εφαρμογής. Ο λόγος είναι καθαρά για ταχύτητα. Μόλις ανοίγει η γραφική διεπαφή το πρόγραμμα κάνει indexing τη βάση (χρειάζεται μερικά δευτερόλεπτα για τα 18.316 αρχεία της βάσης). Αν υπήρχε επιλογή της βάσης από άλλο φάκελο, η εφαρμογή δεν θα μπορούσε να εκμεταλλευτεί το χρόνο που χρειάζεται ο χρήστης για να ορίσει τις υπόλοιπες παραμέτρους της εφαρμογής.

Η εφαρμογή περιμένει να βρει στον φάκελο του SearchEngine.jar ένα αρχείο με το όνομα “Queries.xml” το οποίο περιέχει τα ερωτήματα προς τη βάση. Είναι ωστόσο επιτρεπτή η επιλογή ενός διαφορετικού αρχείου μέσω της γραφικής διεπαφής.

Τα αποτελέσματα της αναζήτησης εξάγονται στον ίδιο φάκελο με το SearchEngine.jar στο αρχείο με τίτλο ”@RESULTS.txt”. Σε κάθε αναζήτηση τα νέα αποτελέσματα επικαλύπτουν τυχόν αποτελέσματα προηγούμενης αναζήτησης.