Συστήματα Ανάκτησης Πληροφοριών

Φάση 2 Χρήση όρων με μεγάλη συχνότητα εμφάνισης για την εύρεση  παρόμοιου περιεχομένου

Γεώργιος- Στέφανος Μεϊδάνης – p3170107

Σταυρούλα Λέκκα - 3090108

Για την δεύτερη φάση της εργασίας επιχειρήσαμε να αξιοποιήσουμε την δυνατότητα MoreLikeThis της Lucene προκειμένου να βελτιώσουμε τα αποτελέσματα της αναζήτησής μας. Υλοποιήσαμε 2 διαφορετικές μεθόδους, χρησιμοποιόντας σε κάθε περίπτωση English Analyzer και Classic Similarity.

Στην πρώτη μας δοκιμή εφαρμόσαμε την MoreLikeThis πάνω στα querries προκειμένου να μας επιστραφούν κείμενα σχετικά με τους πιο σημαντικούς όρους από τα queries.

Απόσπασμα από trecEval για:

**20 ,**  **30**

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματαΕικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

**Και 50 αποτελέσματα.**

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

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

|  |  |  |  |
| --- | --- | --- | --- |
| Φάση 1 | K=20 | K=30 | K=50 |
| MAP | 0.6472 | 0.6944 | 0.7132 |

|  |  |  |  |
| --- | --- | --- | --- |
| Φάση 2 | K=20 | K=30 | K=50 |
| MAP | 0.3995 | 0.4194 | 0.4476 |

|  |  |
| --- | --- |
| P\_5 | 0.68 |
| P\_10 | 0.56 |
| P\_15 | 0.46 |
| P\_20 | 0.41 |

Παρατηρούμε πως **το map και τα αποτελέσματα της μέσης ακρίβειας είναι σαφώς μειωμένα, αν και το map είναι ανωδικό**. Αυτό συμβαίνει διότι η αναζήτηση γίνεται με βάση των πιο σημαντικών λέξεων για κάθε ερώτημα, οπότε είναι πιο εύκολο να έρθουν ερωτήματα που δεν είναι τόσο σχετικά με το αρχικό ερώτημα.

Στην δεύτερη μας δοκιμή (ενναλακτική χρήση της MoreLikeThis) τρέξαμε ένα ερώτημα με τους σημαντικότερους όρους για κάθε ένα από τα 20, 30 και 50 διαφορετικά ερωτήματα που μας επέστρεψε η αναζήτηση στο ευρετήριο. Από αυτά επιλέξαμε τα 20,30 και 50 με το καλύτερο σκόρ για να παρουσιάσουμε στον χρήστη και να αξιολογήσουμε.

Σημειώσεις υλοποίησης:

* Για την κάθε μας αναζήτηση χρησιμοποιήσαμε το Field “contents” το οποίο περιλαμβάνει τον τίτλο και τα περιεχόμενα του κειμένου.
* Το Contents χρησιμοποιήθηκε και για την επιλογή των σημαντικότερων όρων από τα αποτελέσματα κατά την 2η δοκιμή.