Συστήματα ανάκτησης πληροφοριών

Προγραμματιστική εργασία, φάση 2

Δημήτρης Μπάστας, 3130139

Φίλιππος Δουραχαλής, 3170045

Η εκπόνηση της εργασίας πραγματοποιήθηκε στο περιβάλλον Intellij IDEA και PyCharm.

Ένα τμήμα της υλοποίησης στηρίχθηκε σε κώδικα java, παρόμοιο με αυτόν της προηγούμενης φάσης, ενώ οι υπολογισμοί του SVD και της ομοιότητας έγιναν με τη χρήση της Python η οποία παρέχει τις απαραίτητες βιβλιοθήκες για αυτήν την δουλειά.

α) Διαμόρφωση του κώδικα JAVA

Όπως και στη προηγούμενη φάση της εργασίας, φορτώνουμε τις απαραίτητες βιβλιοθήκες της Lucene που αναφέρονται στη προηγούμενη αναφορά στο Intellij. Ωστόσο αυτή τη φορά χρειάζεται να συμπεριλάβουμε ΚΑΙ τη βιβλιοθήκη lucene-classification-7.7.3.jar ώστε να μπορέσουμε να αναπαραστήσουμε τα κείμενα ως διανύσματα.

Αφού φορτώσουμε τις βιβλιοθήκες πρέπει να προσδιορίσουμε τις εξής τοποθεσίες:

1. Την τοποθεσία του φακέλου trec_eval (συαθερά *TREC_EVAL_PATH*)
2. Τη θέση του αρχείου με τη συλλογή των κειμένων και του αρχείου των queries (σταθερά *CACM_PATH*)
3. Την τοποθεσία που θα αποθηκευτεί το index μας

Ο κώδικας της java μοιάζει αρκετά με της πρώτης φάσης, με τη διαφορά ότι πλέον δεν χρησιμοποιείται για την παραγωγή των αποτελεσμάτων, παρά μόνο για την δημιουργία των πινάκων.

Η εκτέλεση του κώδικα παράγει τους πίνακες term x document και term x query οι οποίοι αναπαρίστανται σαν διανύσματα δύο διαστάσεων και αποθηκεύονται στην τοποθεσία που έχουμε ορίσει μέσω του path trev_eval_path.

β) Υπολογισμός ομοιότητας με Python

Σε αυτήν την φάση χρησιμοποιούμε δύο αρχεία: Ένα για τον υπολογισμό του SVD και ένα για τον υπολογισμό της ομοιότητας καθενός query με κάθε από τα κείμενα.

1. Στο αρχείο που υπολογίζεται το SVD (svd.py) αρχικά ορίζουμε μια global μεταβλητή για το path των αρχείων που δημιούργησε η java. Αυτή η τοποθεσία ορίζεται μέσω της *VECTORS_PATH.*
   Η αμέσως επόμενη μεταβλητή x χρησιμοποιείται για να ορίσει τις διαστάσεις (τάξη) της προσέγγισης A_k.

2. Το αρχείο querychecker.py χρησιμοποιείται για τον υπολογισμό της ομοιότητας. Σε αυτό διαβάζουμε το διάνυσμα των ερωτημάτων, υπολογίζουμε την προσέγγιση A_k και στη συνέχεια βγάζουμε την ομοιότητα μέσω του τύπου $\frac{A_k * Q}{\|A_k\| * \|Q\|}$, όπου Q είναι το διάνυσμα που αναπαριστά ένα ερώτημα. Στη συνέχεια ταξινομούμε τα αποτελέσματα και τα γράφουμε σε ένα αρχείο results.txt με συγκεκριμένη μορφοποίηση ώστε να δοθεί στο trec_eval.
   Εδώ σημειώνουμε ότι έχουμε υλοποιήσει δύο τρόπους για τον υπολογισμό της ομοιότητας. Ο πρώτος είναι αυτός που αναφέρθηκε, ενώ ο δεύτερος κάνει χρήση του V_k αντί του A. Παρατηρήσαμε ωστόσο ότι ο πρώτος παράγει καλύτερα αποτελέσματα, άρα ο πίνακας που φαίνεται παρακάτω έχει δημιουργηθεί βάσει αυτού. Ομοίως δοκιμάσαμε να χρησιμοποιήσουμε και την συνημιτονοειδή ομοιότητα προτού καταλήξουμε στον τύπο που αναφέρθηκε ανωτέρω, όμως και πάλι παρατηρήσαμε ότι αυτή δεν δίνει τόσο καλά αποτελέσματα. Ότι δεν χρησιμοποιείται για τον τελικό υπολογισμό έχει κρατηθεί σε σχόλια ώστε να μπορούν γρήγορα να γίνουν επιπλέον δοκιμές και να επαληθευτούν τα αποτελέσματα.

Χρησιμοποιώντας το trec_eval με την εντολή “trec_eval [-q, -m] qrels.text results.text > (αρχείο eval)” βρήκαμε τα Map και P_k ως ορίζονται στον κάτω πίνακα. (Τα πλήρη αποτελέσματα συμπεριλαμβάνονται στα αντίστοιχα αρχεία map50, map150 και map300)

Τάξη	50	150	300
MAP	0.0029	0.0039	0.0043
P_5	0.0047	0.0093	0.0093
P_10	0.0070	0.0116	0.0093
P_15	0.0109	0.0140	0.0155
P_20	0.0128	0.0151	0.0151

Τάξη 50:

```
C:\Users\Philip\Downloads\trec_eval>trec_eval qrels.txt results.txt
      1 [main] trec_eval 8620 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer.  Please report this problem to
the public mailing list cygwin@cygwin.com
runid                   all     1
num_q                   all     43
num_ret                 all     2150
num_rel                 all     719
num_rel_ret             all     38
map                     all     0.0029
gm_map                  all     0.0001
Rprec                   all     0.0148
bpref                   all     0.0404
recip_rank              all     0.0301
iprec_at_recall_0.00    all     0.0376
iprec_at_recall_0.10    all     0.0137
iprec_at_recall_0.20    all     0.0010
iprec_at_recall_0.30    all     0.0000
iprec_at_recall_0.40    all     0.0000
iprec_at_recall_0.50    all     0.0000
iprec_at_recall_0.60    all     0.0000
iprec_at_recall_0.70    all     0.0000
iprec_at_recall_0.80    all     0.0000
iprec_at_recall_0.90    all     0.0000
iprec_at_recall_1.00    all     0.0000
P_5                     all     0.0047
P_10                    all     0.0070
P_15                    all     0.0109
P_20                    all     0.0128
P_30                    all     0.0163
P_100                   all     0.0088
P_200                   all     0.0044
P_500                   all     0.0018
P_1000                  all     0.0009
```

Τάξη 150:

```
C:\Users\Philip\Downloads\trec_eval>trec_eval qrels.txt results.txt
      1 [main] trec_eval 10744 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer.  Please report this problem to
the public mailing list cygwin@cygwin.com
runid                   all     1
num_q                   all     43
num_ret                 all     2150
num_rel                 all     719
num_rel_ret             all     38
map                     all     0.0039
gm_map                  all     0.0001
Rprec                   all     0.0181
bpref                   all     0.0408
recip_rank              all     0.0398
iprec_at_recall_0.00    all     0.0441
iprec_at_recall_0.10    all     0.0168
iprec_at_recall_0.20    all     0.0032
iprec_at_recall_0.30    all     0.0000
iprec_at_recall_0.40    all     0.0000
iprec_at_recall_0.50    all     0.0000
iprec_at_recall_0.60    all     0.0000
iprec_at_recall_0.70    all     0.0000
iprec_at_recall_0.80    all     0.0000
iprec_at_recall_0.90    all     0.0000
iprec_at_recall_1.00    all     0.0000
P_5                     all     0.0093
P_10                    all     0.0116
P_15                    all     0.0140
P_20                    all     0.0151
P_30                    all     0.0155
P_100                   all     0.0088
P_200                   all     0.0044
P_500                   all     0.0018
P_1000                  all     0.0009
```

Τάξη 300:

```
C:\Users\Philip\Downloads\trec_eval>trec_eval qrels.txt results.txt
     1 [main] trec_eval 17716 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer.  Please report this problem to
the public mailing list cygwin@cygwin.com
runid                   all     1
num_q                   all     43
num_ret                 all     2150
num_rel                 all     719
num_rel_ret             all     41
map                     all     0.0043
gm_map                  all     0.0002
Rprec                   all     0.0178
bpref                   all     0.0674
recip_rank              all     0.0380
iprec_at_recall_0.00    all     0.0446
iprec_at_recall_0.10    all     0.0206
iprec_at_recall_0.20    all     0.0031
iprec_at_recall_0.30    all     0.0005
iprec_at_recall_0.40    all     0.0005
iprec_at_recall_0.50    all     0.0005
iprec_at_recall_0.60    all     0.0005
iprec_at_recall_0.70    all     0.0005
iprec_at_recall_0.80    all     0.0005
iprec_at_recall_0.90    all     0.0005
iprec_at_recall_1.00    all     0.0005
P_5                     all     0.0093
P_10                    all     0.0093
P_15                    all     0.0155
P_20                    all     0.0151
P_30                    all     0.0155
P_100                   all     0.0095
P_200                   all     0.0048
P_500                   all     0.0019
```