Συστήματα Ανάκτησης Πληροφοριών

Φάση 1– Baseline-Ανάκτηση χωρίς εύρεση προτεινόμενων κείμενων

Γεώργιος- Στέφανος Μεϊδάνης – p3170107

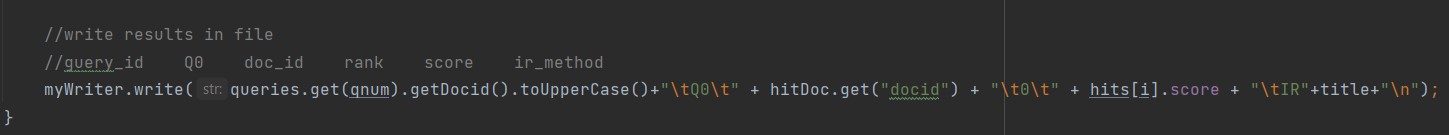
Σταυρούλα Λέκκα - 3090108

1. **Υλοποίηση**

Για την υλοποίηση του συστήματος(μέθοδος MyLucene) βασιστήκαμε στον κώδικα που παρουσιάστηκε στο εργαστήριο. Εφαρμόσαμε αλλαγές προκειμένου να ταιριάζουν στα δωθέντα κείμενα και ερωτήματα.

Το κάθε Document αρχικοποιείται με πεδία docId(Αριθμός Εγγράφου), title(τίτλος), content(Περιεχόμενο).

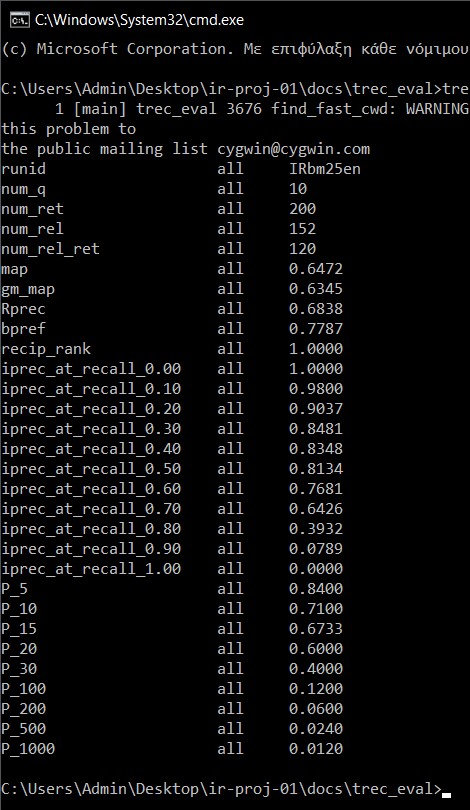
 Το σύστημά μας πραγματοποιεί με την σειρά : indexing των κειμένων, Διάβασμα αρχείων ερωτημάτων(query) και Αναζήτηση-Εξαγωγή αποτελεσμάτων, με παραμέτρους Κ(ανακτηθέντα κείμενα), title( τίτλος αρχείου εξαγωγής), analyzer, similarity.

 Με αυτή την μέθοδο εξασφαλίζουμε ότι τα αποτελέσματά μας θα γραφούν με την απαιτούμενη μορφή για την αξιολόγησή τους από το trec eval.

1. **Εμφάνηση Αποτελεσμάτων**

Για τα παρακάτω παραδείγματα χρησιμοποιήθηκαν EnglishAnalyzer και BM25Similarity.

* **Εικόνα που περιέχει κείμενο

  Περιγραφή που δημιουργήθηκε αυτόματα**Απόσπασμα από την εκτέλεση της μεθόδου (κ=20)
* Διαβάστηκαν και ευρετηριάστηκαν 18316 κείμενα
* Διαβάστηκαν και απαντήθηκαν 10 ερωτήματα , για τα οποία φαίνεται και πόσες φορές είχαμε ταυτοποίηση όρου
* Παραγόμενο αρχείο και απόσπασμα της σύγκρισης των απαντήσεων μας με τα qrels, μέσω του trec\_eval Εικόνα που περιέχει πίνακας

  Περιγραφή που δημιουργήθηκε αυτόματα

1. **Συμπεράσματα**

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=20 | K=30 | K=50 |
| MAP | 0.6472 | 0.6944 | 0.7132 |

|  |  |
| --- | --- |
| P\_5 | 0.84 |
| P\_10 | 0.71 |
| P\_15 | 0.6733 |
| P\_20 | 0.6 |

* Παρατηρούμε αύξηση του MAP καθώς αυξάνουμε τον αριθμό αποτελεσμάτων, καθώς επιτρέπουμε περισσότερα σχετικά κείμενα να αξιολογηθούν.
* Ταυτόχρονα, για κάθε ξεχωριστή αναζήτηση, τα αποτελέσματα της μέσης ακρίβειας για 5, 10, 15 και 20 πρώτα κείμενα μειώνεται. 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 μέση ακρίβεια μειώνεται.