Συστήματα Ανάκτησης – Φάση 1

ΑΝΑΦΟΡΑ ΠΑΡΑΔΟΣΗΣ

(**Όνομα:** Αναστάσιος Παπαπαναγιώτου, **ΑΜ:** 3200143, **Email:** p3200143@aueb.gr)

(**Όνομα:** Φοίβος Παπαθανασίου, **ΑΜ:** 3200138, **Email:** p3200138@aueb.gr)

Βήματα 1, 2

Η μέθοδος **buildIndex()** της κλάσης Index διαβάζει το αρχείο που περιέχει τα κείμενα (documents.txt) και μετατρέπει κάθε κείμενο σε ένα document με δύο fields – body και id, το id χρειάζεται ώστε να μπορέσουμε να κατασκευάσουμε το αρχείο ir_results_file για το trec_eval.

Βήμα 3

Η μέθοδος **getDocsForQueries(...)** της κλάσης Index επιστρέφει ένα ArrayList που περιέχει για κάθε query από ένα queries.txt με το δοθέντο format τα top k documents που έγιναν retrieve.

Πιο συγκεκριμένα, αρχικά κατασκευάζουμε ένα αντικείμενο της κλάσης IndexSearcher σεττάρωντας το similarity σε ClassicSimilarity, κατασκευάζουμε επίσης ένα αντικείμενο της κλάσης QueryParser μόνο για το field “body” και του περνάμε τον EnglishAnalyzer.

Στη συνέχεια κάνουμε extract τα queries από το queries.txt με χρήση της βοηθητικής συνάρτησης readFile της κλάσης Utils και για κάθε query κάνουμε retrieve τα top k documents τα οποία είναι αντικείμενα της κλάσης ScoreDoc.

Εκτός όμως από το score τους χρειαζόμαστε και το αντίστοιχο id τους προκειμένου να κατασκευάσουμε το αρχείο ir_results_file. Για το λόγο αυτό στη συνέχεια για κάθε ένα από αυτά τα queries και για καθένα από τα hits τους παίρνουμε το αντίστοιχο Document instance (και όχι ScoreDoc instance), και με χρήση μίας βοηθητικής κλάσης **QueriedDoc** μαζεύουμε και το Document instance και το score ώστε να μπορέσουμε μετά να παράξουμε το αναγκαίο αρχείο.

Βήμα 4

Λαμβάνωντας το ArrayList από τη μέθοδο **getDocsForQueries(...)** της κλάσης Index, κατασκευάζουμε ένα αντικείμενο της κλάσης Evaluation και χρησιμοποιούμε τη μέθοδο **storeResults(...)** για την κατασκευή του αρχείου ir_results_file προκειμένου να αξιολογήσουμε μετά τις απαντήσεις με τη χρήση του trec_eval.

Τοποθετούμε στο **ίδιο directory** όπου κάναμε extract το trec_eval.zip τα δύο αρχεία **qrels.txt** και **ir_results_file.txt**. Τρέχουμε το trec_eval.exe, στη συνέχεια ανοίγουμε terminal στο directory και τρέχουμε την παρακάτω εντολή για να βρούμε το MAP.

```
PS C:\trec_eval> .\trec_eval -q -m map qrels.txt ir_results_file.txt
      2 [main] trec_eval 21308 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer.  Please report this problem to
the public mailing list cygwin@cygwin.com
map                     Q01     0.7602
map                     Q02     0.2872
map                     Q03     0.6106
map                     Q04     0.0534
map                     Q05     0.2946
map                     Q06     0.0425
map                     Q07     0.2712
map                     Q08     0.8822
map                     Q09     0.1370
map                     Q10     0.3183
map                     all     0.3657
```

Για το Precision@k γράφουμε την παρακάτω εντολή (π.χ. για k=5):

```
PS C:\trec_eval> .\trec_eval -q -m P.5 qrels.txt ir_results_file.txt
      1 [main] trec_eval 4148 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer.  Please report this problem to
the public mailing list cygwin@cygwin.com
P_5                     Q01     0.8000
P_5                     Q02     0.6000
P_5                     Q03     1.0000
P_5                     Q04     0.0000
P_5                     Q05     0.4000
P_5                     Q06     0.0000
P_5                     Q07     0.2000
P_5                     Q08     1.0000
P_5                     Q09     0.4000
P_5                     Q10     0.6000
P_5                     all     0.5000
```

Ο παρακάτω πίνακας παρουσιάζει τα αποτελέσματα που βρήκαμε.

Q	MAP	Precision@5	Precision@10	Precision@15	Precision@20
Q01	0.7602	0.8	0.9	0.8	0.65
Q02	0.2872	0.6	0.4	0.2667	0.2
Q03	0.6106	1	0.6	0.4667	0.4
Q04	0.0534	0	0.1	0.2	0.15
Q05	0.2946	0.4	0.2	0.4667	0.35
Q06	0.0425	0	0.1	0.2	0.15
Q07	0.2712	0.2	0.2	0.4	0.4
Q08	0.8822	1	1	0.8	0.6
Q09	0.137	0.4	0.3	0.2667	0.2
Q10	0.3183	0.6	0.4	0.2667	0.2
all	0.3657	0.5	0.42	0.4133	0.33

Παρατηρούμε ότι το μέσο precision είναι χαμηλό και μάλιστα καθώς το k αυξάνεται μειώνεται ακόμα περισσότερο.

Πηγές: eclass, Lucene's documentation