Συστήματα Ανάκτησης Πληροφοριών

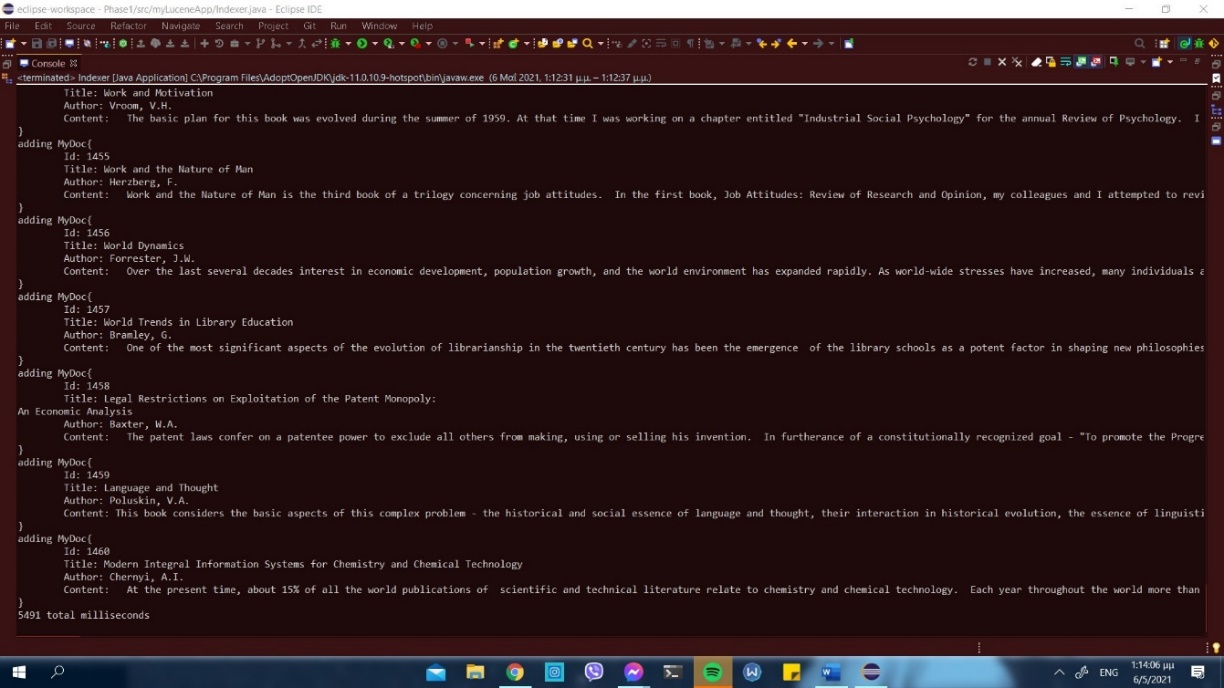
Φάση 1 – Baseline – Μοντέλο Ανάκτησης Διανυσματικού Χώρου

Στυλιανή Δούκα – p3170042

Βασίλειος Μπάλλας – p170115

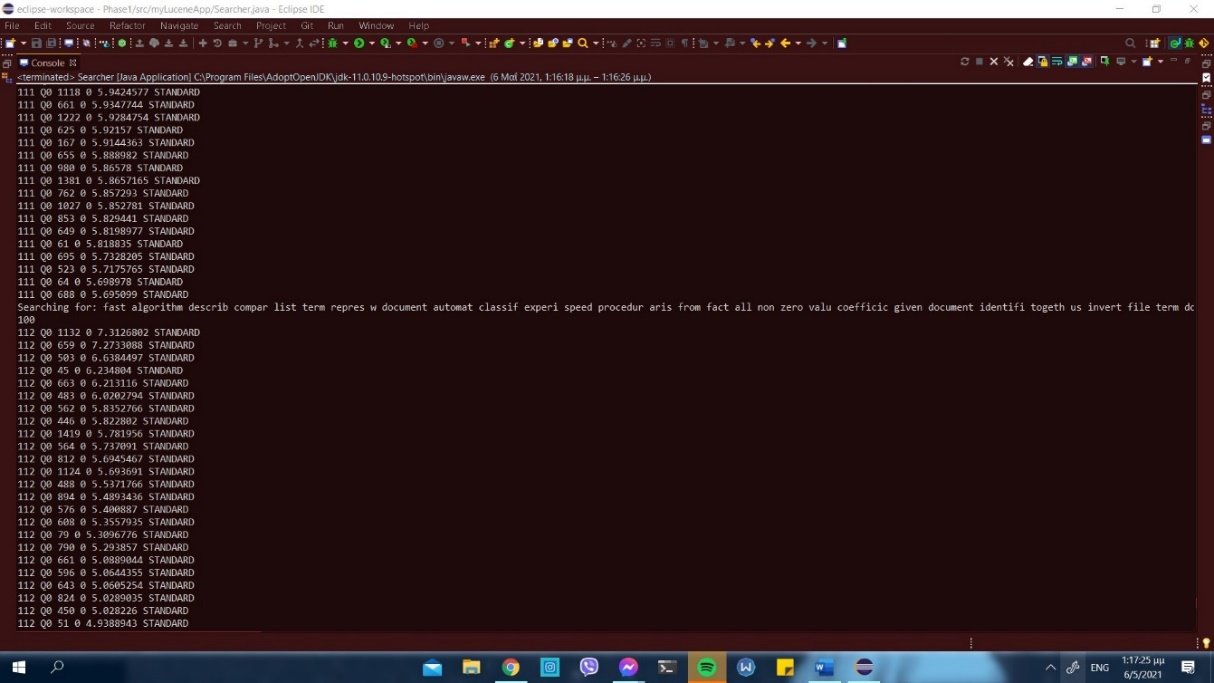
**Περιγραφή Υλοποίησης**

1. Για την αποθήκευση των κειμένων στο ευρετήριο χρησιμοποιήσαμε τα πεδία *.I, .T, .A, .W.* Πιο συγκεκριμένα, για κάθε κείμενο αποθηκεύσαμε το μοναδικό κωδικό (id) του: *.I*, τον τίτλο (title) του: *.Τ*, τον συγγραφέα(author) του: *.A* και το κείμενο (content) του: *.W*. Όλα τα παραπάνω αποθηκεύτηκαν σε *TextFields.* Για κάθε ερώτημα (*query*) αποθηκεύσαμε το μοναδικό κωδικό (id) του: *.I* και το περιεχόμενο (content) του: *.W*.
2. 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. Ως συνάρτηση ομοιότητας, μεταξύ κειμένων και ερωτημάτων, χρησιμοποιήθηκε η *ClassicSimilarity*.
3. Απόσπασμα από την εκτέλεση του Indexer για τη δημιουργία του ευρετηρίου.

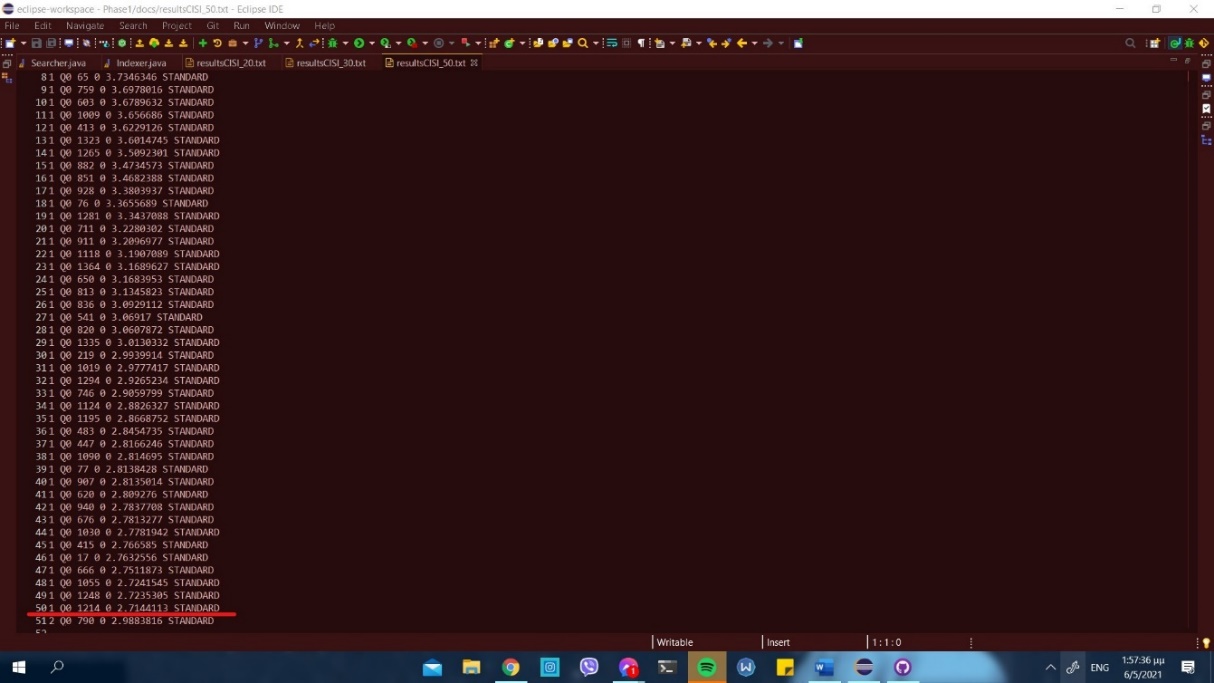
Βλέπουμε ότι αποθηκεύονται στο ευρετήριο 1460 κείμενα (Documents) από τη συλλογή CISI.

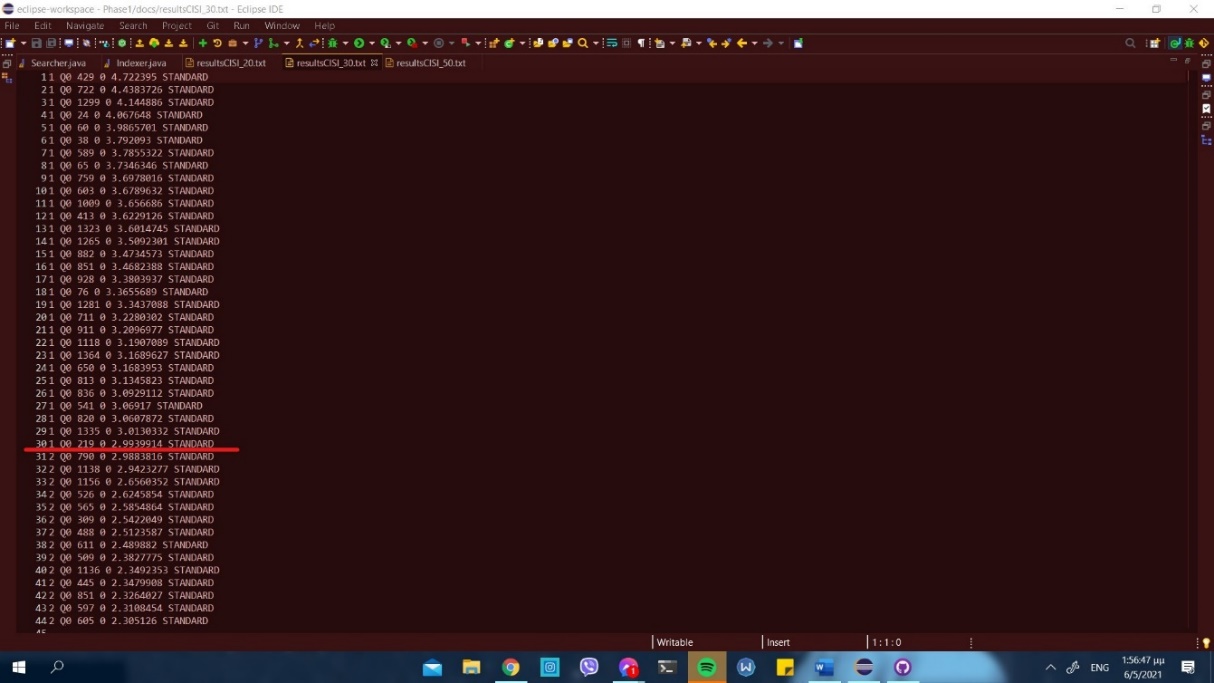
Απόσπασμα από την εκτέλεση του Searcher για την αναζήτηση των ερωτημάτων.

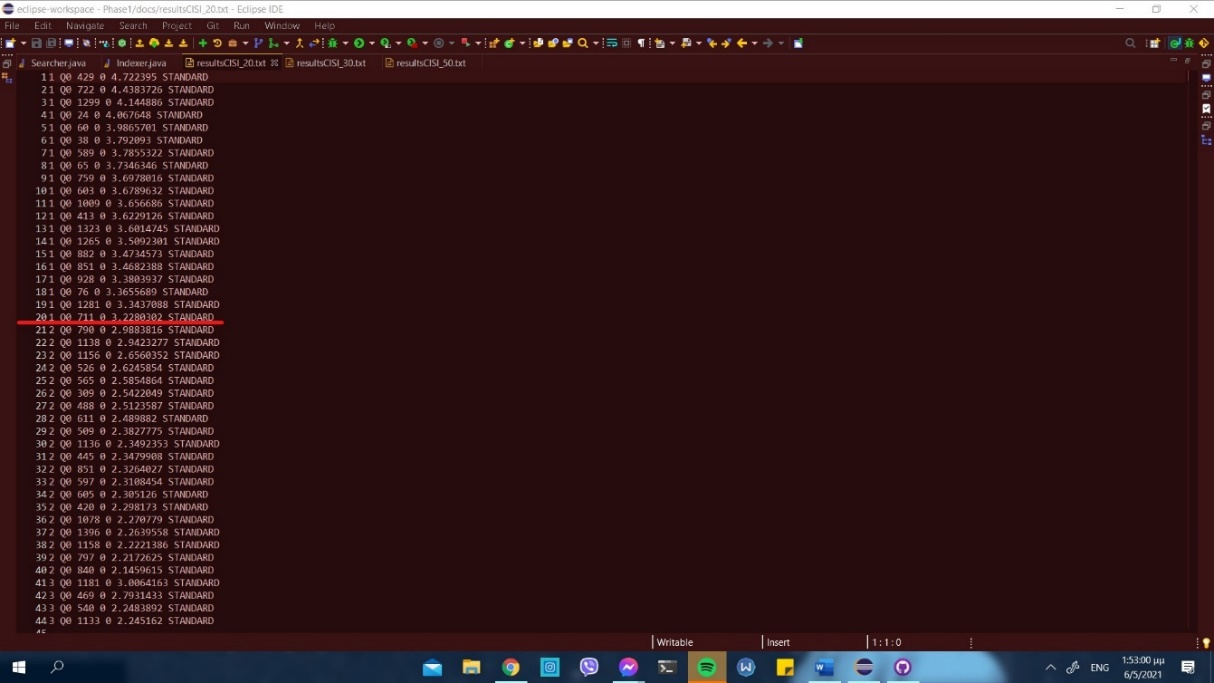
Βλέπουμε ότι αναζητούνται κείμενα σχετικά με τα ερωτήματα 111 και 112 και τυπώνονται τα κείμενα και ο βαθμός ομοιότητας τους σύμφωνα με τη μορφοποίηση που δέχεται το εργαλείο trec\_eval.



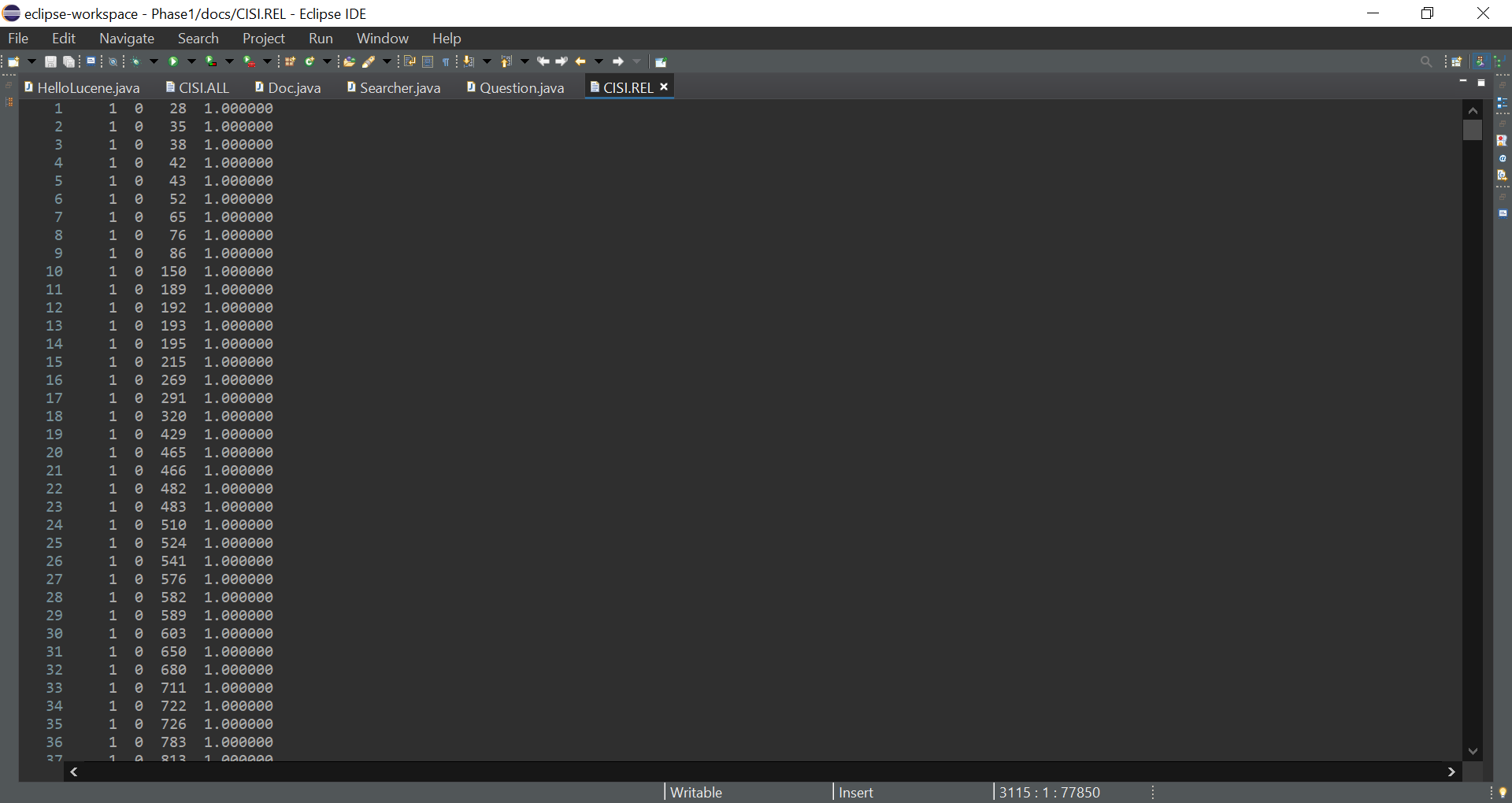
Απόσπασμα από τα αποτελέσματα της αναζήτησης για τα πρώτα 50 κείμενα.

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

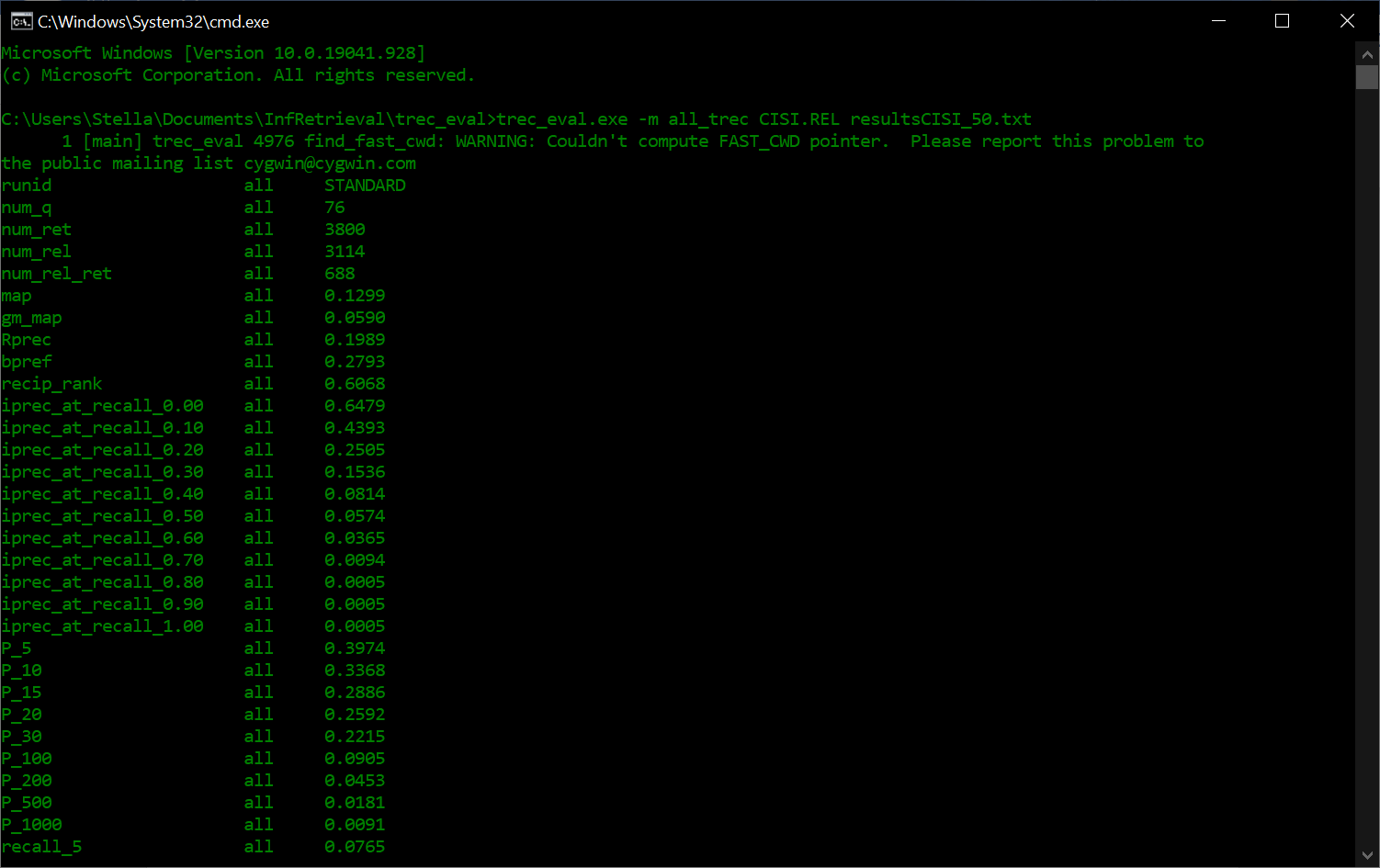
Απόσπασμα από τα αποτελέσματα της αναζήτησης για τα πρώτα 30 κείμενα. Παρατηρούμε ότι το πρώτο ερώτημα με κωδικό 1 έχει μέχρι 30 σχετικά κείμενα.

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. 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. Παρακάτω φαίνεται η τελική μορφή του CISI.REL.



Εκτελέσαμε την εντολή trec\_eval.exe -m all\_trec CISI.REL resultsCISI\_50.txt



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | k=5 | k=10 | k=15 | k=20 |
| avgPre@k | 0.3974 | 0.3368 | 0.2886 | 0.2592 |

Επιπλέον η τιμή του mean average precision (map) είναι 0.1299

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.

Πηγές:

<http://www.rafaelglater.com/en/post/learn-how-to-use-trec_eval-to-evaluate-your-information-retrieval-system>