ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΑΝΑΦΟΡΑ PROJECT

Συστήματα Ανάκτησης Πληροφοριών

ANASTASIOS ZACHARIOUDAKIS (3170048) – NIKOLAOS VATTIS (3170203)

9/5/2021

Φαση 1: Baseline (Συλλογη LISA)

ΒΗΜΑΤΑ 1,2

Για τη προεπεξεργασια του κειμένου φτιαξαμε τη κλαση FileParser η οποια κανει parse τα txt αρχεια της συλλογης, τα οποια οριζονται στη μεταβλητη txtfile της Baseline, την οποια καλουμε 8 φορες για κάθε ξεχωριστο αρχειο txt από documents. Οποτε ο FileParser κανει parse το αρχειο από documents και ξεχωριζει κάθε doc και κάθε field του doc φτιαχνοντας αντικειμενα MyDoc.

Επειτα, επιστρεφει στην List<MyDoc> **docs** της baseline όλα τα documents του txt και με την indexDoc() γραφει στο ευρετηριο τα documents, ενώ με την CREATE_OR_APPEND ειτε δημιουργουμε το ευρετηριο ειτε προσθετουμε στο ευρετηριο τα επομενα.

Τελος, εχουμε επιλεξει ως συναρτηση ομοιοτητας την ClassicSimilarity() η οποία υλοποιει το vector space μοντελο και ως αναλυτη τον EnglishAnalyzer().

Σημειωτέον ότι στην υλοποιηση μας στην Baseline, κανουμε run μονο το 1ο μερος για κάθε αρχειο μεχρι να ολοκληρωθει το ευρετηριο, δηλαδη εχουμε σε σχολια τον κωδικα των επομενων βηματων της εργασιας, ενώ αντιστροφα για κανουμε το αναποδο για να τρεξουμε τα υπολοιπα μερη της εργασιας, δηλαδη το (2.)parse των queries και το (3. search) matching.

*Ακομη, υπαρχει η κλαση ReadIndex() με την οποια μπορουμε να κανουμε print ολοκληρο το index.

*Επίσης, εχουν αφαιρεθει ολες οι κενες τελευταιες γραμμες από όλα τα αρχεια LISA.

ΒΗΜΑΤΑ 3,4

Στη συνεχεια υλοποιησαμε την κλαση QueriesParser την οποια καλουμε για να ξεχωρισει τα queries του αρχειου LISA.QUE και να αποθηκευσει κάθε query ξεχωριστα στην List<QueryDoc> **queries** οπου το queryDoc είναι ένα αντικειμενο query το οποιο αποτελειται από το νουμερο του και το περιεχομενο του.

Επειτα κανουμε το matching με την συναρτηση search για το κάθε query, επιλεγοντας το **hitsPerPage** το οποιο είναι το k μας, το οποιο το τρεχουμε 3 φορες για k=20,30,50 και αποθηκευουμε τα αποτελεσματα στα αρχεια results20.txt , results30.txt , results50.txt.

Επισης εχουμε υλοποιησει την κλαση RelConverter() η οποια μετατρεπει το αρχειο LISA.REL σε ένα αρχειο qrel.txt το οποιο βρισκεται στη μορφη που χρειαζεται το trec_eval.

Παρακάτω παραθετουμε τον πινακα με τα αποτελεσματα του trec_eval για τις διαφορες τιμες του k:

1. για k=20:	1. για k=30:	1. για k=50:
2. runid all STANDARD	2. runid all STANDARD	2. runid all STANDARD
3. num_q all 18	3. num_q all 18	3. num_q all 18
4. num_ret all 308	4. num_ret all 458	4. num_ret all 753
5. num_rel all 225	5. num_rel all 225	5. num_rel all 225
6. num_rel_ret all 42	6. num_rel_ret all 43	6. num_rel_ret all 50
7. map all 0.1139	7. map all 0.1150	7. map all 0.1178
8. gm_map all 0.0028	8. gm_map all 0.0043	8. gm_map all 0.0045
9. Rprec all 0.1556	9. Rprec all 0.1556	9. Rprec all 0.1556
10. bpref all 0.1768	10. bpref all 0.2045	10. bpref all 0.2235
11. recip_rank all 0.3988	11. recip_rank all 0.4010	11. recip_rank all 0.4010
12. iprec_at_recall_0 .00 all 0.4222	12. iprec_at_recall_0 .00 all 0.4244	12. iprec_at_recall_0 .00 all 0.4244
13. iprec_at_recall_0 .10 all 0.3006	13. iprec_at_recall_0 .10 all 0.3028	13. iprec_at_recall_0 .10 all 0.3028
14. iprec_at_recall_0 .20 all 0.2196	14. iprec_at_recall_0 .20 all 0.2218	14. iprec_at_recall_0 .20 all 0.2347
15. iprec_at_recall_0 .30 all 0.1918	15. iprec_at_recall_0 .30 all 0.1940	15. iprec_at_recall_0 .30 all 0.2023
16. iprec_at_recall_0 .40 all 0.1204	16. iprec_at_recall_0 .40 all 0.1226	16. iprec_at_recall_0 .40 all 0.1226
17. iprec_at_recall_0 .50 all 0.0833	17. iprec_at_recall_0 .50 all 0.0856	17. iprec_at_recall_0 .50 all 0.1011
18. iprec_at_recall_0 .60 all 0.0556	18. iprec_at_recall_0 .60 all 0.0556	18. iprec_at_recall_0 .60 all 0.0556
19. iprec_at_recall_0 .70 all 0.0000	19. iprec_at_recall_0 .70 all 0.0000	19. iprec_at_recall_0 .70 all 0.0000
20. iprec_at_recall_0 .80 all 0.0000	20. iprec_at_recall_0 .80 all 0.0000	20. iprec_at_recall_0 .80 all 0.0000

21. iprec_at_recall_0.90 all 0.0000	21. iprec_at_recall_0.90 all 0.0000	21. iprec_at_recall_0.90 all 0.0000
22. iprec_at_recall_1.00 all 0.0000	22. iprec_at_recall_1.00 all 0.0000	22. iprec_at_recall_1.00 all 0.0000
23. P_5 all 0.2444	23. P_5 all 0.2444	23. P_5 all 0.2444
24. P_10 all 0.1667	24. P_10 all 0.1667	24. P_10 all 0.1667
25. P_15 all 0.1370	25. P_15 all 0.1370	25. P_15 all 0.1370
26. P_20 all 0.1167	26. P_20 all 0.1167	26. P_20 all 0.1167
27. P_30 all 0.0778	27. P_30 all 0.0796	27. P_30 all 0.0796
28. P_100 all 0.0233	28. P_100 all 0.0239	28. P_100 all 0.0278
29. P_200 all 0.0117	29. P_200 all 0.0119	29. P_200 all 0.0139
30. P_500 all 0.0047	30. P_500 all 0.0048	30. P_500 all 0.0056
31. P_1000 all 0.0023	31. P_1000 all 0.0024	31. P_1000 all 0.0028
32. runid all STANDARD	32. runid all STANDARD	32. runid all STANDARD
33. num_q all 18	33. num_q all 18	33. num_q all 18