Αναγνώριση Προτύπων

2η Εργασία Μαθήματος – Ομαδοποίηση

Σκοπός της Εργασίας:
Στο πλαίσιο της παρούσας εργασίας, θα κληθείτε να εφαρμόσετε τεχνικές ομαδοποίησης για να κατανοήσετε τα χαρακτηριστικά που εξάγονται από τον κώδικα γνωστών βιβλιοθηκών προγραμματισμού.

Πρόβλημα:
Η ομαδοποίηση και η εύρεση παρόμοιων κειμένων αποτελεί ένα πολύ ενδιαφέρον πρόβλημα της ανάκτησης δεδομένων. Στο πλαίσιο της παρούσας εργασίας, τα κείμενα αντιστοιχούν σε κώδικα γνωστών βιβλιοθηκών ενώ οι ομάδες (clusters) που αναμένονται είναι οι κατηγορίες των βιβλιοθηκών αυτών. Το dataset που σας δίνεται περιέχει τις λέξεις που εμφανίζονται σε 80 βιβλιοθήκες της Java. Για κάθε λέξη (που αποτελεί feature), η κάθε βιβλιοθήκη (sample) έχει μια τιμή που αντιστοιχεί στη συχνότητα της λέξης στον κώδικα της βιβλιοθήκης. Ζητείται από εσάς να ομαδοποιήσετε τις βιβλιοθήκες σε κατηγορίες με βάση το σκοπό τους (π.χ. testing-frameworks, json-libraries, android, κτλ.). Σημειώστε ότι ο διαχωρισμός σε κατηγορίες δίνεται στη στήλη category, η οποία όμως δίνεται για τον υπολογισμό μετρικών ποιότητας της ομαδοποίησης, και **όχι για να τη χρησιμοποιήσετε στην ομαδοποίηση που θα κάνετε**. Σαφείς οδηγίες για τα δεδομένα δίνονται στο αρχείο Dataset.pdf.

Διαδικασία:
Ακολουθήστε την παρακάτω διαδικασία:

Βήμα 1. Εξερευνήστε τα δεδομένα και κάντε οποιοσδήποτε μορφής καθαρισμό/κανονικοποίηση θεωρείτε απαραίτητο. Μπορείτε επίσης να κάνετε συνάθροιση των δεδομένων σας σε όποιο επίπεδο λεπτομέρειας θέλετε, να δημιουργήσετε ζώνες δεδομένων και οτιδήποτε θεωρείτε ότι μπορεί να σας βοηθήσει.

Βήμα 2. Επιλέξτε τουλάχιστον δύο τεχνικές ομαδοποίησης (Ιεραρχικοί, Πυκνωτικοί, Διαχωρισμού, SOMs), και εφαρμόστε τουλάχιστον τρία σετ παραμέτρων όπου αυτό είναι δυνατό. **Δικαιολογήστε την επιλογή σας.**

Βήμα 3. Κάντε αποτίμηση της ποιότητας της ομαδοποίησής σας με βάση τις εσωτερικές μετρικές αξιολόγησης (SSE, Cohesion, Separation, Silhouette).

Βήμα 4. Κάντε μια γενική αποτίμηση της ομαδοποίησης με βάση τα δεδομένα που δίνονται. Σχολιάστε κατά πόσο η ομαδοποίηση είναι λογική, και σχολιάστε κατά περίπτωση τις ομάδες και τις βιβλιοθήκες που παρουσιάζουν ενδιαφέρον.

Χρήσιμες παρατηρήσεις:

1. Είναι ιδιαίτερα σημαντικό να γίνει σωστή προ-επεξεργασία για να έχουν νόημα και τα αποτελέσματα της ανάλυσης. Πάρτε υπόψη σας τα παρακάτω:
    a. Όταν μια λέξη εμφανίζεται σε πολλές βιβλιοθήκες, ενδέχεται να είναι πολύ «γενική», π.χ. java ή util, οπότε είναι πιθανό να μην είναι ιδιαίτερα χρήσιμη για την ομαδοποίηση.
    b. Όταν μια λέξη εμφανίζεται σε πολύ λίγες βιβλιοθήκες, ενδέχεται να είναι πολύ «ειδική», π.χ. ThisIsMyClassThatIOnlyUseInMyLibrary, οπότε είναι επίσης πιθανό να μην είναι ιδιαίτερα χρήσιμη για την ομαδοποίηση.

c. Οι βιβλιοθήκες μπορεί να έχουν αρκετά διαφορετικό μέγεθος, π.χ. μια βιβλιοθήκη μπορεί να έχει 2500 αρχεία .java ενώ κάποια άλλη μπορεί να έχει 250. Η ύπαρξη και η συχνότητα εμφάνισης μιας λέξης ενδέχεται να έχει διαφορετική σημασία/βαρύτητα στις δύο αυτές βιβλιοθήκες.

2. Για το βήμα 4, θα χρειαστεί να μελετήσετε την ομαδοποίηση που πραγματοποιήθηκε. Η μελέτη απαιτεί ουσιαστικά τον έλεγχο του κατά πόσο οι ομάδες είναι «λογικές». Ένας τρόπος π.χ. θα ήταν να επιλέξετε κάποιους σημαντικούς όρους κάθε ομάδας και να δείτε κατά πόσο οι βιβλιοθήκες που ανήκουν στην ομάδα είναι λογικό να έχουν αυτούς τους όρους. Επίσης, μπορείτε π.χ. να δείτε ποιες από τις βιβλιοθήκες είναι παρόμοιες ίσως με κάποιο distance metric.

Ομάδες Εργασίας:
Οι ομάδες είναι 3 ατόμων. Θα τις δημιουργήσετε στον δικτυακό τόπο του μαθήματος (και οι συνάδελφοί σας θα κάνουν join), στην Ενότητα "Υποβολή Εργασιών -> Ομαδοποίηση".
Προθεσμία δήλωσης ομάδας: Παρασκευή 04/12/2015, 23:59 (**Αυστηρή προθεσμία!**)

Παραδοτέα:

1. Ο πηγαίος κώδικας που θα υλοποιήσετε. Η υλοποίηση μπορεί να γίνει σε Matlab, WEKA, R, Python ή συνδυασμούς αυτών.
2. Όλα τα πειράματα που θα κάνετε.
3. Έγγραφη αναφορά η οποία θα περιέχει: α) σύντομη περιγραφή του προβλήματος που καλείστε να επιλύσετε, β) τις λύσεις που δώσατε (παραμέτρους και πειράματα) και γ) αξιολόγηση των αποτελεσμάτων και συμπεράσματα. Στο εξώφυλλο μη ξεχάσετε να περιλάβετε τον τίτλο της εργασίας, ημερομηνία, καθώς και τα ονόματα και τις ηλεκτρονικές διευθύνσεις των μελών της ομάδας σας.
4. Το τελικό μοντέλο.

Βαθμολογία:
Όπως έχει ήδη αναφερθεί, η εργασία αυτή αποτελεί το 25% του τελικού βαθμού.

Εξέταση εργασίας:
Η εργασία θα πρέπει να ολοκληρωθεί μόνο κατά την περίοδο Φεβρουαρίου.

Υποβολή:
Τα παραδοτέα πρέπει να γίνουν ένα zip αρχείο, με όνομα Επίθ1_Επίθ2_Επίθ3.zip. Η υποβολή του αρχείου αυτού θα γίνει μέσω του δικτυακού τόπου του μαθήματος.
Προθεσμία υποβολής: Κυριακή, 13/12/2015, 23:59.

Για περισσότερες πληροφορίες:

Ανδρέας Λ. Συμεωνίδης	**Βάγια Καλτσά**	**Θεμιστοκλής Διαμαντόπουλος**
Επ. Καθηγητής	Μεταπτυχιακή Φοιτήτρια	Μεταπτυχιακός Φοιτητής
E-mail. asymeon@eng.auth.gr	E-mail: vkaltsa@auth.gr	E-mail: thdiaman@auth.gr