**Συστήματα ανάκτησης πληροφοριών**

**4η Φάση Προγραμματιστικής Εργασίας**

**Βιβλιοθήκη LISA**

**Ον/μο:** ΒΙΤΑΛΗΣ ΙΩΑΝΝΗΣ

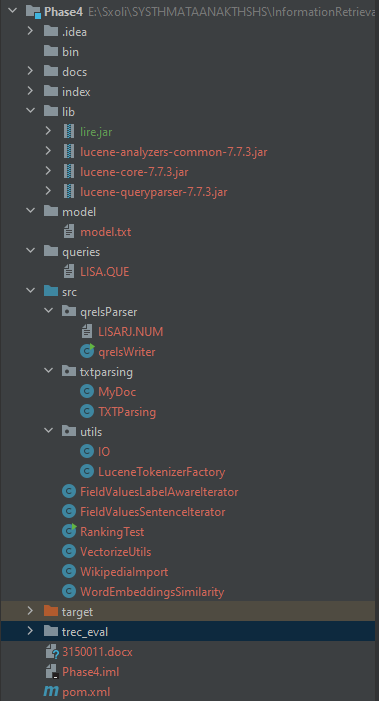
ΑΜ: 3150011

**Εγκατάσταση**

Για το setup του project χρησιμοποιήσα τον κώδικα που δώθηκε στην τελευταία διάλεξη αφού πραγματοποιήθηκε εγκατάσταση του Maven. Συγκεκριμένα:

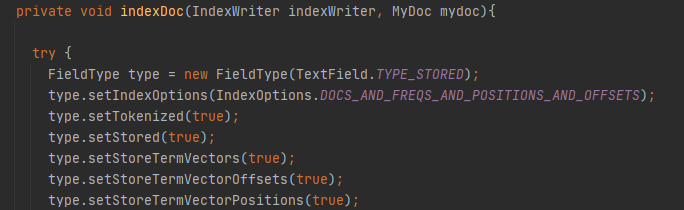
1. Εγκατάσταση Maven
2. Download τον κώδικα που μας δώθηκε και δημιουργία νέου project στο intellij
3. mvn clean install στο φάκελο του project
4. Έλεγχος οτι το project των διαλέξεων λειτουργεί
5. Προσθήκη των .jar άρχειων, lire.jar και των τριών που χρειάζονται για την lucene
6. Προσθήκη κώδικα/αρχείων όπου κρίνω απαραιτητο για την υλοποιήση της εργασίας

Η τελική μορφή της εργασίας με τα απαιραίτητα directories στο Intellij είναι η εξής:

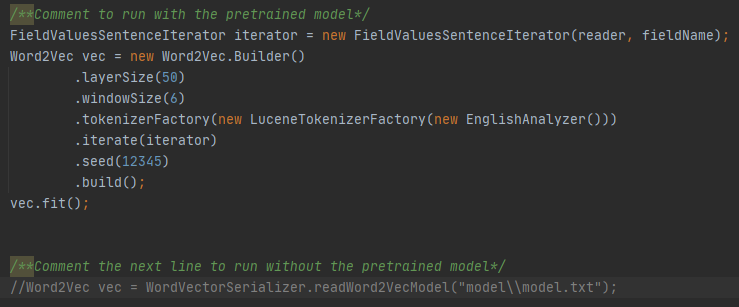
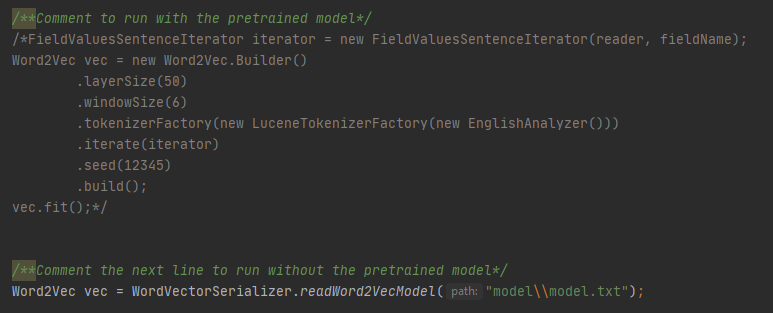


**Δομή Κώδικα**

Αρχικά χρησιμοποιώ κώδικα που υλοποίησα σε προηγούμενες φάσεις της εργασίας για την δημιουργία του Index με χρήση της Lucene με σημαντική αλλαγή στα πεδία του Fieldtype όπως μας δώθηκε και παρουσιαζεται εδώ:



Ο κώδικας της εργασίας με προεκπαιδευμένο μοντέλο και χωρίς είναι σχεδόν ο ίδιος για αυτό και δεν ήταν απαραίτητη η δημιουργία ενός δεύτερου project

* Για το πρώτο μέρος τις εργασίας στην κλάση RankingTest, το function testRanking() πρέπει να έχει την εξής μορφή: 
* Ενώ για το δεύτερο μέρος για να γίνει η φόρτωση του προεκπαιδευμένου μοντέλου πρέπει να έχει την εξής μορφή: 

Οι υπόλοιπες κλάσεις του κώδικα παραμένουν ίδιες με τις προηγούμενες φάσεις (TXTParsing, MyDoc, IO) ενώ οι υπόλοιπες έκρινα πως δεν χρειάζονται αλλαγές για την υλοποίηση της εργασίας και χρησιμοποιήθηκαν όπως μας δώθηκαν.

Η qrelsWriter στον φάκελο qrelsParser είναι ίδια με προηγούμενες φάσεις και χρησιμοποιείται για την δημιοργία του αρχείου LISAQRELS.test στον φάκελο trec\_eval που θα χρησιμοποιήσουμε με το trec\_eval για την αξιολόγηση.

**Λεπτομέρειες στον κώδικα**

Οι τιμές για την δημιουργία του Word2Vec (layerSize, windowSize) έγιναν μετά την δοκιμή διαφόρων τιμών και κράτησα τις καλύτερες, η τιμή του layer εξαρτάται από το μέγεθος των δεδομένων οπότε τιμές κάτω του 50 επέστρεφαν χειρότερα αποτελέσματα.

Για την δημιουργία του Index χρησιμοποίησα τον EnglishAnalyzer, οπότε για την δημιουργία του Word2Vec αντικειμένου ο tokenizerFactory πρέπει να χρησιμοποιεί EnglishAnalyzer().

Για το Similarity του IndexSearcher δοκίμασα MEAN και TF\_IDF smoothing με το πρώτο να επιστρέφει ελαφρώς καλύτερα αποτελέσματα

**Χωρις προεκπαιδευμένο μοντέλο αποτελέσματα στον φάκελο trec\_eval αρχεία k20.txt, k30.txt, k50.txt**

|  |  |  |
| --- | --- | --- |
| **Μέτρο**  **Αξιολόγησης**  **K πρώτα**  **Ανακτηθέντα κείμενα** | MAP  (Mean Average Precision) | avgPre@k  (Average Precision @k) |
| Κ = 20 | 0.1206 | P\_05 = 0.1657  P\_10 = 0.1200  P\_15 = 0.0990  P\_20 = 0.0914 |
| Κ = 30 | 0.1262 | P\_05 = 0.1657  P\_10 = 0.1200  P\_15 = 0.0990  P\_20 = 0.0914 |
| Κ = 50 | 0.1319 | P\_05 = 0.1657  P\_10 = 0.1200  P\_15 = 0.0990  P\_20 = 0.0914 |

Παρατηρούμε χειρότερα αποτελέσματα συγκριτικά με την 1η Φάση της προγραμματιστικής εργασίας.

Παρατηρούμε καλύτερα αποτελέσματα συγκριτικά με την 2η Φάση της προγραμματιστικής εργασίας.

Παρατηρούμε χειρότερα αποτελέσματα συγκριτικά με την 3η Φάση της προγραμματιστικής εργασίας και με την χρηση BM25 αλλά και με LMJ.

Ίσως να έχουμε χειρότερα αποτελέσματα από τις φάσεις 1 και 3 επειδή η συλλογή αποτελείται από πολλά κείμενα άρα τα βάρη δεν μπορούν να ρυθμιστούν πολύ αποτελεσματικά.

**Με προεκπαιδευμένο μοντέλο αποτελέσματα στον φάκελο trec\_eval αρχεία pretrained\_k20.txt, pretrained\_k30.txt, pretrained\_k50.txt**

|  |  |  |
| --- | --- | --- |
| **Μέτρο**  **Αξιολόγησης**  **K πρώτα**  **Ανακτηθέντα κείμενα** | MAP  (Mean Average Precision) | avgPre@k  (Average Precision @k) |
| Κ = 20 | 0.0084 | P\_05 = 0.0171  P\_10 = 0.0143  P\_15 = 0.0152  P\_20 = 0.0114 |
| Κ = 30 | 0.0087 | P\_05 = 0.0171  P\_10 = 0.0143  P\_15 = 0.0152  P\_20 = 0.0114 |
| Κ = 50 | 0.0087 | P\_05 = 0.0171  P\_10 = 0.0143  P\_15 = 0.0152  P\_20 = 0.0114 |

Παρατηρούμε χειρότερα αποτελέσματα συγκριτικά με όλες τις Φάσεις της προγραμματιστικής εργασίας.

Αυτό μπορεί να ωφείλεται σε ένα προεκπαιδευμενο μοντέλο που δεν είναι κατάλληλο για την συλλογή LISA αλλά και σε δικό μου προγραμματιστικό λάθος καθώς δεν πρόλαβα να ελέγξω τυχον παραμετροποιήσεις που μπορεί να χρειάζονταν για την ενσωμάτωση ενός προεκπαιδευμένου μοντέλου ώστε να παραχθούν ακριβής αποτελέσματα