Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αλγοριθμική Επιστήμη Δεδομένων, 2018-2019

Διδάσκοντες: Δ. Φωτάκης, Α. Παγουρτζής

Προγραμματιστική Άσκηση

Έχουμε καταγραφές συνδέσεων δικτυακών συσκευών σε ένα δίκτυο διάφορων εγκαταστάσεων. Στόχος είναι να εξάγουμε πληροφορίες για τα άτομα στα οποία ανήκουν οι συσκευές.

Ειδικότερα, οι εγκαταστάσεις χωρίζονται σε τρεις κατηγορίες: Οικίες, χώροι εργασίας και ξενοδοχεία. Κάθε εγκατάσταση έχει σε κάθε χρονική στιγμή μια IP (που είναι ένας ακέραιος). Οι IP όλων των εγκαταστάσεων αλλάζουν στο χρόνο, με ρυθμό που πιθανώς να εξαρτάται από τον τύπο της εγκατάστασης. Επιπλέον οι IP είναι μοναδικές ανά πάσα χρονική στιγμή, δηλαδή δυο εγκαταστάσεις δεν έχουν ποτέ την ίδια IP ταυτόχρονα (αλλά μπορεί η ίδια IP να ανατεθεί σε διαφορετικές εγκαταστάσεις, σε διαφορετικές χρονικές στιγμές).

Κάθε συσκευή έχει ένα νούμερο (ακέραιος), την mac address της, το οποίο είναι μοναδικό – δεν έχει το ίδιο καμία άλλη συσκευή – και δεν αλλάζει ποτέ. Οι συσκευές χωρίζονται σε δυο κατηγορίες: Σταθερές και κινητές. Κάθε σταθερή συσκευή βρίσκεται μόνιμα σε μια εγκατάσταση, ενώ κάθε κινητή συσκευή ανήκει σε κάποιο άτομο. Σε κάθε άτομο μπορεί να ανήκουν περισσότερες από μία κινητές συσκευές.

Κάθε άτομο ζει σε μια οικία. Επιπλέον, τα άτομα μπορεί να επισκέπτονται άλλες οικίες, χώρους εργασίας και ξενοδοχεία, μεταφέροντας μαζί τους κάποιες από (μπορεί και όλες) τις φορητές τους συσκευές. Η συχνότητα επίσκεψης άλλων εγκαταστάσεων εξαρτάται από το προφίλ των ατόμων (δεν έχουν όλοι την ίδια συμπεριφορά) και τη χρονική στιγμή. Ένας τύπος ατόμων είναι οι εργαζόμενοι. Κάθε εργαζόμενος δουλεύει σε μία μόνο εταιρία, χωρίς αυτό να σημαίνει ότι δεν μπορεί να επισκεφθεί άλλες εταιρίες/εγκαταστάσεις.

Περιγραφή Δεδομένων: Τα δεδομένα δίνονται σε ένα csv αρχείο με τέσσερις στήλες (μέρα, ώρα, IP, mac address). Κάθε εγγραφή (γραμμή) δηλώνει ότι τη συγκεκριμένη μέρα και ώρα, η συσκευή με τη συγκεκριμένη mac address συνδέθηκε από τη συγκεκριμένη IP. Τα δεδομένα καταγράφουν βάθος χρόνου 35 ημερών (5 εβδομάδων) με καταγραφές ανά ώρα και εμφανίζονται με τυχαία σειρά στο αρχείο. Η μέρα 1, αντιστοιχεί σε Δευτέρα. Το αρχείο έχει πλήθος εγγραφών της τάξης του 10^6. Γενικώς, θέλουμε οι αλγόριθμοι που θα υλοποιηθούν να είναι αποδοτικοί για αρχεία τέτοιας τάξης μεγέθους.

Ζητούμενα: Με βάση τα παραπάνω, θέλουμε να υπολογίσουμε τα εξής:

Ερώτημα 1. (α) Ποιες συσκευές ανήκουν στην ίδια οικογένεια; Αν θεωρήσουμε ότι όσοι ζουν στην ίδια οικία είναι οικογένεια, θέλουμε να βρούμε σύνολα συσκευών που ανήκουν στην ίδια οικογένεια. Στα σύνολα αυτά, θέλουμε να συμπεριλάβουμε και τις σταθερές συσκευές των αντίστοιχων οικιών. Η απάντηση θα είναι ένα csv αρχείο του οποίου κάθε γραμμή θα αντιστοιχεί σε μια οικογένεια και θα περιέχει τις mac addresses των συσκευών της, χωρισμένες με "," (κόμμα), ταξινομημένες σε αύξουσα σειρά. Για παράδειγμα, αν έχουμε βρει ότι κάποιο σπίτι έχει δυο κατοίκους, με μια συσκευή ο πρώτος και δυο συσκευές ο δεύτερος, με mac addresses 3 και 1, 4 αντίστοιχα, και το σπίτι έχει και μια σταθερή συσκευή με mac address 2, θα πρέπει να υπάρχει η γραμμή: "1,2,3,4" στο csv αρχείο εξόδου.

(β) Ποιοι εργάζονται στον ίδιο χώρο εργασίας; Εδώ θέλουμε να ομαδοποιήσουμε τις συσκευές των ατόμων που δουλεύουν στον ίδιο χώρο, πάλι μαζί με τις σταθερές συσκευές του χώρου. Χρησιμοποιούμε την ίδια μορφή αρχείου εξόδου με το 1.α για τα αποτελέσματα.

Ερώτημα 2. (bonus) Πόσα και ποια είναι τα ξεονοδοχεία; Κάθε ξενοδοχείο έχει τουλάχιστον μια σταθερή συσκευή. Η απάντηση θα είναι πάλι ένα csv αρχείο, όπου κάθε γραμμή θα αφορά ένα ξενοδοχείο και θα περιέχει τις mac addresses των σταθερών συσκευών του.

Κριτήρια Αξιολόγησης και Αρχεία Εισόδου: Το βασικό αρχείο εισόδου της άσκησης είναι το **recordings.csv**. Δίνεται ακόμη ένα σημαντικά μικρότερο αρχείο εισόδου, το **recordings_example.csv**, με δυο αρχεία εξόδου, τα **solution_1a_example.csv** και **solution_2_example.csv**, για τα ερωτήματα 1α και 2 αντίστοιχα. Τα αρχεία αυτά περιέχουν τις ενδεικτικές λύσεις για το μικρότερο στιγμιότυπο εισόδου και μπορείτε να τα χρησιμοποιήσετε για να δοκιμάσετε τον κωδικά σας. Το δείγμα του παραδείγματος, αφορά σε καταγραφή 21 ημερών (3 εβδομάδων), και η ημέρα 1 είναι πάλι Δευτέρα.

Η βασική συνάρτηση που θα χρησιμοποιηθεί για την αξιολόγηση της ομοιότητας των λύσεων σας με τις ενδεικτικές λύσεις είναι η adjusted rand index. Μια υλοποίηση της σε βιβλιοθήκη της python είναι η: sklearn.metrics.adjusted_rand_score.

Προσοχή: Οι συχνότητες των δεδομένων του παραδείγματος (π.χ. πόσο συχνά κάποιος επισκέπτεται ένα ξενοδοχείο), όπως και οι συνήθειες των ατόμων (π.χ. τι ώρες πηγαίνουν στη δουλειά), μπορεί να διαφέρουν από αυτές του αρχείου της άσκησης.

Προθεσμία Υποβολής και Οδηγίες: Οι απαντήσεις θα πρέπει να υποβληθούν έως τις **24/5/2019**, σε ηλεκτρονική μορφή. Πρέπει να παραδόσετε τον κώδικά σας (συνοδευόμενο από οδηγίες για το πως μπορούμε να τον χρησιμοποιήσουμε και από σχόλια για τις βασικές λειτουργίες και συναρτήσεις) και την απάντησή που έχετε υπολογίσει για το αρχείο εισόδου **recordings.csv**. Είναι καλό να σημειώσετε στην υποβολή σας το χρόνο και την μνήμη που χρειάστηκε ο κώδικά σας για τον υπολογισμό της απάντησης. Ο κώδικάς σας θα αξιολογηθεί και με άλλα αρχεία εισόδου, αντίστοιχου μεγέθους, αλλά με ελαφρώς διαφορετικά χαρακτηριστικά για τις συνήθειες των ατόμων.

Για απορίες ή διευκρινίσεις, μπορείτε να στείλετε μήνυμα στη διεύθυνση `ads@corelab.ntua.gr`.

Καλή Επιτυχία!