Εργασία Ανάκτησης Πληροφορίας

Στην εργασία θα χρησιμοποιηθεί η βιβλιοθήκη Lucene που είναι γραμμένη σε Java.

Η Lucene σας επιτρέπει να προσθέσετε, εύκολα, αναζήτηση σε οποιαδήποτε εφαρμογή. Τα τελευταία χρόνια η Lucene έχει γίνει εξαιρετικά δημοφιλής και είναι η πλέον ευρέως χρησιμοποιούμενη βιβλιοθήκη ανάκτησης πληροφοριών. Δίνει τη δυνατότητα αναζήτησης πίσω από ιστοσελίδες και εφαρμογές γραφείου. Παρόλο που είναι γραμμένη σε Java, χάρη στη δημοτικότητά της αλλά και την αποφασιστικότητα των απαιτητικών προγραμματιστών, υπάρχει διαθέσιμη σε αρκετές θύρες αλλά και ενσωματωμένη σε άλλες γλώσσες προγραμματισμού (C / C ++, C #, Ruby, Perl, Python, PHP κ.λπ.). Εσείς μπορείτε να επιλέξετε την αρχική έκδοση (σε Java) αλλά αν το επιθυμείτε μπορείτε να χρησιμοποιήσετε το PyLucene (επέκταση της Python).

Βοήθεια για τη Lucene μπορείτε να βρείτε:

- στο βιβλίο Lucene in Action [1] που θα βρείτε στο eclass στην ενότητα ΕΡΓΑΣΤΗΡΙΟ/Έγγραφα.
- https://www.tutorialspoint.com/lucene/lucene_environment.htm
- http://www.lucenetutorial.com/index.html

Την εργασία είναι ατομική

Παραδοτέα: *Αναφορά σε μορφή* ***.pdf*** *με όνομα αρχείου* **`Surname1 AM1-Surname2 AM2.pdf`** *και κώδικας σε μορφή* ***.py****. Το* ***σύνολο*** *των αρχείων συμπιεσμένα (μόνο zip και rar) με όνομα αρχείου* **`Surname1 AM1-Surname2 AM2.zip`**

Ημ/νία κατάθεσης: *σύμφωνα με τις ημ/νίες που φαίνονται στο eclass για την κάθε ομάδα. Δεν υπάρχει δυνατότητα εκπρόθεσμης υποβολής ούτε υποβολής σε άλλη ομάδα.*

Βήμα 1 Γνωριμία με τη Lucene

Μελετήστε το Chapter 1 "Meet Lucene"του βιβλίου Lucene in Action [1] και απαντήστε στις παρακάτω ερωτήσεις.

Ερώτηση 1: Περιγράψτε συνοπτικά τι είναι η Lucene και τι μπορεί να κάνει.

Ερώτηση 2: Αναφέρατε ποια είναι τα στοιχεία-συνιστώσες που απαρτίζουν μία εφαρμογή αναζήτησης και ποια από αυτά μπορεί να υλοποιήσει η Lucene. **Περιγράψτε συνοπτικά** καθένα από τα στοιχεία-συνιστώσες που υλοποιεί η Lucene.

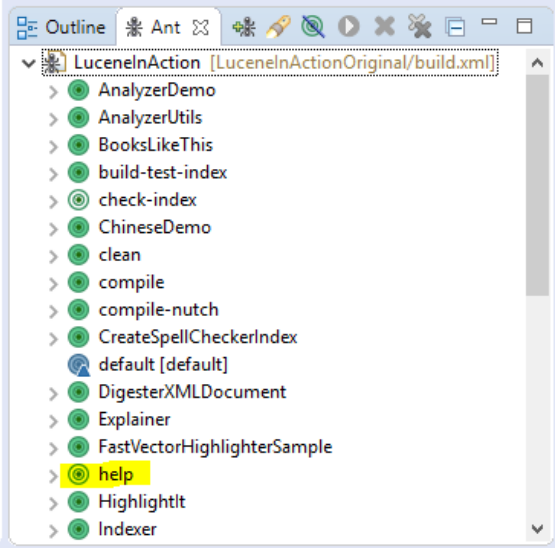
Βήμα 2 Απαιτήσεις συστήματος- Εγκατάσταση

Για την εγκατάσταση της Lucene θα χρειαστεί να έχετε προεγκατεστημένα:

- Το Java Development Kit (JDK)
- Ένα Java IDE όπως π.χ. το Eclipse

Κατεβάστε το Ant project LuceneInAction source code και αποσυμπιέστε το. Εάν θέλετε να το φορτώσετε στο Eclipse κάντε New Project→Java→Java Project from Existing Ant Buildfile.

Από το menu επιλέξτε Window→Show View→Ant και κάντε Add Buildfile το build.xml του LuceneInAction. Θα εμφανιστεί η παρακάτω καρτέλα.



Διαβάστε το help και τρέξτε όλα τα buildfiles ώστε να δείτε την λειτουργία τους.

Ερώτηση 3: Τρέξτε την **Indexer** και περιγράψτε πως γίνεται το indexing παρουσιάζοντας εικόνα με τα indexed αρχεία.

Ερώτηση 4: Τρέξτε την **Searcher** και περιγράψτε πως γίνεται το searching παρουσιάζοντας αποτελέσματα από 4 χαρακτηριστικά ερωτήματα (queries) που κάνατε.

Ερώτηση 5: Τρέξτε το SortingExample και περιγράψτε τη λειτουργία της

Βήμα 3 Μηχανή αναζήτησης σε Python

Υλοποιήστε μία μηχανή αναζήτησης σε Python που να υλοποιεί indexing, searching και ranking με δύο τουλάχιστον αλγόριθμους π.χ. BM25 και TF-IDF. Στο διαδίκτυο μπορείτε να βρείτε **διάφορα** παραδείγματα όπως:

How to build a search engine και How to build a smart search engine (Part II)

Ακόμα και με διεπαφή όπως στο Building a search engine with Python, Tornado and Strus

Μπορείτε επίσης, να χρησιμοποιήσετε και ένα οποιοδήποτε άλλο. **Σε κάθε περίπτωση** θα πρέπει να αναφέρατε την **πηγή ή τις πηγές** (αν κάνατε συνδυασμό) του κώδικα που χρησιμοποιήσατε.

Ερώτηση 6: Παρουσιάστε τον κώδικα και τυχόν αλλαγές που κάνατε.

Ερώτηση 7: Φορτώστε κάποια documents π.χ. από το Wikipedia abstracts, project gutenberg (για τα οποία πρέπει να περιγράψετε τι περιέχουν) και τρέξτε τουλάχιστον 5 ερωτήματα (queries). Παρουσιάστε και σχολιάστε τα αποτελέσματα για τον κάθε αλγόριθμο.

Ερώτηση 8: Περιγράψτε τις διαφορές στη λειτουργία των δύο διαφορετικών προσεγγίσεων (του προγράμματος σε python και του αντίστοιχου στη Lucene).

Ερώτηση 9: Πως πιστεύετε ότι μπορεί να βελτιωθεί το καθένα από τα δύο προγράμματα (υλοποιημένα σε Python και σε Lucene). Ποιες δυνατότητες ή/και αλγόριθμοι θα βελτίωναν το κάθε project.

Αναφορές-Βιβλιογραφία

[1] Erik Hatcher and Otis Gospodnetic. 2005. *Lucene in Action（SECOND EDITION）*. Retrieved from http://www.imamu.edu.sa/dcontent/IT_Topics/java/luceneinaction.pdf