Κεχρής Κωνσταντίνος

ΑΜ:3150071

ΦΑΣΗ Β

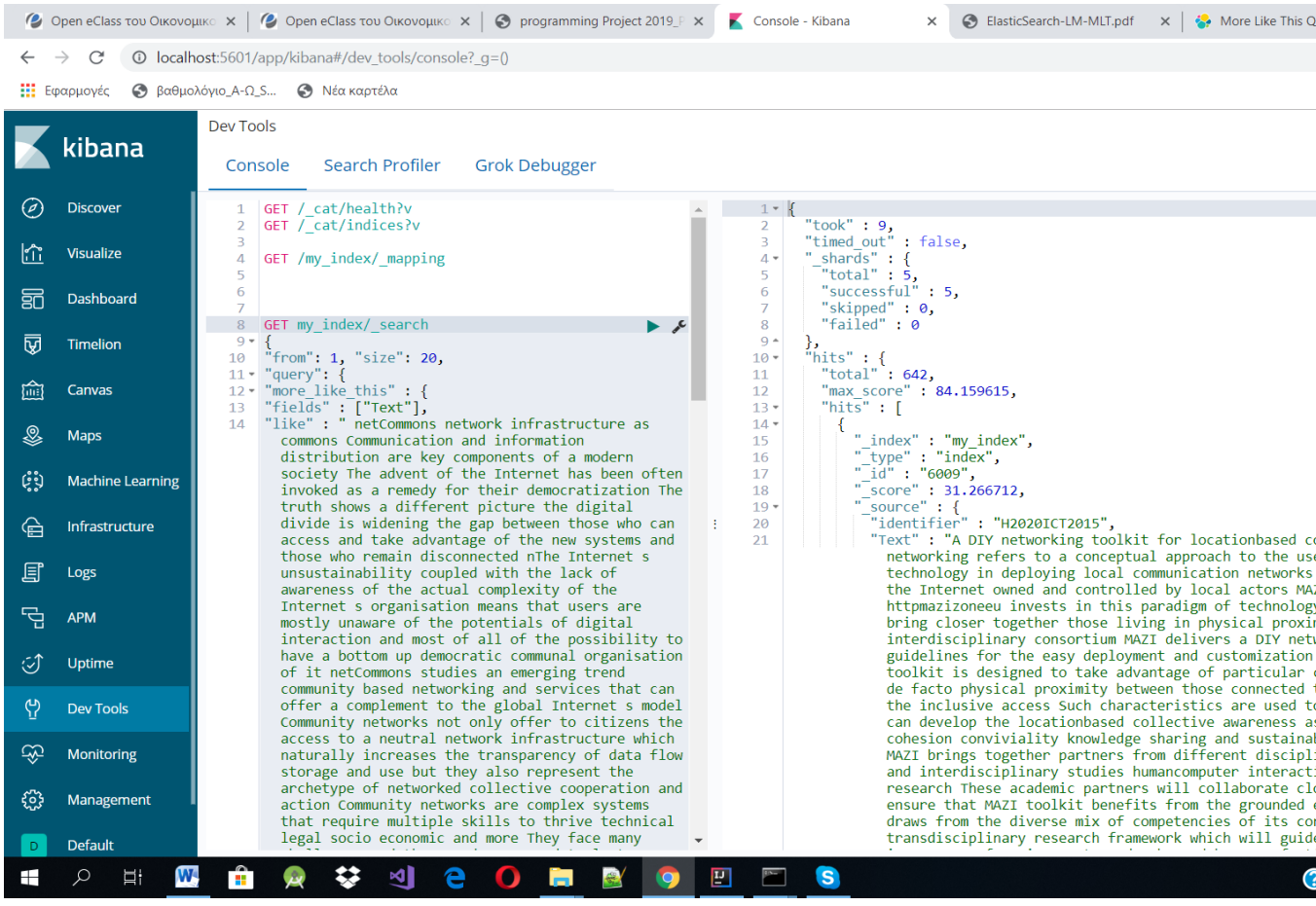
1) Α) **Ανάλυση κώδικα**: Αρχικά στην κλάση Main γίνεται ο εντοπισμός των txt αρχείων που μας ενδιαφέρουν. Αφού τα εντοπίσουμε γράφουμε 3 νέα αρχεία χρησιμοποιώντας ως βάση το υπάρχον αρχείο που διαβάσαμε. Πολλαπλασιάζουμε τις γραμμές του με το 0.3 (30%), με το 0.6 (60%) και με το 0.9 (90%). Τα νέα αρχεία τα αποθηκεύουμε στο φάκελο New_Collection_1. Τα τρία πρώτα αρχεία αφορούν το πρώτο κείμενο που μας ενδιαφέρει (193378.txt) δηλαδή το αρχείο 1.txt είναι το 30% του 193378.txt το 2.txt είναι το 60% του 193378.txt και το 3.txt είναι το 90% του 193378.txt. Όμοια δουλεύουμε για τα υπόλοιπα αρχεία που μας ενδιαφέρουν 213164.txt, 204146.txt κ.λ.π.

Για το πρώτο ερώτημα της δεύτερης φάσης χρησιμοποιήθηκε το ίδιο index με την πρώτη φάση!

Εκτελούμε τα ερωτήματα δηλαδή για να βγάλουμε πόρισμα για τα 30% των κειμένων χρησιμοποιούμε τα αρχεία 1.txt,4.txt,7.txt,10.txt,13.txt,16.txt,19.txt,22.txt,25.txt,28.txt. Εκτελούμε τα ερωτήματα στο kibana και παίρνουμε το αποτέλεσμα copy paste και το βάζουμε στο φάκελο Results/30% 1,2 Κλπ για κάθε αρχείο. Ομοίως δουλεύουμε και για τα άλλα ερωτήματα.

Στη συνέχεια, για να εξάγουμε τα αποτελέσματα δηλαδή το RCN και το score χρησιμοποιούμε τις κλάσεις Test (για τον φάκελο 30%) Test2 (για τον φάκελο 60%) και την κλάση Test3 (για τον φάκελο 90%). Παίρνουμε το αποτέλεσμα copy paste από την κονσόλα κάθε φορά που τρέχουμε τις κλάσεις και τα βάζουμε στα αρχεία 30%,60% και 90%. Συνεπώς το κάθε αρχείο έχει 200 γραμμές κάθε γραμμή έχει σκορ και RCN. Κατόπιν, τρέχουμε τις κλάσεις WriteResults, WriteResults2 και WriteResults3 όπου φτίαχνουν τα αρχεία RESULTS30%.txt, RESULTS60%.txt και RESULTS90%.txt. Τα αποτελέσματα του trec_eval είναι τα αρχεία 3150071_result30%.txt, 3150071_result60%.txt, 3150071_result90%.txt.

2) Για το δεύτερο ερώτημα χρησιμοποιήσαμε το ίδιο index που είχαμε και στην πρώτη φάση. Εκτελούμε αρχικά τα ερωτήματα όπως φαίνονται παρακάτω.



Αξίζει να αναφερθεί πως για να “καθαρίσουμε” τα αρχικά ερωτήματα από ειδικούς χαρακτήρες ή /n χρησιμοποιούμε την κλάση New_Test όπου αποβάλλει τους ειδικούς χαρακτήρες. Στη συνέχεια εκτελούμε το query στο kibana. Παίρνουμε copy paste τα αποτελέσματα και τα βάζουμε στο φάκελο Results/Default. Μετά τρέχουμε την κλάση Test_Default όπου επεξεργάζεται τα αρχεία και μας επιστρέφει στην κονσόλα 200 RCN και τα σκορ τους. Παίρνουμε copy-paste τα αποτελέσματα από κονσόλα και τα βάζουμε στο αρχείο Default. Ύστερα, τρέχουμε την κλάση WriteResultsDefault όπου φτιάχνει το αρχείο RESULTSDEFAULT και στην συνέχεια τρέχοντας το trec_eval. Το αποτέλεσμα είναι στο αρχείο 3150071_result_Default.txt

Με τον ίδιο ακριβώς τρόπο εργαζόμαστε και για να τροποποιήσουμε τις παραμέτρους του εργαλείου mlt χρησιμοποιώντας τις υπόλοιπες κλάσεις.

Η πρωτη δοκιμή που κάναμε ήταν να αλλάξουμε το max_query_terms σε 20, το min_term_freq σε 1 και το min_doc_freq σε 4.

Η δεύτερη δοκιμή που κάναμε ήταν να αλλάξουμε το max_query_terms σε 15, το min_term_freq σε 0 και το min_doc_freq σε 3.

Η Τρίτη δοκιμή που κάναμε ήταν να αλλάξουμε το max_query_terms σε 5, το min_term_freq σε 0 και το min_doc_freq σε 1.

Τα αποτελέσματα του trec_eval είναι στα αρχεία 3150071_results_parameter1, 3150071_results_parameter2 και 3150071_results_parameter3 για τα αντίστοιχα πειράματα που έκανα με τις παραμέτρους.

Παρατηρούμε ότι όταν μειώνουμε λίγο το max_query_terms από 25 σε 20 έχουμε καλύτερο σκορ, αυτό δεν συμβαίνει όμως όταν το μειώνουμε αρκετά σε 5 δηλαδή όπως το τρίτο πειραμα.