Ανάκτηση Πληροφοριών και Εξόρυξη Δεδομένων
2023-2024

1η Εργασία --- 19/03/2024 v 0.97
Ανάκτηση Κειμένων, Φράσεις και Συνώνυμα

Η εργασία είναι ομαδική, με ομάδες των δύο φοιτητών αποκλειστικά.
Σχηματίστε και δηλώστε άμεσα ομάδες στο eClass.

Μέρος Α

Με χρήση της μηχανή αναζήτησης Terrier, κατασκευάστε ένα ευρετήριο με τη μέθοδο Classical two-pass indexing για τις συλλογές κειμένων που βρίσκονται στους φακέλους: fbis, fr94, ft, latimes. Για τη δημιουργία του ευρετηρίου, χρησιμοποιήστε τον ενσωματωμένο Porter stemming και την λίστα stopwords που παρέχεται από το Terrier. Για την κατασκευή ευρετηρίου, σας δίνεται ένα αρχείο παράδειγμα το οποίο πρέπει προσαρμόσετε στο περιβάλλον σας (π.χ να αλλάξετε τα paths κλπ).

Στα 3 αρχεία topics.XXX-XXX.trecX σας δίνονται συνολικά 150 ανάγκες πληροφορίας (information needs / topics). Σχηματίστε και τρέξτε 150 ερωτήσεις (queries), με τρεις διαφορετικούς τρόπους χρησιμοποιώντας τα πεδία title, title+desc, title+desc+narr. Για το πεδίο title, σας δίνεται ένα αρχείο παράδειγμα που χρησιμοποιεί το προεπιλεγμένο (default) μοντέλο ανάκτησης του Terrier με τις προεπιλεγμένες του παραμέτρους (`matching.retrieved_set_size = 1000`---τρέξτε όλα τα πειράματά σας σε αυτήν την εργασία με αυτήν την παράμετρο όπως σας δίνεται). Για την μορφοποίηση των λιστών των αποτελεσμάτων, χρησιμοποιήστε το trec format (υποστηρίζεται στο Terrier) με το πολύ 1,000 αποτελέσματα ανά ερώτηση.

Στα 3 αρχεία qrels.XXX-XXX.trecX.adhoc δίνονται οι κρίσεις συνάφειας (relevance judgments) των παραπάνω ερωτήσεων αναφορικά με τις συλλογές. Αξιολογήστε την ποιότητα ανάκτησης με τους αριθμητικούς μέσους όρους των μετρικών Precision@10, R-Precision, και Average Precision (AP) χρησιμοποιώντας το πρόγραμμα/εργαλείο trec_eval. Σχολιάστε τα αποτελέσματα και καταλήξτε σε συμπεράσματα.

Τα παραπάνω αποτελούν τα baseline runs, των οποίων την ποιότητα ανάκτησης θα προσπαθήσετε παρακάτω να βελτιώσετε.

Μέρος Β

Επιλέξτε μία από τις παρακάτω τέσσερις κατευθύνσεις και ενημερώστε τον κ. Αραμπατζή Γεώργιο με email (geoaramp@ee.duth.gr) για επιβεβαίωση. Περίπου ο ίδιος αριθμός από ομάδες πρέπει να ακολουθήσουν την κάθε κατεύθυνση, οπότε οι επιλογές θα γίνουν με σειρά προτεραιότητας (first-come-first-served).

Κατεύθυνση Β.1 – Επέκταση Ερωτήματος με Θησαυρό Λέξεων

Χρησιμοποιήστε το WordNet για να εμπλουτίστε τα ερωτήματα με συνώνυμα. Τρέξτε τις εμπλουτισμένες ερωτήσεις και αξιολογήστε τις μεθόδους σας με τις μετρικές του Μέρους Α. Παρουσιάστε τα αποτελέσματα σε μορφή πίνακα, συγκρίνετε και σχολιάστε τα καταλήγοντας σε συμπεράσματα.

Κατεύθυνση Β.2 – Επέκταση Ερωτήματος με Φράσεις

Βρείτε έναν αριθμό από καλές στατιστικές φράσεις (πχ bi-words) επεξεργάζοντας τη συλλογή, και εμπλουτίστε τα ερωτήματα. Τρέξτε τις εμπλουτισμένες ερωτήσεις και αξιολογήστε τις μεθόδους σας με τις μετρικές του Μέρους Α. Παρουσιάστε τα αποτελέσματα σε μορφή πίνακα, συγκρίνετε και σχολιάστε τα καταλήγοντας σε συμπεράσματα.

Κατεύθυνση Β.3 – Επέκταση Ερωτήματος με Διαχωρισμό των Ερμηνειών των λέξεων

Χρησιμοποιήστε το WordNet και εφαρμόστε τον αλγόριθμο του Lesk, για τον διαχωρισμό των ερμηνειών των λέξεων, για να εμπλουτίστε τα ερωτήματα με συνώνυμα. Τρέξτε τις εμπλουτισμένες ερωτήσεις και αξιολογήστε τις μεθόδους σας με τις μετρικές του Μέρους Α. Παρουσιάστε τα αποτελέσματα σε μορφή πίνακα, συγκρίνετε και σχολιάστε τα καταλήγοντας σε συμπεράσματα.

Κατεύθυνση Β.4 – Επέκταση Ερωτήματος με Ανάδραση Σχετικότητας με Όρους της Συλλόγης και επέκταση με θησαυρό λέξεων

Χρησιμοποιήστε το μοντέλο Ανάδρασης Σχετικότητας της μηχανής αναζήτησης Terrier για να επεκτείνεται τα αρχικά ερωτήματα με είκοσι (20) όρους από τα δεκαπέντε (15) πρώτα κείμενα της αρχικής κατάταξης. Στην συνέχεια χρησιμοποιήστε το WordNet για να εμπλουτίστε τα νέα ερωτήματα (αρχικοί και νέοι όροι)με συνώνυμα. Τρέξτε τις εμπλουτισμένες ερωτήσεις (αρχικοί + νέοι όροι, αρχικοί + νέοι όροι + συνώνυμα) και αξιολογήστε τις μεθόδους σας με τις μετρικές του Μέρους Α. Παρουσιάστε τα αποτελέσματα σε μορφή πίνακα, συγκρίνετε και σχολιάστε τα καταλήγοντας σε συμπεράσματα.

Γενικά:
Γράψτε ότι ενδιάμεσα μικρά προγράμματα ή scripts θα (και αν) σας χρειαστούν για την προ-επεξεργασία των δεδομένων και αποτελεσμάτων, σε οτιδήποτε γλώσσα θέλετε. Χρησιμοποιήστε όποια πλατφόρμα θέλετε (Linux, Windows, MacOS, κ.α.), αφού πρώτα ενημερωθείτε για το τι δυνατότητες σας προσφέρουν τα προαναφερόμενα εργαλεία---πρέπει να είστε δημιουργικοί και ευέλικτοι.

Παράδοση και Εξέταση: Τρίτη 14 Μαΐου
Η παράδοση και εξέταση θα γίνει ανά ομάδα. Το παραδοτέο περιλαμβάνει την παρουσίαση των αποτελεσμάτων αξιολόγησης με την ανάλυση και τα σχόλιά σας, σε μορφή PDF. Επίσης, να περιλάβετε σε ξεχωριστά αρχεία τα καλύτερα δύο runs σας του Μέρους Α και τα καλύτερα δύο του Μέρους Β. Θα πρέπει να ανεβάσετε το παραδοτέο στο e-class μέχρι τις 23:59 της 13 Μαΐου.

Η εξέταση θα πραγματοποιηθεί κατά την διάρκεια των ωρών της Θεωρίας (Τρίτη 14/05/24 και ώρα 17:15-20:00, στο γραφείο του κ. Αραμπατζή Γεώργιου, στο ισόγειο στα γραφεία του Τομέα Λογισμικού) και περιλαμβάνει επίδειξη του setup σας (πχ, τρέξιμο ερωτήσεων). Στην ακόλουθη διάλεξη θα γίνει συζήτηση με όλες τις ομάδες και συγκριτική παρουσίαση των μεθόδων και αποτελεσμάτων.

Πόροι:

- Terrier: http://terrier.org/
- Java: https://www.oracle.com/java/technologies/downloads/
- Οδηγίες εγκατάστασης και χρήσης του Terrier: http://terrier.org/download/
- Τα δεδομένα σας: http://lethe.nonrelevant.net/datasets/IR-2023-2024-Project-1.zip
(το συμπιεσμένο αρχείο έχει password---θα σας το δώσει ο κ. Αραμπατζής Γεώργιος στο εργαστήριο)
- WordNet: http://wordnet.princeton.edu/
- Η τελευταία έκδοση του trec_eval: http://trec.nist.gov/trec_eval/
- Video Tutorial: https://drive.google.com/file/d/1i5bW7Gs0SuJHhRRA_ULZnsXiLaBaXPp2/view?usp=sharing

Σχετική ύλη από το βιβλίο "Introduction to Information Retrieval":

- Ch.1 (εκτός από τους αλγόριθμους των Figures 1.6 και 1.7)
- Ch.2 (εκτός από το Section 2.3 και τον αλγόριθμο του Figure 2.12)
- Ch.5 (μόνο το Section 5.1)
- Ch.6 (εκτός από το Section 6.1)
- Ch.8
- Ch.9 (εκτός από το Section 9.1.2)

Ο Διδάσκων