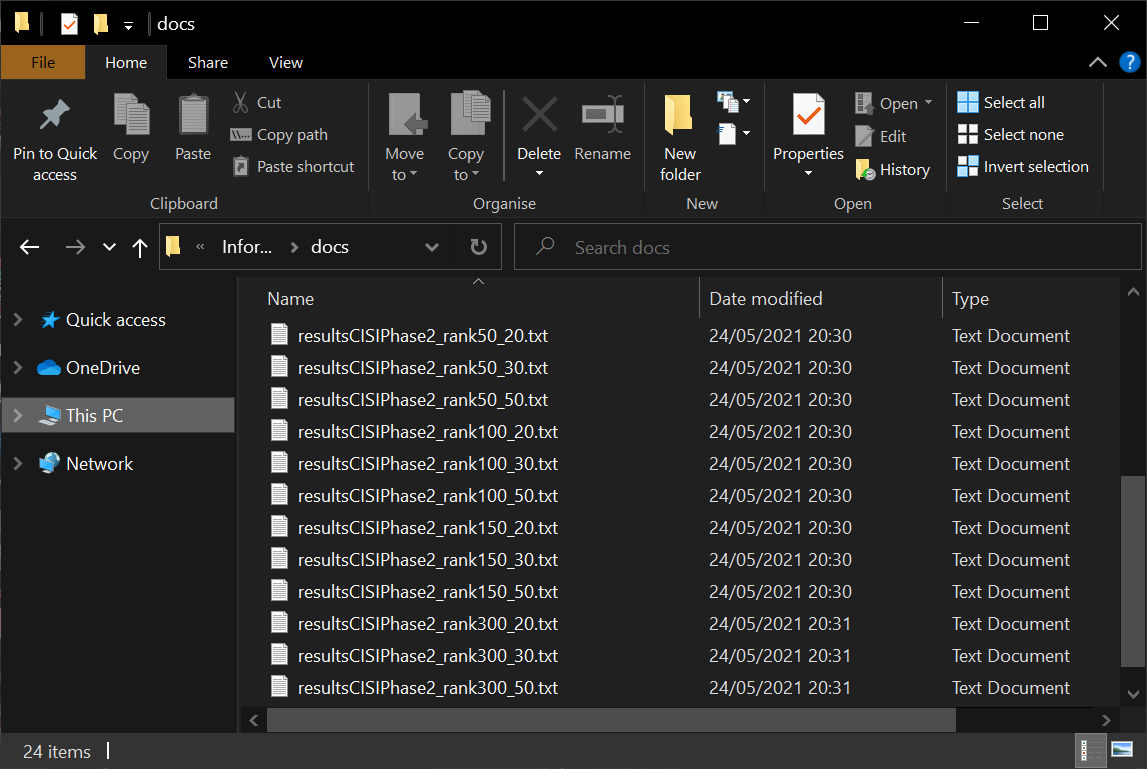
Συστήματα Ανάκτησης Πληροφοριών

Φάση 2 – Εφαρμογή Λανθάνουσας Σημασιολογικής Ευρετηρίασης στην ανάκτηση

Στυλιανή Δούκα – p3170042

Βασίλειος Μπάλλας – p170115

**Περιγραφή Υλοποίησης**

1. Για την αποθήκευση των κειμένων στο ευρετήριο χρησιμοποιήσαμε τα πεδία *.I, .T, .A, .W* όπως και στο προηγούμενο στάδιο της εργασίας*.* Η διαφορά εδώ είναι ότι το πεδίο .W αποθηκεύτηκε ως πίνακας συχνοτήτων (TermXDoc Matrix).Για κάθε ερώτημα (*query*) αποθηκεύσαμε το μοναδικό κωδικό (id) του: *.I* και το περιεχόμενο (content) του: *.W*. Για την επεξεργασία των κειμένων χρησιμοποιήθηκε ο *EnglishAnalyzer*,τοσο για τα κείμενα που αποθηκεύονται στο ευρετήριο, όσο και για τα ερωτήματα (*queries*) που θα χρησιμοποιηθούν για την αναζήτηση.
2. Για την ανάλυση του πίνακα TermXDocs χρησιμοποιήθηκε η βιβλιοθήκη Jama. Με τον αλγόριθμο SVD υπολογίστηκαν οι πίνακες U, S και V. Εφαρμόσαμε την παραλλαγή 2 όπως παρουσιάζεται στο σετ διαφανειών. Συγκεκριμένα, χρησιμοποιήθηκε ο πίνακας Vk για την αναπαράσταση των κειμένων, με το k να αντιπροσωπεύει τις διαστάσεις του διανύσματος κάθε κειμένου. Εκτελέσαμε για k = 50, 100, 150, 300.
3. Για την αναπαράσταση των ερωτημάτων δημιουργήσαμε αραιούς πίνακες μεγέθους termsX1 που αντιπροσωπεύουν τα terms που εμφανίζονται στο ερώτημα. Οι πίνακες αυτοί μετασχηματίστηκαν σε πυκνούς αφού πολλαπλασιάστηκαν με τους πίνακες Uk και Sk. Τα ερωτήματα είναι τώρα στη σωστή μορφή ώστε να εφαρμοστεί η συνημιτονοειδή ομοιότητα. Στο στάδιο αυτό δεν χρησιμοποιήθηκε η lucene καθώς θέλαμε η σειρά των terms να είναι η ίδια με αυτή που εφαρμόστηκε στα κείμενα.
4. Για κάθε ερώτημα υπολογίσαμε τη συνημιτονοειδή ομοιότητα της αναπαράστασης του με κάθε στήλη του πίνακα Vk, δηλαδή με την αναπαράσταση κάθε κειμένου στο νέο χώρο. Ο τύπος που χρησιμοποιήθηκε είναι: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ . Τα αποτελέσματα ταξινομήθηκαν σε φθίνουσα σειρά και κρατήθηκαν τα πρώτα k = 20, 30, 50 πιο σχετικά κείμενα.
5. Για την χρήση του εργαλείου trec\_eval τροποποιήσαμε κατάλληλα το αρχείο CISI.REL. Πιο συγκεκριμένα παρατηρήσαμε ότι η τοποθέτηση των κωδικών των αρχείων ήταν λανθασμένη και πως όλες οι τιμές ομοιότητας ήταν μηδενισμένες.

Εκτελέσαμε την εντολή trec\_eval.exe -m all\_trec CISI.REL resultsCISIPhase2\_rank**X**\_50.txt όπου Χ η τάξη που χρησιμοποιήθηκε για την προσέγγιση των πινάκων.

**Εικόνα που περιέχει κείμενο, υπολογιστής

Περιγραφή που δημιουργήθηκε αυτόματαresultsCISIPhase2\_rank50\_50.txt**

|  |  |  |  |
| --- | --- | --- | --- |
| k = 5 | k = 10 | k = 15 | k = 20 |
| 0.0263 | 0.0342 | 0.0351 | 0.0336 |

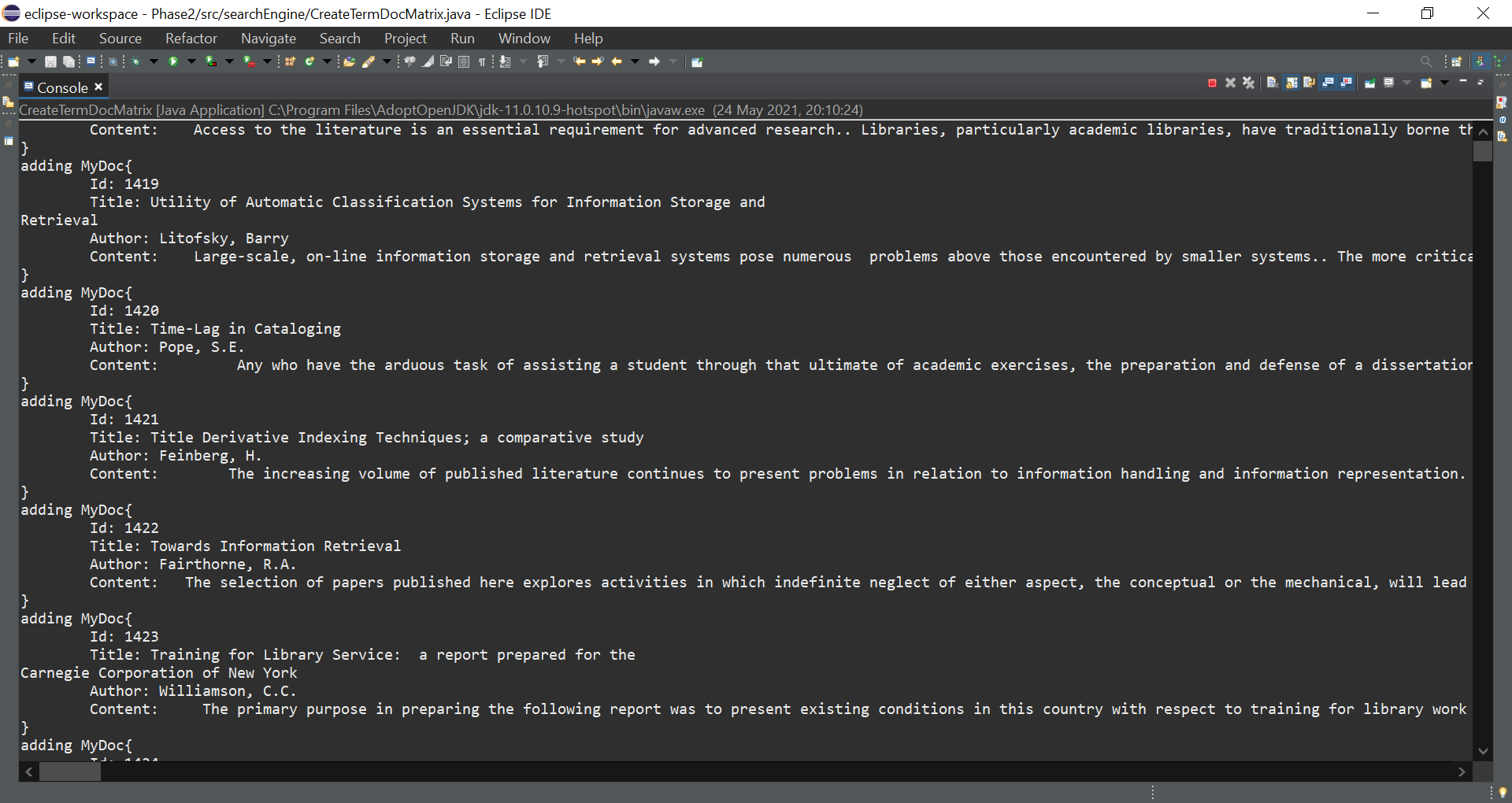
Επιπλέον η τιμή του mean average precision (MAP) για rank = 50 είναι 0.0051.

*(*

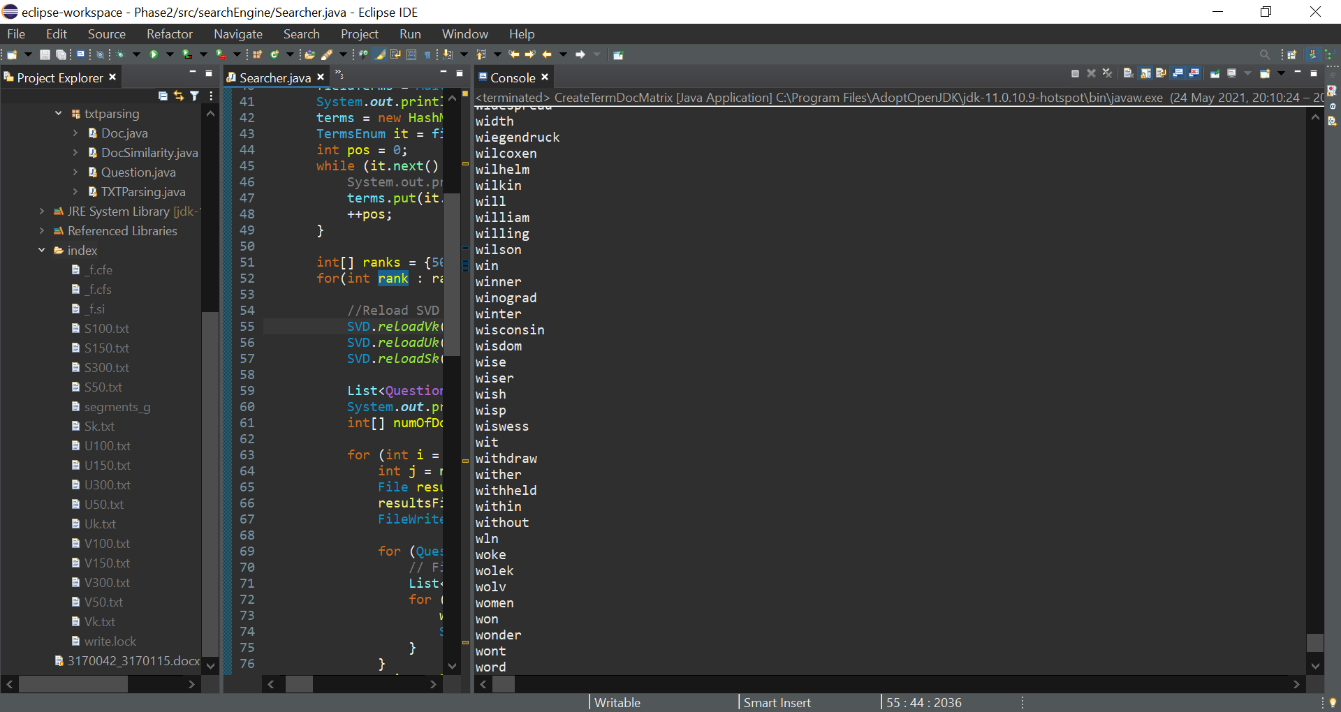
*Παρατηρούμε πως όσο αυξάνεται το k τόσο μειώνεται η μέση ακρίβεια στα k πρώτα ανακτηθέντα κείμενα. Αυτό συμβαίνει καθώς όταν τα κείμενα που συγκρίνονται στο trec\_eval είναι λίγα, η μεση ακρίβεια αυξάνεται αρκετά μόλις βρεθεί μια σωστή απάντηση. Όσο όμως αυξάνεται το πλήθος των κειμένων ο αντίκτυπος που έχουν οι σωστές απαντήσεις στην μεση ακρίβεια μειώνεται.*

*)*

Απόσπασμα από την εκτέλεση του αρχείου CreateTermDocMatrix.java όπου δημιουργείται ο αραιός πίνακας. Βλέπουμε ότι τα documents προστίθενται στο ευρετήριο.

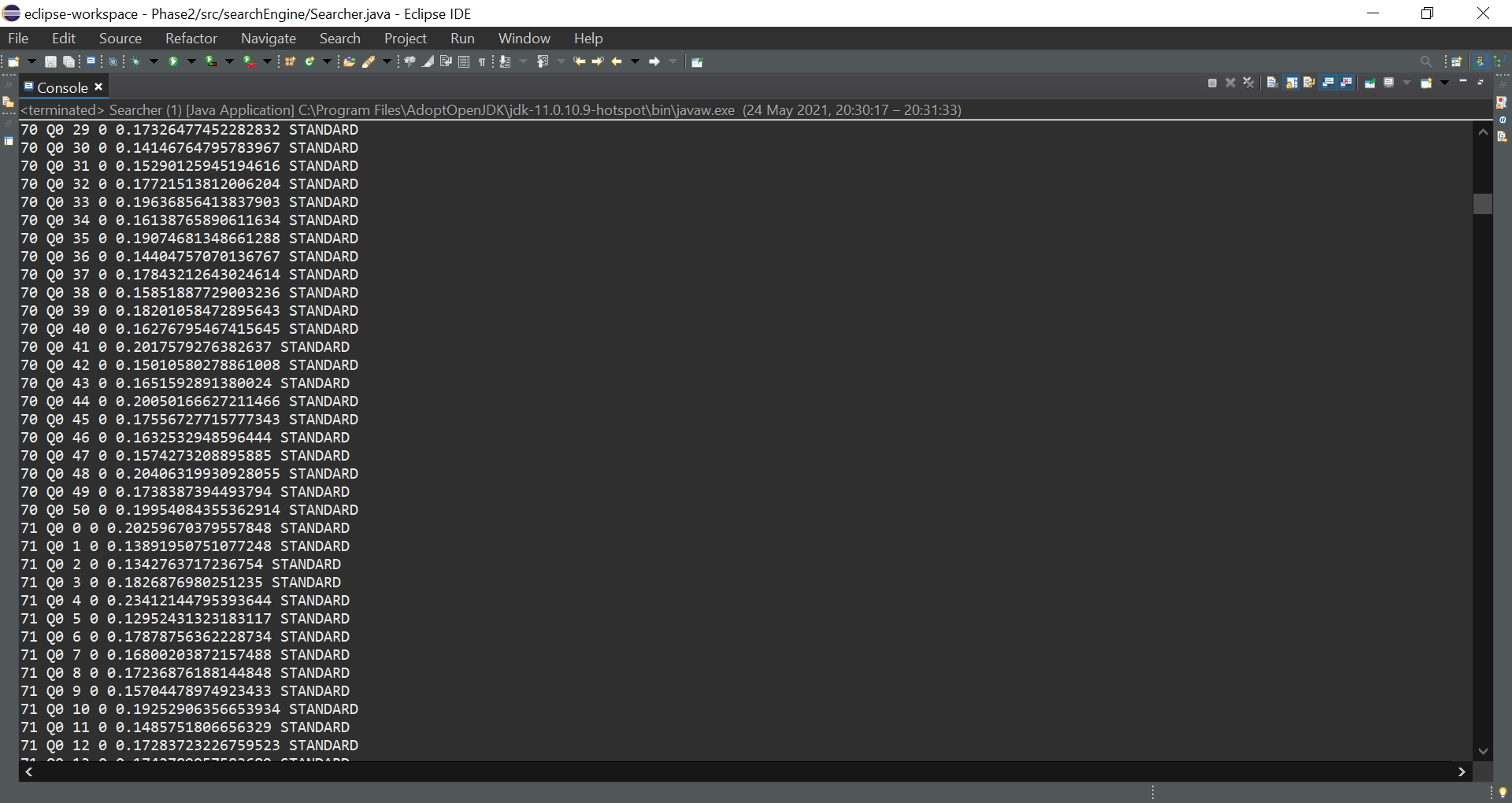


Μέρος των μοναδικών λέξεων που εντόπισε η Lucene, επεξεργασμένων με τον English Analyser, και αποτελούν το λεξιλόγιο μας.



Απόσπασμα από την εκτέλεση του Searcher για την αναζήτηση των ερωτημάτων.

Βλέπουμε ότι αναζητούνται κείμενα σχετικά με τα ερωτήματα 70 και 71 και τυπώνονται τα κείμενα και ο βαθμός ομοιότητας τους σύμφωνα με τη μορφοποίηση που δέχεται το εργαλείο trec\_eval.



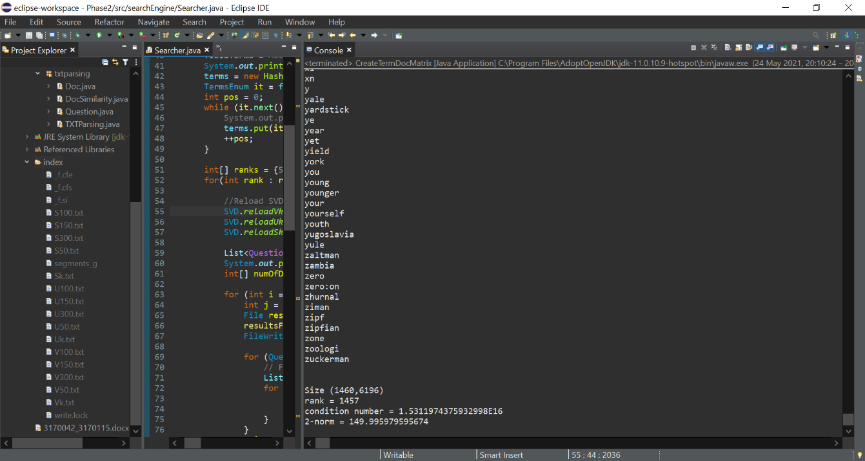
(

Απόσπασμα από τα αποτελέσματα της αναζήτησης για τα πρώτα 50 κείμενα.

Και πάλι η μορφοποίηση είναι συμβατή με το εργαλείο trec\_eval. Παρατηρούμε ότι το πρώτο ερώτημα με κωδικό 1 έχει μέχρι 50 σχετικά κείμενα.

Απόσπασμα από τα αποτελέσματα της αναζήτησης για τα πρώτα 30 κείμενα. Παρατηρούμε ότι το πρώτο ερώτημα με κωδικό 1 έχει μέχρι 30 σχετικά κείμενα.

Απόσπασμα από τα αποτελέσματα της αναζήτησης για τα πρώτα 20 κείμενα. Παρατηρούμε ότι το πρώτο ερώτημα με κωδικό 1 έχει μέχρι 20 σχετικά κείμενα.



)

1. Αν επαναλάβουμε τα προηγούμενα βήματα για τάξη = 50, 100, 150 και 300 θα λάβουμε διαφορετικά αποτελέσματα τα οποία φαίνονται τις επόμενες εικόνες:

Εικόνα που περιέχει κείμενο, ηλεκτρονικές συσκευές, υπολογιστής

Περιγραφή που δημιουργήθηκε αυτόματα**resultsCISIPhase2\_rank100\_50.txt**

|  |  |  |  |
| --- | --- | --- | --- |
| k = 5 | k = 10 | k = 15 | k = 20 |
| 0.0237 | 0.0237 | 0.0398 | 0.0322 |

Map = 0.0050

Εικόνα που περιέχει κείμενο, ηλεκτρονικές συσκευές, υπολογιστής

Περιγραφή που δημιουργήθηκε αυτόματα**resultsCISIPhase2\_rank150\_50.txt**

|  |  |  |  |
| --- | --- | --- | --- |
| k = 5 | k = 10 | k = 15 | k = 20 |
| 0.0263 | 0.0376 | 0.0281 | 0.0322 |

**Map = 0.0052**

**resultsCISIPhase2\_rank300\_50.txt** Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

|  |  |  |  |
| --- | --- | --- | --- |
| k = 5 | k = 10 | k = 15 | k = 20 |
| 0.0316 | 0.0303 | 0.0307 | 0.0336 |

Map = 0.0056

Σύγκριση:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | k = 5 | k = 10 | k = 15 | k = 20 | Map |
| Rank = 50 | 0.0263 | 0.0342 | 0.0351 | 0.0336 | 0.0051 |
| Rank = 100 | 0.0237 | 0.0237 | 0.0398 | 0.0322 | 0.0050 |
| Rank = 150 | 0.0263 | 0.0376 | 0.0281 | 0.0322 | 0.0052 |
| Rank = 300 | 0.0316 | 0.0303 | 0.0307 | 0.0336 | 0.0056 |

Παρατηρούμε ότι δεν υπάρχει δεν υπάρχει κάποια σταθερά με τη χρήση της οποία οι ακρίβεια βελτιώνεται. Συνεπώς είναι στο χέρι μας να δοκιμάσουμε αρκετές υπερπαραμέτρους ώστε να βρούμε αυτές που ταιριάζουν καλύτερα στο πρόβλημα και τα δεδομένα μας.

Πηγές:

<https://introcs.cs.princeton.edu/java/95linear/SVD.java.html>

https://math.nist.gov/javanumerics/jama/doc/

<https://stackoverflow.com/questions/520241/how-do-i-calculate-the-cosine-similarity-of-two-vectors>