

Συστήματα Ανάκτησης Πληροφοριών

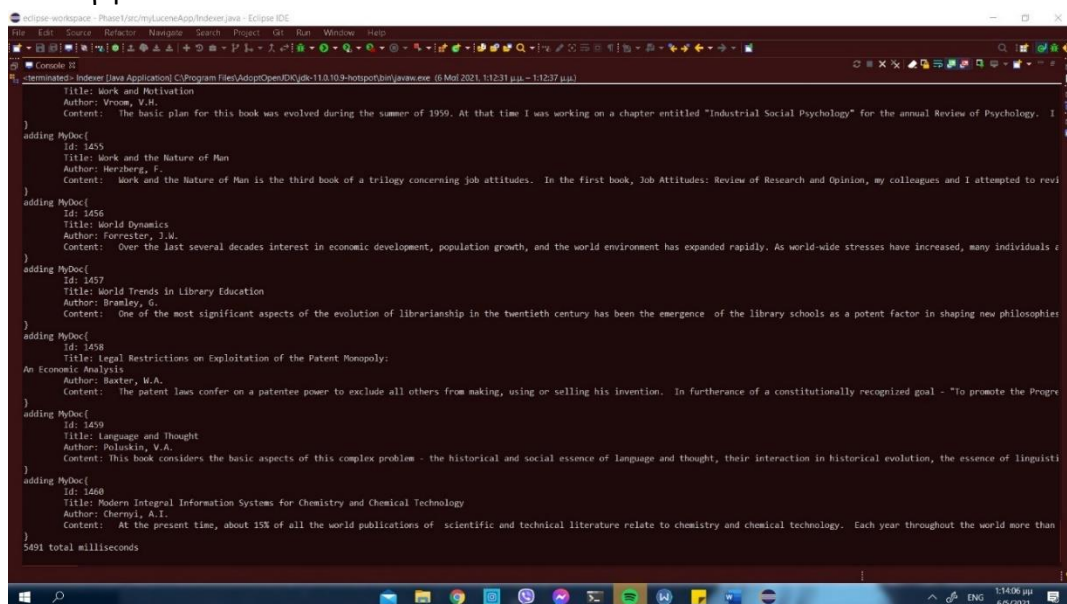
Φάση 1 – Baseline – Μοντέλο Ανάκτησης Διανυσματικού Χώρου

Στυλιανή Δούκα – p3170042

Βασίλειος Μπάλλας – p170115

Περιγραφή Υλοποίησης

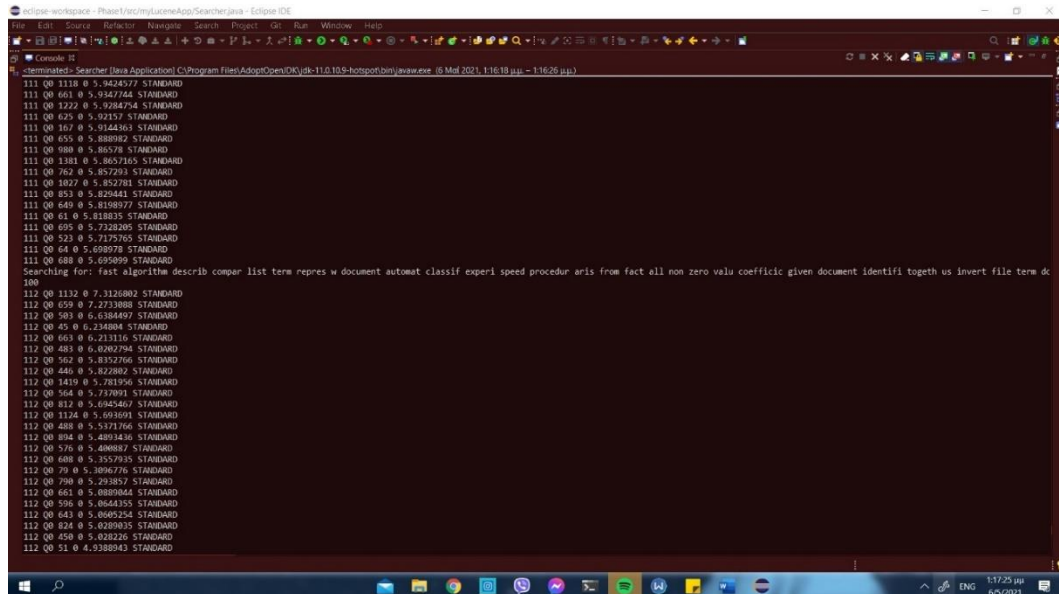
1. Για την αποθήκευση των κειμένων στο ευρετήριο χρησιμοποιήσαμε τα πεδία *.I*, *.T*, *.A*, *.W*. Πιο συγκεκριμένα, για κάθε κείμενο αποθηκεύσαμε το μοναδικό κωδικό (id) του: *.I*, τον τίτλο (title) του: *.T*, τον συγγραφέα(author) του: *.A* και το κείμενο (content) του: *.W*. Όλα τα παραπάνω αποθηκεύτηκαν 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*, τόσο 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.



```
terminated: Indexer [Java Application] C:\Program Files\Adobe\OpenID\jdk-11.0.9-hotspot\bin\java.exe (6 Mol 2021 11:23:11 μμ - 11:23:17 μμ)

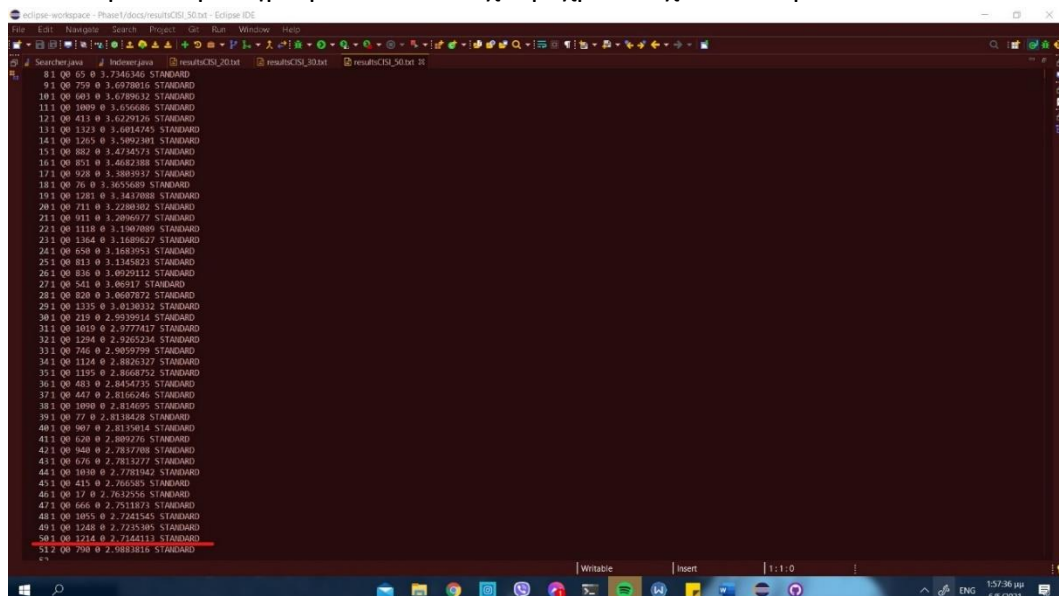
}
Title: Work and Motivation
Author: Vroom, V.H.
Content: The basic plan for this book was evolved during the summer of 1959. At that time I was working on a chapter entitled "Industrial Social Psychology" for the annual Review of Psychology. I
}
adding MyDoc {
  Id: 1455
  Title: Work and the Nature of Man
  Author: Herzberg, F.
  Content: Work and the Nature of Man is the third book of a trilogy concerning job attitudes. In the first book, Job Attitudes: Review of Research and Opinion, my colleagues and I attempted to revi
}
adding MyDoc {
  Id: 1456
  Title: World Dynamics
  Author: Forrester, J.W.
  Content: Over the last several decades interest in economic development, population growth, and the world environment has expanded rapidly. As world-wide stresses have increased, many individuals e
}
adding MyDoc {
  Id: 1457
  Title: World Trends in Library Education
  Author: Bramley, G.
  Content: One of the most significant aspects of the evolution of librarianship in the twentieth century has been the emergence of the library schools as a potent factor in shaping new philosophies
}
adding MyDoc {
  Id: 1458
  Title: Legal Restrictions on Exploitation of the Patent Monopoly:
  An Economic Analysis
  Author: Baxter, W.A.
  Content: The patent laws confer on a patentee power to exclude all others from making, using or selling his invention. In furtherance of a constitutionally recognized goal - "To promote the Progre
}
adding MyDoc {
  Id: 1459
  Title: Language and Thought
  Author: Poluskin, V.A.
  Content: This book considers the basic aspects of this complex problem - the historical and social essence of language and thought, their interaction in historical evolution, the essence of linguisti
}
adding MyDoc {
  Id: 1460
  Title: Modern Integral Information Systems for Chemistry and Chemical Technology
  Author: Chernyi, A.I.
  Content: At the present time, about 15% of all the world publications of scientific and technical literature relate to chemistry and chemical technology. Each year throughout the world more than
}
5491 total milliseconds
```

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.



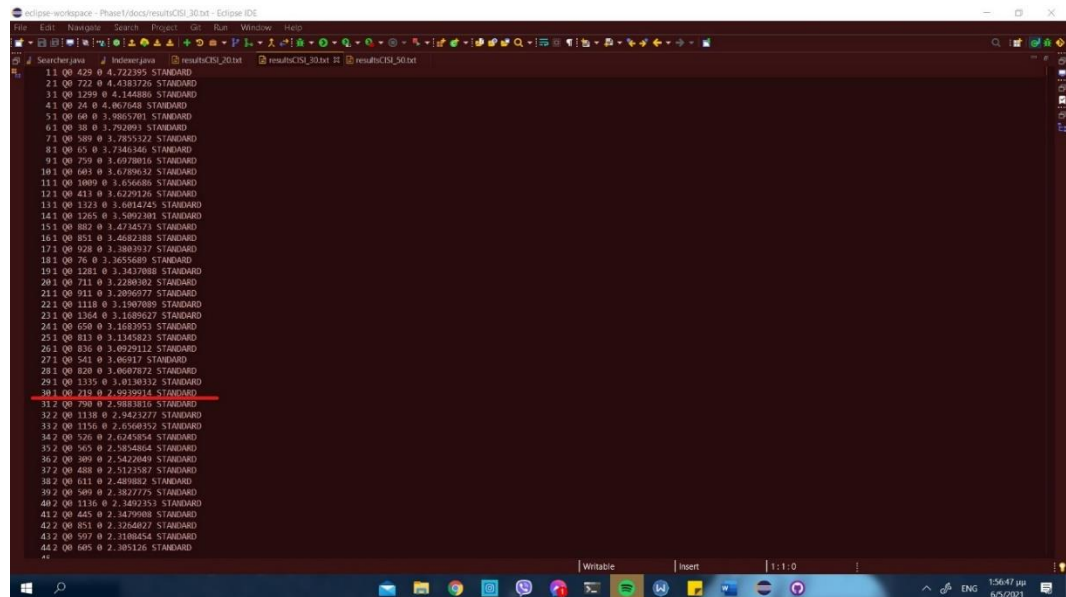
```
<terminated> Searcher [Java Application] C:\Program Files\AdoptOpenJDK\jdk-11.0.10-hotspot\bin\java.exe (8 Mtd 2021, 1.1618 μμ - 1.1626 μμ)
111 Q0 1118 0 5.942497 STANDARD
111 Q0 661 0 5.934774 STANDARD
111 Q0 1222 0 5.928475 STANDARD
111 Q0 625 0 5.92157 STANDARD
111 Q0 167 0 5.914493 STANDARD
111 Q0 655 0 5.888082 STANDARD
111 Q0 980 0 5.86578 STANDARD
111 Q0 1381 0 5.865216 STANDARD
111 Q0 762 0 5.857293 STANDARD
111 Q0 1827 0 5.852781 STANDARD
111 Q0 853 0 5.822441 STANDARD
111 Q0 649 0 5.819897 STANDARD
111 Q0 61 0 5.818835 STANDARD
111 Q0 695 0 5.722828 STANDARD
111 Q0 523 0 5.717576 STANDARD
111 Q0 64 0 5.698978 STANDARD
111 Q0 688 0 5.694989 STANDARD
Searching for: fast algorithm describ compen list repres w document automat classif experi speed procedur aris from fact all non zero valu coeffic given document identifi togeth us Invert file term d
100
112 Q0 1132 0 7.312588 STANDARD
112 Q0 659 0 7.273888 STANDARD
112 Q0 583 0 6.638449 STANDARD
112 Q0 45 0 6.234884 STANDARD
112 Q0 653 0 6.213115 STANDARD
112 Q0 483 0 6.026279 STANDARD
112 Q0 562 0 5.835276 STANDARD
112 Q0 448 0 5.822882 STANDARD
112 Q0 1419 0 5.781956 STANDARD
112 Q0 564 0 5.737891 STANDARD
112 Q0 812 0 5.695467 STANDARD
112 Q0 1124 0 5.693691 STANDARD
112 Q0 488 0 5.537176 STANDARD
112 Q0 898 0 5.489146 STANDARD
112 Q0 576 0 5.488887 STANDARD
112 Q0 608 0 5.355793 STANDARD
112 Q0 79 0 5.389676 STANDARD
112 Q0 798 0 5.283857 STANDARD
112 Q0 661 0 5.088084 STANDARD
112 Q0 596 0 5.064355 STANDARD
112 Q0 643 0 5.060254 STANDARD
112 Q0 824 0 5.028035 STANDARD
112 Q0 450 0 5.028226 STANDARD
112 Q0 51 0 4.938893 STANDARD
```

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.



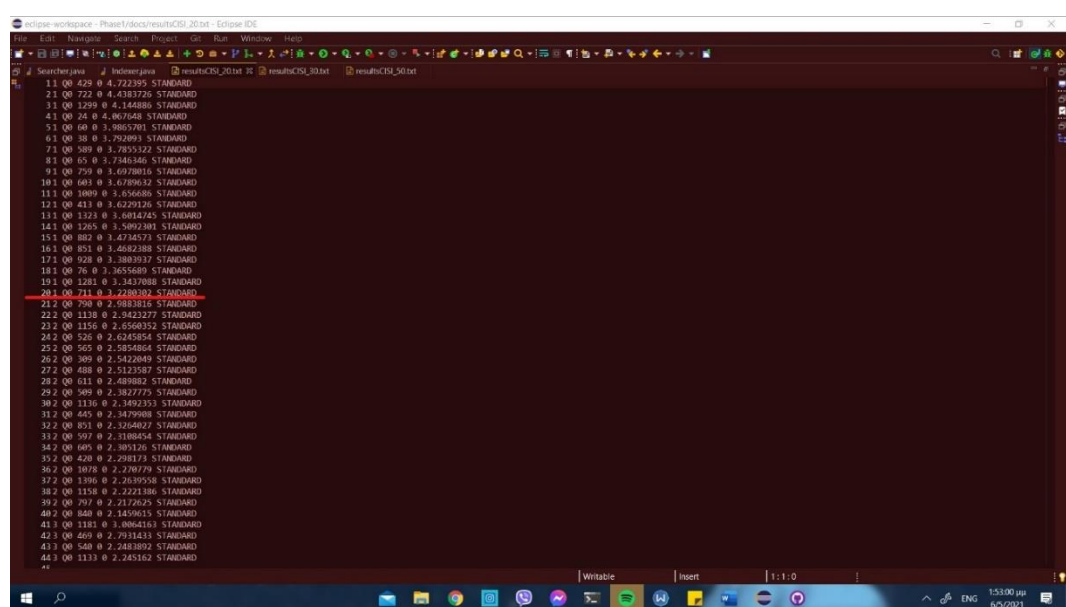
```
Searcher.java | Interpreter.java | resultCSI_0.txt | resultCSI_10.txt | resultCSI_50.txt
81 Q0 65 0 3.7346346 STANDARD
91 Q0 759 0 3.6978016 STANDARD
101 Q0 603 0 3.6789632 STANDARD
111 Q0 1009 0 3.656686 STANDARD
121 Q0 413 0 3.6229126 STANDARD
131 Q0 1323 0 3.6014745 STANDARD
141 Q0 1205 0 3.5992381 STANDARD
151 Q0 882 0 3.4734573 STANDARD
161 Q0 851 0 3.4682388 STANDARD
171 Q0 928 0 3.3883931 STANDARD
181 Q0 76 0 3.3655689 STANDARD
191 Q0 1281 0 3.3437888 STANDARD
201 Q0 711 0 3.2288382 STANDARD
211 Q0 911 0 3.2096927 STANDARD
221 Q0 1118 0 3.1907889 STANDARD
231 Q0 1364 0 3.1680627 STANDARD
241 Q0 650 0 3.1683953 STANDARD
251 Q0 813 0 3.1345823 STANDARD
261 Q0 836 0 3.0925112 STANDARD
271 Q0 541 0 3.06517 STANDARD
281 Q0 820 0 3.0607872 STANDARD
291 Q0 1335 0 3.0130332 STANDARD
301 Q0 215 0 2.9939514 STANDARD
311 Q0 1019 0 2.9777417 STANDARD
321 Q0 1284 0 2.9265234 STANDARD
331 Q0 746 0 2.9059799 STANDARD
341 Q0 1124 0 2.8826327 STANDARD
351 Q0 1195 0 2.8668752 STANDARD
361 Q0 403 0 2.8454735 STANDARD
371 Q0 447 0 2.8166246 STANDARD
381 Q0 1090 0 2.814699 STANDARD
391 Q0 77 0 2.813828 STANDARD
401 Q0 987 0 2.8137014 STANDARD
411 Q0 620 0 2.809276 STANDARD
421 Q0 940 0 2.7837708 STANDARD
431 Q0 676 0 2.7833272 STANDARD
441 Q0 1030 0 2.7781942 STANDARD
451 Q0 415 0 2.765585 STANDARD
461 Q0 77 0 2.7632556 STANDARD
471 Q0 666 0 2.7511873 STANDARD
481 Q0 1055 0 2.7241545 STANDARD
491 Q0 1248 0 2.7235385 STANDARD
501 Q0 1214 0 2.7144113 STANDARD
512 Q0 750 0 2.6881816 STANDARD
```

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



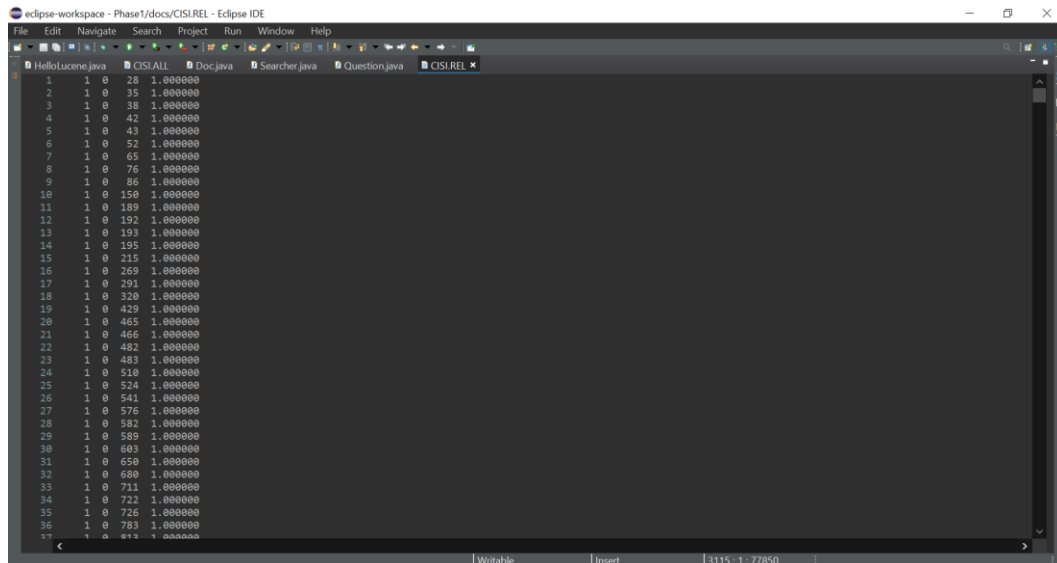
```
11 Q0 429 0 4.722395 STANDARD
21 Q0 722 0 4.4383726 STANDARD
31 Q0 1299 0 4.144888 STANDARD
41 Q0 26 0 4.867648 STANDARD
51 Q0 68 0 3.9865781 STANDARD
61 Q0 38 0 3.792093 STANDARD
71 Q0 389 0 3.785532 STANDARD
81 Q0 65 0 3.7546346 STANDARD
91 Q0 759 0 3.6978016 STANDARD
101 Q0 683 0 3.6789632 STANDARD
111 Q0 3889 0 3.656486 STANDARD
121 Q0 413 0 3.6229126 STANDARD
131 Q0 1323 0 3.6014745 STANDARD
141 Q0 3265 0 3.5992381 STANDARD
151 Q0 882 0 3.4734573 STANDARD
161 Q0 851 0 3.4682388 STANDARD
171 Q0 928 0 3.3883932 STANDARD
181 Q0 76 0 3.3655689 STANDARD
191 Q0 1281 0 3.3437888 STANDARD
201 Q0 711 0 3.2289882 STANDARD
211 Q0 911 0 3.2096977 STANDARD
221 Q0 1118 0 3.1987889 STANDARD
231 Q0 1364 0 3.1688627 STANDARD
241 Q0 658 0 3.1683953 STANDARD
251 Q0 813 0 3.1345823 STANDARD
261 Q0 836 0 3.0929112 STANDARD
271 Q0 541 0 3.06917 STANDARD
281 Q0 820 0 3.0607872 STANDARD
291 Q0 1335 0 3.0130332 STANDARD
301 Q0 718 0 2.9939914 STANDARD
312 Q0 790 0 2.9883816 STANDARD
322 Q0 1138 0 2.9423277 STANDARD
332 Q0 1156 0 2.6580392 STANDARD
342 Q0 526 0 2.6245854 STANDARD
352 Q0 565 0 2.5854864 STANDARD
362 Q0 389 0 2.5420849 STANDARD
372 Q0 488 0 2.5121587 STANDARD
382 Q0 611 0 2.488882 STANDARD
392 Q0 589 0 2.3827775 STANDARD
402 Q0 1110 0 2.3462333 STANDARD
412 Q0 445 0 2.3479988 STANDARD
422 Q0 851 0 2.3264027 STANDARD
432 Q0 597 0 2.3180454 STANDARD
442 Q0 685 0 2.305126 STANDARD
45
```

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



```
11 Q0 429 0 4.722395 STANDARD
21 Q0 722 0 4.4383726 STANDARD
31 Q0 1299 0 4.144888 STANDARD
41 Q0 26 0 4.867648 STANDARD
51 Q0 68 0 3.9865781 STANDARD
61 Q0 38 0 3.792093 STANDARD
71 Q0 389 0 3.785532 STANDARD
81 Q0 65 0 3.7546346 STANDARD
91 Q0 759 0 3.6978016 STANDARD
101 Q0 683 0 3.6789632 STANDARD
111 Q0 3889 0 3.656486 STANDARD
121 Q0 413 0 3.6229126 STANDARD
131 Q0 1323 0 3.6014745 STANDARD
141 Q0 3265 0 3.5992381 STANDARD
151 Q0 882 0 3.4734573 STANDARD
161 Q0 851 0 3.4682388 STANDARD
171 Q0 928 0 3.3883932 STANDARD
181 Q0 76 0 3.3655689 STANDARD
191 Q0 1281 0 3.3437888 STANDARD
201 Q0 711 0 3.2289882 STANDARD
212 Q0 790 0 2.9883816 STANDARD
222 Q0 1138 0 2.9423277 STANDARD
232 Q0 1156 0 2.6580392 STANDARD
242 Q0 526 0 2.6245854 STANDARD
252 Q0 565 0 2.5854864 STANDARD
262 Q0 389 0 2.5420849 STANDARD
272 Q0 488 0 2.5121587 STANDARD
282 Q0 611 0 2.488882 STANDARD
292 Q0 589 0 2.3827775 STANDARD
302 Q0 1110 0 2.3462333 STANDARD
312 Q0 445 0 2.3479988 STANDARD
322 Q0 851 0 2.3264027 STANDARD
332 Q0 597 0 2.3180454 STANDARD
342 Q0 685 0 2.305126 STANDARD
352 Q0 428 0 2.293173 STANDARD
362 Q0 3898 0 2.278778 STANDARD
372 Q0 1396 0 2.2639558 STANDARD
382 Q0 1158 0 2.2221386 STANDARD
392 Q0 797 0 2.217626 STANDARD
402 Q0 848 0 2.1450615 STANDARD
413 Q0 1181 0 3.0864183 STANDARD
423 Q0 469 0 2.793183 STANDARD
431 Q0 540 0 2.481892 STANDARD
443 Q0 1133 0 2.245162 STANDARD
45
```

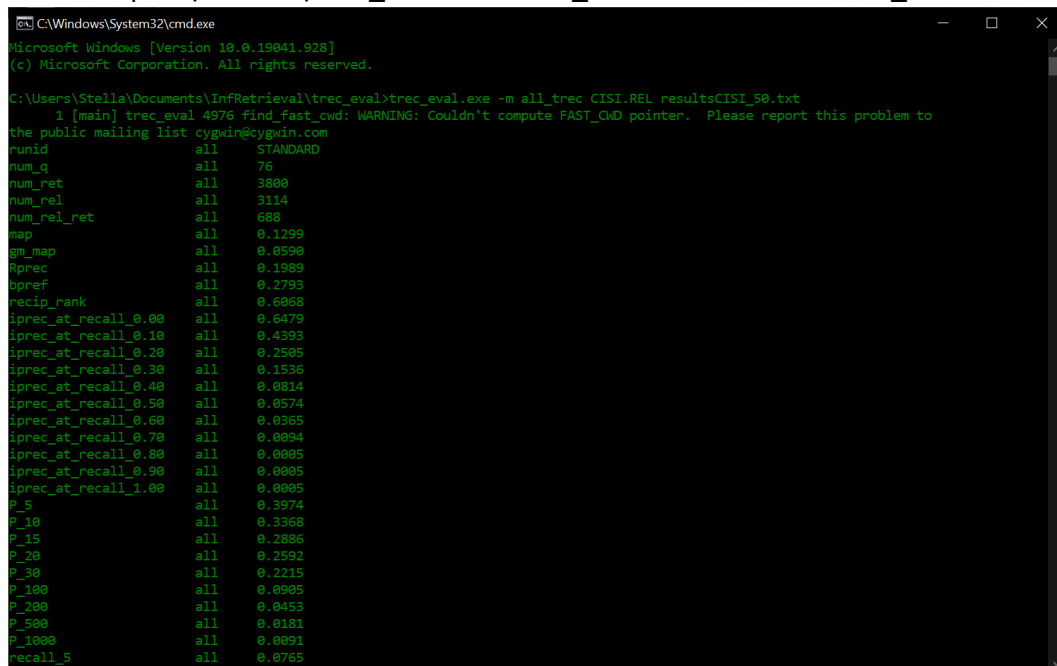
4. Για την χρήση του εργαλείου `trec_eval` τροποποιήσαμε κατάλληλα το αρχείο CISI.REL. Πιο συγκεκριμένα παρατηρήσαμε ότι η τοποθέτηση των κωδικών των αρχείων ήταν λανθασμένη και πως όλες οι τιμές ομοιότητας ήταν μηδενισμένες. Παρακάτω φαίνεται η τελική μορφή του CISI.REL.



The screenshot shows the Eclipse IDE with the file `CISI.REL` open. The file contains a list of 36 lines, each representing a document pair. Each line has four columns: a line number, a document ID, a relevance score, and a similarity score. The similarity scores are all 1.000000.

Line	Doc1	Doc2	Relevance	Similarity
1	1	0	28	1.000000
2	1	0	35	1.000000
3	1	0	38	1.000000
4	1	0	42	1.000000
5	1	0	43	1.000000
6	1	0	52	1.000000
7	1	0	65	1.000000
8	1	0	76	1.000000
9	1	0	86	1.000000
10	1	0	150	1.000000
11	1	0	189	1.000000
12	1	0	192	1.000000
13	1	0	193	1.000000
14	1	0	195	1.000000
15	1	0	215	1.000000
16	1	0	269	1.000000
17	1	0	291	1.000000
18	1	0	320	1.000000
19	1	0	429	1.000000
20	1	0	465	1.000000
21	1	0	466	1.000000
22	1	0	482	1.000000
23	1	0	483	1.000000
24	1	0	510	1.000000
25	1	0	524	1.000000
26	1	0	541	1.000000
27	1	0	576	1.000000
28	1	0	582	1.000000
29	1	0	589	1.000000
30	1	0	603	1.000000
31	1	0	650	1.000000
32	1	0	680	1.000000
33	1	0	711	1.000000
34	1	0	722	1.000000
35	1	0	726	1.000000
36	1	0	783	1.000000

Εκτελέσαμε την εντολή `trec_eval.exe -m all_trec CISI.REL resultsCISI_50.txt`



The screenshot shows a Windows command prompt window with the output of the `trec_eval.exe` command. The output displays various performance metrics for the 'all' model, including precision, recall, and F1 score at different thresholds.

Metric	Value
num_q	76
num_ret	3800
num_rel	3114
num_rel_ret	688
map	0.1299
gn_map	0.0590
aprec	0.1989
bpref	0.2793
recip_rank	0.6068
iprec_at_recall_0.00	0.6479
iprec_at_recall_0.10	0.4393
iprec_at_recall_0.20	0.2505
iprec_at_recall_0.30	0.1536
iprec_at_recall_0.40	0.0814
iprec_at_recall_0.50	0.0574
iprec_at_recall_0.60	0.0365
iprec_at_recall_0.70	0.0094
iprec_at_recall_0.80	0.0005
iprec_at_recall_0.90	0.0005
iprec_at_recall_1.00	0.0005
P_5	0.3974
P_10	0.3368
P_15	0.2886
P_20	0.2592
P_30	0.2215
P_100	0.0905
P_200	0.0453
P_500	0.0181
P_1000	0.0091
recall_5	0.0765

	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 μέση ακρίβεια αυξάνεται αρκετά μόλις βρεθεί μια σωστή απάντηση. Όσο όμως αυξάνεται το πλήθος των κειμένων ο αντίκτυπος που έχουν οι σωστές απαντήσεις στην μέση ακρίβεια μειώνεται.



Πηγές:

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