

Πρώτη Σειρά Ασκήσεων

Αυτό είναι το δεύτερο μέρος της πρώτης σειράς ασκήσεων. Η προθεσμία για την παράδοση αυτού του κομματιού είναι στις 25 Νοεμβρίου, στο ξεκίνημα του μαθήματος. Κάνετε turn-in τον κώδικα για την άσκηση 3, και παραδώστε τις υπόλοιπες ερωτήσεις είτε ηλεκτρονικά, είτε σε χαρτί. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Οι λεπτομέρειες για το turn-in είναι στη σελίδα του μαθήματος.

Ερώτηση 1

Η Άσκηση 6.3.1 από το [Κεφάλαιο 6](#) του βιβλίου [Mining Massive Datasets](#) των Anand Rajaraman, Jeff Ullman, and Jure Leskovec.

Ερώτηση 2

Υποθέστε ότι σας δίνεται σαν είσοδος ο γράφος του Twitter με πληροφορία ποιος ακολουθεί ποιον. Θέλετε να βρείτε (s,t) -διμερείς κλίκες στο γράφημα, δηλαδή υποσύνολα από χρήστες όπου s χρήστες ακολουθούν ακριβώς τους ίδιους t χρήστες. Οι s χρήστες είναι χρήστες με κοινά ενδιαφέροντα, ενώ οι t χρήστες είναι χρήστες ίδιου «τύπου». Δώσετε ένα αποτελεσματικό αλγόριθμο για το πρόβλημα χρησιμοποιώντας την ιδέα του APriori.

Ερώτηση 3

Ο στόχος αυτής της άσκησης είναι να χρησιμοποιήσετε την εξόρυξη συχνών στοιχειοσυνόλων στην πράξη.

Για τα δεδομένα εισόδου, χρησιμοποιήστε μια συλλογή από 57K περίπου “tips” που άφησαν χρήστες του FourSquare σε διάφορα εστιατόρια στη Νέα Υόρκη. Οι λεπτομέρειες για τα δεδομένα είναι στη σελίδα του μαθήματος. Σε αυτό το dataset ένα “καλάθι” είναι ένα tip ενός χρήστη, και τα “αντικείμενα” είναι οι λέξεις. Τα δεδομένα πρέπει να περάσουν μια προεπεξεργασία, ώστε να αφαιρεθεί θόρυβος, να «κανονικοποιηθούν» οι λέξεις (μετατροπή σε μικρά, αφαίρεση σημείων στίξεως), και να αφαιρεθούν οι Αγγλικές stop-words. Λίστα με αγγλικές stop-words μπορείτε να βρείτε στη σελίδα του μαθήματος.

Θα χρησιμοποιήσετε ένα υπάρχον πακέτο για την εξόρυξη συχνών στοιχειοσυνόλων. Μπορείτε να χρησιμοποιήσετε μια υλοποίηση από τη σελίδα του FIMI, ή το WEKA. Λεπτομέρειες για το FIMI και το

WEKA είναι στη σελίδα του μαθήματος. Υπάρχει περίπτωση να πρέπει να μετατρέψετε τα δεδομένα στο format εισόδου που απαιτεί το πακέτο.

Χρησιμοποιείτε ένα αρκετά υψηλό κατώφλι για την εύρεση συχνών στοιχειοσύνολων, ώστε να μην πάρετε ένα πολύ μεγάλο αριθμό από στοιχειοσύνολα. Αν το υποστηρίζει το πακέτο που επιλέξατε υπολογίστε τα κλειστά και μεγιστικά στοιχειοσύνολα. Μετατρέψτε την έξοδο ώστε να διαβάζεται (π.χ. μετατρέψτε τα στοιχεία από αύξοντες αριθμούς σε λέξεις), επιθεωρήστε τα αποτελέσματα και σχολιάστε για στοιχειοσύνολα που βρήκατε.

Θα πρέπει να παραδώσετε τα ακόλουθα:

- Όλο τον κώδικα που θα γράψετε εσείς.
- Το αρχείο με τα συχνά στοιχειοσύνολα
- Μια σύντομη αναφορά όπου θα περιγράφετε πώς κάνατε την προεπεξεργασία, την επιλογή για το κατώφλι υποστήριξης, και ένα σχολιασμό για το τι βρήκατε.