Ανάκτηση Πληροφορίας
Προγραμματιστική Άσκηση 1

Διδάσκων:
Χ. Τρυφωνόπουλος

Παράδοση μέχρι: Παρασκευή 21/12/2018 ώρα 23.59
Προσωπική εξέταση: στο τέλος του εξαμήνου

ΣΗΜΑΝΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ:

1. Αφού έχετε ολοκληρώσει την άσκηση που θέλετε να παραδώσετε, την υποβάλετε στο eclass στο υποσύστημα «Εργασίες». Η υποβολή πρέπει να γίνει ΠΡΙΝ την ημερομηνία παράδοσης. Παραδίδετε όλα τα απαραίτητα αρχεία σε ένα zip (κώδικας, εκτελέσιμο, συνοδευτικά αρχεία, και αναφορά).

2. Η άσκηση συνίσταται να υλοποιηθεί από ομάδα δύο ατόμων, αλλά μπορεί να υλοποιηθεί και ατομικά. Περιπτώσεις αντιγραφής θα μηδενίζονται και το ζήτημα θα πηγαίνει στη συνέλευση Τμήματος. Η ημερομηνία παράδοσης δεν αλλάζει, και η παράδοση γίνεται μόνο μέσω του eclass και όχι με email στον διδάσκοντα. Εκπρόθεσμες ασκήσεις δεν θα γίνονται δεκτές για κανένα λόγο.

Στην άσκηση αυτή καλείστε να υλοποιήσετε τη μηχανή αναζήτησης RASTA (Researcher Automated Search & Text Analytics), η οποία θα ευρετηριάζει και θα αναζητά ερευνητές με βάση τα ερευνητικά τους ενδιαφέροντα και το δημοσιευμένο τους έργο. Η υλοποίησή σας θα βασιστεί στη βιβλιοθήκη Apache Lucene, η οποία προσφέρει βασικές λειτουργίες Ανάκτησης Πληροφορίας, όπως οργάνωση της συλλογής κειμένων με τη δημιουργία ευρετήριου, αναζήτηση κειμένων με ερωτήσεις εκφρασμένες σε διάφορα μοντέλα (όπως Boolean, Vector Space, φράσεων, κλπ), και κατάταξη των σχετικών αποτελεσμάτων. Η βιβλιοθήκη Lucene είναι ένα από τα πιο δημοφιλή projects της Apache και χρησιμοποιείται ευρέως για την ανάπτυξη εφαρμογών που απαιτούν λειτουργίες ανάκτησης κειμένων.

Ως γλώσσα υλοποίησης προτείνεται η Java για λόγους ευκολίας και συμβατότητας με τη βιβλιοθήκη Apache Lucene. Μαζί με τον πηγαίο κώδικα του προγράμματός σας θα παραδώσετε και τυχόν συνοδευτικά αρχεία, το εκτελέσιμο αρχείο, και μια αναφορά. Έμφαση θα δοθεί στην κομψότητα της υλοποίησης, στην ταχύτητα εκτέλεσης των ζητούμενων, στη χρηστικότητα και ορθότητα του προγράμματός σας, και στην πληρότητα και σαφήνεια των λοιπών παραδοτέων. ΜΗΝ συμπεριλάβετε στην αναφορά σας μέρος ή σύνολο του κώδικα!

Περιγραφή του προβλήματος και των δεδομένων

Στο πρόβλημα που έχετε να αντιμετωπίσετε, κάθε ερευνητής έχει ένα σύνολο δημοσιεύσεων, οι οποίες έχουν μία συγκεκριμένη δομή που χρησιμοποιείται από τον κειμενογράφο Latex για να χαρακτηρίσει τα διάφορα μέρη μιας δημοσίευσης. Κάθε δημοσίευση ξεκινά με το σύμβολο "@", το οποίο ακολουθείται από ένα αλφαριθμητικό που περιγράφει το είδος της δημοσίευσης (π.χ., inproceedings για άρθρα σε πρακτικά συνεδρίων, article για άρθρα σε περιοδικά, κλπ) και το άνοιγμα αγκύλης "{", ενώ τελειώνει με το κλείσιμο της αγκύλης "}". Σε κάθε δημοσίευση περιέχονται πεδία που περιγράφουν τα επιμέρους στοιχεία της. Από τα πεδία αυτά, για τη συγκεκριμένη άσκηση, θα χρειαστούμε μόνο το title, το οποίο μας δίνει τον τίτλο μιας δημοσίευσης, και το booktitle ή journal (ανάλογα με το είδος της δημοσίευσης), το οποίο μας δίνει τον τίτλο του συνεδρίου ή περιοδικού.

Στην προσέγγισή σας, κάθε ερευνητής θα θεωρείται ένα αντικείμενο, το οποίο θα περιέχει ως κειμενική πληροφορία το σύνολο των τίτλων των δημοσιεύσεών του και των ονομάτων των συνεδρίων στα οποία έχουν γίνει δεκτές οι δημοσιεύσεις αυτές. Οπότε στη μηχανή αναζήτησης RASTA κάθε ερευνητής είναι σαν ένα «κείμενο» της συλλογής, το οποίο περιέχει προτάσεις με τους τίτλους των άρθρων του και των αντίστοιχων συνεδρίων/περιοδικών. Προφανώς θα πρέπει να υπάρχει τρόπος κάποιος να προσθέσει, να αφαιρέσει, ή να τροποποιήσει ερευνητές στη συλλογή που ευρετηριάζει το RASTA. Τα ζητήματα αυτά περιγράφονται αναλυτικά στις παρακάτω ενότητες.

Λειτουργικότητα της μηχανής αναζήτησης RASTA

Η μηχανή αναζήτησης RASTA θα πρέπει να παρέχει τη λειτουργικότητα που περιγράφεται παρακάτω.

Προεπεξεργασία περιεχομένου ερευνητών [10%]

Για κάθε ερευνητή θα παίρνετε ένα αρχείο με το σύνολο των δημοσιεύσεών του με τη δομή που περιγράφηκε στην προηγούμενη ενότητα και θα πρέπει να προεπεξεργαστείτε το αρχείο αυτό ώστε να κρατήσετε μόνο τα χρήσιμα πεδία και να κατασκευάσετε ένα αρχείο κειμένου το οποίο θα περιγράφει τον κάθε ερευνητή. Η διαδικασία αυτή μπορεί να γίνει είτε με κώδικα που θα γράψετε εσείς, είτε με κάποιον parser για .bib αρχεία (έτσι ονομάζεται το συγκεκριμένο format αρχείων) τρίτου, καθώς το Lucene δεν παρέχει έτοιμη τέτοια λειτουργικότητα. Κατόπιν σκεφτείτε αν χρειάζεται κάποιο άλλο είδος προεπεξεργασίας, όπως αν θα παραλείψετε τις λέξεις αποκλεισμού, αν θα μετατρέψετε τις λέξεις σε μικρά/κεφαλαία γράμματα (case folding), αν θα κάνετε λημματοποίηση (stemming), και αν θα σβήσετε τα σημεία στίξης (punctuation removal). Η λειτουργικότητα αυτή παρέχεται από το Lucene, αλλά μπορείτε να χρησιμοποιήσετε και κώδικα δικό σας ή τρίτων, αρκεί να δίνετε την πηγή στη γραπτή αναφορά σας.

Κατασκευή ευρετηρίου και εισαγωγή/διαγραφή κειμένων [30%]

Η μηχανή αναζήτησης RASTA θα πρέπει να παρέχει τη δυνατότητα εισαγωγής και διαγραφής ενός ερευνητή. Η εισαγωγή ενός νέου ερευνητή στη μηχανή αναζήτησης θα πρέπει να γίνεται με δύο τρόπους: (α) μέσα από ένα αρχείο το οποίο θα υποδεικνύει ο χρήστης και θα περιέχει όλες τις δημοσιεύσεις ενός ερευνητή και (β) τραβώντας απευθείας πληροφορία από τη βιβλιογραφική βάση δεδομένων DBLP (dblp.org). Σε κάθε περίπτωση η πληροφορία θα δίνεται σε αρχείο .bib το οποίο θα είναι και το μόνο format αρχείου που θα μπορεί να αναγνώσει το RASTA.

Για να προστεθεί ένας νέος ερευνητής στο RASTA με τον πρώτο τρόπο, ο χρήστης θα επιλέγει ένα αρχείο με κατάλληλα δεδομένα και εσείς στη συνέχεια θα πρέπει να ανοίγετε, να διαβάζετε, να προεπεξεργάζεστε το αρχείο, και να προσθέτετε την πληροφορία για το νέο ερευνητή στα κατάλληλα αντεστραμμένα ευρετήρια του Lucene.

Για να προστεθεί ένας νέος ερευνητής στο RASTA με το δεύτερο τρόπο, ο χρήστης θα δίνει το όνομα και το επώνυμο ενός ερευνητή και εσείς με βάση αυτό θα κατασκευάζετε ένα κατάλληλο URL από το οποίο θα κατεβάζετε το αρχείο με τις δημοσιεύσεις του συγκεκριμένου ερευνητή από το DBLP. Το URL που θα φτιάχνετε θα σας οδηγεί στο σύστημα DBLP και έχει την παρακάτω μορφή:

- Στην αρχή έχει το αλφαριθμητικό http://dblp.org/pers/tb2/,
- το οποίο ακολουθείται από το πρώτο γράμμα του επωνύμου του ερευνητή,
- στη συνέχεια από μία κάθετο /,
- κατόπιν από το επώνυμο και το όνομα του ερευνητή χωρισμένα με άνω κάτω τελεία,
- και τέλος την κατάληξη .bib.

Επομένως, αν ο ερευνητής που θέλετε να προστεθεί είναι ο Christos Tryfonopoulos το URL που θα κατασκευάσετε για να κατεβάσετε τις δημοσιεύσεις του από το DBLP θα είναι το:

http://dblp.org/pers/tb2/t/Tryfonopoulos:Christos.bib

Παρατηρείστε το /t/ λόγω του πρώτου γράμματος τους επωνύμου, καθώς και τη γραφή του ονόματος και του επωνύμου με το πρώτο γράμμα κεφαλαίο. Προσοχή, το URL είναι case sensitive! Αν κάποιος ερευνητής έχει δύο ονόματα τότε αυτά μπαίνουν με underscore ανάμεσά τους, π.χ.,

http://dblp.org/pers/tb2/p/Politi:Christina_Tanya.bib

Αν δοθεί λάθος όνομα ερευνητή, θα πρέπει να χειρίζεστε το λάθος με κατάλληλο μήνυμα. Σημειώστε ότι η εισαγωγή νέων ερευνητών στο RASTA θα πρέπει να συνεπάγεται και ενημέρωση των ευρετηρίων που κρατούνται στο Lucene (ή τη δημιουργία τους αν αυτά είναι κενά). Η κατασκευή του ευρετηρίου αυτού θα επιτρέψει την εκτέλεση των ερωτημάτων που περιγράφονται στην παρακάτω ενότητα. Σημειώστε ότι αν ο χρήστης προσπαθήσει να κάνει εισαγωγή ενός ερευνητή που υπάρχει ήδη στη μηχανή αναζήτησης, τότε αυτό σημαίνει ενημέρωση των στοιχείων. Σε κάθε περίπτωση θεωρείται ότι στην

ενημέρωση δίνεται το σύνολο των δημοσιεύσεων (όχι μόνο οι νέες). Ο απλούστερος τρόπος να χειριστείτε την ενημέρωση είναι η διαγραφή και η εκ νέου εισαγωγή του εν λόγω ερευνητή στο RASTA.

Η διαγραφή των βιβλιογραφικών αντικειμένων θα γίνεται επιλέγοντας έναν ή περισσότερους ερευνητές και δίνοντας κατάλληλη εντολή διαγραφής από το γραφικό περιβάλλον. Στην περίπτωση της διαγραφής θα πρέπει επίσης να γίνεται ενημέρωση των ευρετηρίων.

Αναζήτηση και συνάφεια ερευνητών [30%]

Θα πρέπει να υποστηρίζεται η δυνατότητα αναζήτησης ερευνητών και να παρέχεται στο χρήστη η δυνατότητα διατύπωσης ερωτημάτων Boolean (τελεστές AND, OR, NOT), εγγύτητας, φράσεων, και VSM. Σημειώστε ότι το Lucene υποστηρίζει εγγενώς όλους τους παραπάνω τύπους ερωτήσεων, επομένως για να υποστηρίξετε τέτοια ερωτήματα θα πρέπει απλώς να καλέσετε την κατάλληλη μέθοδο από αυτές που παρέχονται έτοιμες.

Εκτός από την αναζήτηση ερευνητών το σύστημα RASTA θα πρέπει να υποστηρίζει και την εύρεση της συνάφειας μεταξύ ερευνητών. Η συνάφεια θα πρέπει να υπολογίζεται ως η ομοιότητα συνημιτόνου μεταξύ των διανυσμάτων που περιγράφουν τα κείμενα κάθε ζεύγους ερευνητών τους. Οπότε, για τη λειτουργία αυτή ο χρήστης α) επιλέγει ένα ζεύγος ερευνητών και βλέπει την ομοιότητά τους, ή β) επιλέγει έναν ερευνητή και το σύστημα φέρνει ως απαντήσεις μια λίστα με τους top-K πιο συναφείς ερευνητές (ανάλογα με το K που έδωσε ή έχει θέσει ως ρύθμιση ο χρήστης).

Γραφικό περιβάλλον χρήσης και προβολή αποτελεσμάτων [20%]

Το RASTA θα πρέπει να προβάλει τους ερευνητές που ανακτήθηκαν συνεπεία ενός ερωτήματος κατά σειρά σχετικότητας (από τον πιο σχετικό προς τον λιγότερο σχετικό) μαζί με το σκορ σχετικότητας, και ένα μικρό απόσπασμα από το κείμενο που περιέχει τις λέξεις κλειδιά που υπέβαλλε ο χρήστης (κατά τη συνήθη πρακτική των μηχανών αναζήτησης). Η υλοποίησή σας θα πρέπει να δίνει στο χρήστη τη δυνατότητα να περιορίσει τον αριθμό των αποτελεσμάτων που θα λάβει για κάθε ερώτημα στα K πρώτα/πιο σχετικά (να εμφανίζει δηλαδή μόνο τα top-K αποτελέσματα ανάλογα με το K που έδωσε ή έχει θέσει ως ρύθμιση ο χρήστης). Σημειώστε ότι το Lucene, για κάθε ερώτημα που υποβάλλεται υπολογίζει και επιστρέφει αυτόματα τα σκορ αυτά, επομένως η δική σας δουλειά είναι κυρίως στην σωστή παρουσίαση των αποτελεσμάτων και όχι στον υπολογισμό των σκορ.

Τέλος, φροντίστε το γραφικό περιβάλλον που θα σχεδιάσετε να είναι απλό και κατανοητό, να δίνει στο χρήστη τη δυνατότητα να υποβάλλει με έναν εύκολο τρόπο το ερώτημά του και να βρει τη συνάφεια μεταξύ δύο ερευνητών, ενώ θα πρέπει να έχει έναν εύληπτο τρόπο παρουσίασης των αποτελεσμάτων.

Αναφορά [10%]

Η αναφορά σας θα πρέπει να έχει έκταση τουλάχιστον πέντε σελίδες χωρίς το εξώφυλλο και τις αναφορές/παραπομπές και θα πρέπει να περιέχει λεπτομέρειες της υλοποίησής σας, οδηγίες εκτέλεσης του κώδικά σας, παραδείγματα επίδειξης του προγράμματός σας, και επεξήγηση των βασικών στοιχείων του προγράμματός σας (user manual). ΜΗΝ συμπεριλάβετε στην αναφορά σας μέρος ή σύνολο του κώδικα!

Θέματα υλοποίησης και bonus

Μπορείτε στην υλοποίησή σας να χρησιμοποιήσετε κώδικα τρίτων ή έτοιμες βιβλιοθήκες (και εκτός αυτών που παρέχονται από το Lucene) αρκεί να περιλαμβάνετε κατάλληλη παραπομπή της πηγής σας στην αναφορά.

Σε συνεννόηση με το διδάσκοντα μπορείτε να πάρετε μέχρι 10% bonus για διάφορα επιπλέον χαρακτηριστικά ή λειτουργικότητα που θα υλοποιήσετε. Τέτοια μπορεί να είναι η υλοποίηση του προγράμματος ως web εφαρμογη, η υποστήριξη διαφορετικών τύπων αρχείων βιβλιογραφίας (π.χ., RIS – χρησιμοποιείται σε βιβλιογραφικές βάσεις δεδομένων, RDF, XML, κλπ.) που παρέχονται από το DBLP, ή μία δική σας πρόταση βελτίωσης.

Καλή δουλειά!