Programming Project 1 – Lucene

[Ολα τα αρχεία που αναφέρονται παρακάτω βρίσκονται στον κατάλογο "Using Lucene" στα Εγγραφα της Διάλεξης 06]

Σας δίνεται το αρχείο data.zip που περιέχει περίπου 100000 άρθρα από το PubMed Central. Κάθε γραμμή του αρχείου είναι ένα άρθρο σε μορφή JSON. Κάθε άρθρο έχει έναν τίτλο, μια περίληψη, έναν μοναδικό προσδιοριστή (το PubMed ID, ή PMID) και βιβλιογραφικά μεταδεδομένα (λίστα συγγραφέων, πληροφορία για το περιοδικό που δημοσιεύτηκε το άρθρο, κλπ).

Σας δίνεται ένα πολύ απλό αρχείο πηγαίου κώδικα java (IndexTester.java) που χρησιμοποιεί το Lucene και κατασκευάζει ένα απλό ανεστραμμένο ευρετήριο στην παραπάνω συλλογή άρθρων. Ευρετηριοποιεί μόνο τους τίτλους των άρθρων και μπορεί να διαχειριστεί πολύ απλά ερωτήματα. Οι αλλαγές που πρέπει να κάνετε και περιγράφονται παρακάτω, απαιτούν πολύ μικρές παρεμβάσεις στον κώδικα που σας δίνεται και αυτή είναι η αξία του Lucene. Το Lucene διαθέτει πολύ καλό documentation (https://lucene.apache.org/core/5_0_0/) από το οποίο θα πάρετε τις πληροφορίες που θα χρειαστείτε.

Πρέπει να τροποποιήσετε τον κώδικα ως εξής:

1. Ο κώδικας χρησιμοποιεί τον default StandardAnalyzer που μετατρέπει τα tokens σε μικρά γράμματα και αφαιρεί τα stopwords. Το Lucene διαθέτει έναν πιο εξελιγμένο αναλυτή, τον EnglishAnalyzer, ο οποίος κάνει και stemming, διαχειρίζεται τις γενικές καλύτερα, κλπ. Βρείτε πως θα αντικαταστήσετε τον StandardAnalyzer με τον EnglishAnalyzer (δείτε το documentation).
2. Ο κώδικας ευρετηριοποιεί μόνο τους τίτλους (title) των άρθρων. Τροποποιήστε τον κώδικα ώστε να ευρετηριοποιεί και τις περιλήψεις των άρθρων (abstract).
3. Ο κώδικας χρησιμοποιεί έναν απλό query parser, τον QueryParser. Το Lucene διαθέτει έναν πιο εξελιγμένο και ευέλικτο query parser, τον SimpleQueryParser. Βρείτε πως θα αντικαταστήσετε τον QueryParser με τον SimpleQueryParser και αξιοποιείστε τις δυνατότητές του (δείτε το documentation).

Ο κώδικας σας δίνεται σε ένα αρχείο code.zip. Αποσυμπιέστε το για να πάρετε τον κατάλογο code όπου θα υπάρχει ο κατάλογος lib που περιέχει το Lucene 5.0 και το αρχείο IndexTester.java. Συνεπώς, δεν χρειάζεται να κατεβάσετε το Lucene. Φυσικά, μπορείτε να χρησιμοποιήσετε το Eclipse ή όποιο άλλο IDE επιθυμείτε, αλλά επειδή πρόκειται για πολύ απλό project με ένα μόνο αρχείο πηγαίου κώδικα, σας δείχνω πως να δουλέψετε από τερματικό (δείτε το σχετικό video).

Υποθέτοντας ότι βρίσκεστε μέσα στον κατάλογο code, μεταγλωττίστε το αρχείο με τις παρακάτω εντολές:

```
$ export CLASSPATH=$CLASSPATH:./lib/lucene-core-5.0.0.jar:./lib/lucene-analyzers-common-5.0.0.jar:./lib/
lucene-queryparser-5.0.0.jar:./lib/jodd-3.6.5.jar
```

```
$ javac IndexTester.java
```

Τώρα μπορείτε να εκτελέσετε το αρχείο IndexTester το οποίο δέχεται τρία ορίσματα:

1. should_clear_index: αν θα δημιουργηθεί το ευρετήριο (1), ή θα χρησιμοποιηθεί το υπάρχον (0).
2. path_to_data: η διαδρομή στο corpus (αρχείο data.txt που σας δίνεται).
3. path_to_index: η διαδρομή στην οποία θα αποθηκευτεί το ευρετήριο (ή η διαδρομή σε ένα υπάρχον ευρετήριο όπου θα γίνει η αναζήτηση αν το should_clear_index είναι 0).

Για παράδειγμα, δώστε

```
$ java IndexTester 1 corpus/data.txt index/
```

για να δημιουργήσετε το ευρετήριο μέσα στον κατάλογο index που προηγουμένως δημιουργήσατε μέσα στον κατάλογο code (και αφού προηγουμένως αντιγράψατε το data.txt μέσα στον κατάλογο corpus) και

```
$ java IndexTester 0 corpus/data.txt index/
```

για να χρησιμοποιήσετε το υπάρχον ευρετήριο.

Θα παραδώσετε τον κώδικά σας με σχόλια και ένα κείμενο όπου θα περιγράφετε την εμπειρία σας με το Lucene και θα σχολιάζετε διάφορα queries που τρέξατε με τον κώδικά σας.

ΣΗΜΕΙΩΣΗ: Εννοείται πως άμα θέλετε να χρησιμοποιήσετε νεότερη έκδοση του Lucene μπορείτε να το κάνετε, έχετε όμως υπόψιν ότι ο κώδικας που σας δίνω μπορεί να μην τρέχει καθώς θα υπάρχουν διαφορές ανάμεσα στις εκδόσεις του Lucene – θα πρέπει να κάνετε αλλαγές στον κώδικα μελετώντας σχετικά tutorial και το documentation του Lucene.