Πανεπιστήμιο Ιωαννίνων,
Τμήμα Πληροφορικής

Εξόρυξη Δεδομένων – ΠΛΕ 059
Χειμερινό Εξάμηνο, 2012

Πρώτη Σειρά Ασκήσεων

Αυτό είναι το δεύτερο μέρος της πρώτης σειράς ασκήσεων. Η προθεσμία για την παράδοση αυτού του κομματιού είναι στις 6 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.

Άσκηση 1

Ένας χαρακτηριστικός κανόνας είναι ένας κανόνας της μορφής $\{q\} \rightarrow \{p_1, p_2, \dots, p_n\}$, όπου το αριστερό μέρος αποτελείται από ένα μόνο στοιχείο. Για ένα στοιχειοσύνολο με k στοιχεία, μπορούμε να δημιουργήσουμε k τέτοιους χαρακτηριστικούς κανόνες. Ορίζουμε το μέτρο ξ για ένα στοιχειοσύνολο, ως το ελάχιστο confidence c αυτών των κανόνων . Δηλαδή:

$$\xi(p_1, p_2, \dots, p_k) = \min\{c(\{p_1\} \rightarrow \{p_2, \dots, p_k\}), c(\{p_2\} \rightarrow \{p_1, \dots, p_k\}), \dots, c(\{p_k\} \rightarrow \{p_1, \dots, p_{k-1}\})\}$$

Είναι το μέτρο ξ μονότονο, αντί-μονότονο, ή μη μονότονο, (ούτε μονότονο, ούτε αντί-μονότονο)? Υπενθυμίζουμε ότι για οποιαδήποτε δύο σύνολα X και Y, έτσι ώστε $X \subseteq Y$, ένα μέτρο f είναι μονότονο αν $f(X) \leq f(Y)$, και αντί-μονότονο αν $f(X) \geq f(Y)$.

Άσκηση 2

Υποθέστε ότι τα δεδομένα μας αποτελούνται από μία *διατεταγμένη ακολουθία γεγονότων* (π.χ., λέξεις σε ένα κείμενο, ή χαρακτήρες σε ένα αλφαριθμητικό, ή ερωτήματα σε ένα ημερολόγιο ερωτημάτων). Θέλουμε να βρούμε όλες τις συχνές *υπο-ακολουθίες* (γνωστές και ως *σειριακά επεισόδια*) γεγονότων στα δεδομένα μας, που συμβαίνουν μέσα σε ένα χρονικό παράθυρο μεγέθους *W* (δηλαδή η χρονική διαφορά μεταξύ του πρώτου και του τελευταίου γεγονότος της υπο-ακολουθίας είναι το πολύ *W*). Για παράδειγμα, <A,B,C> είναι μια υπο-ακολουθία μήκους 3, η οποία εμφανίζεται σε ένα παράθυρο μεγέθους *W = 5* το οποίο περιέχει τα γεγονότα <A,A,B,D,C>. Για την ακρίβεια υπάρχουν δύο εμφανίσεις της υπο-ακολουθίας σε αυτό το παράθυρο, μία που ξεκινάει στην πρώτη θέση και μία που ξεκινάει στη δεύτερη θέση. Το support μιας υπο-ακολουθίας είναι ο αριθμός των φορών που εμφανίζεται μέσα στην συνολική ακολουθία γεγονότων.

Περιγράψτε ένα αλγόριθμο ο οποίος λειτουργεί σε επίπεδα (level-wise) και υπολογίζει όλες τις συχνές υπο-ακολουθίες με support τουλάχιστον *minsup*, για δεδομένο μέγεθος παραθύρου W. Περιγράψτε ξεκάθαρα τη δημιουργία υποψήφιων υπο-ακολουθιών, και τον υπολογισμό των συχνών υπο-ακολουθιών. Ο αλγόριθμος σας θα πρέπει να είναι αποδοτικός ως προς μνήμη και χρόνο, όπως ο Apriori αλγόριθμος. Περιγράψτε τον αλγόριθμο σας με ψευδοκώδικα, ή στα ελληνικά.

Εφαρμόστε τον αλγόριθμο σας στην ακολουθία ABCACDBAAC με *W=4* και *minsup* = 2, και αναφέρετε την έξοδο όλων των ενδιάμεσων βημάτων του αλγορίθμου.

Άσκηση 3

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

Έχετε τις εξής επιλογές για τα δεδομένα εισόδου:

1. Μια συλλογή από 6500 προφίλ χρηστών του Twitter. Οι λεπτομέρειες για τα δεδομένα είναι στη σελίδα του μαθήματος. Σε αυτό το dataset ένα "καλάθι" είναι ένα προφίλ ενός χρήστη, και τα "στοιχεία" 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.
2. Κατεβάστε emails που έχετε στείλει και βρείτε emails που έχουν σταλεί σε τουλάχιστον δύο διευθύνσεις. Δημιουργήστε ένα dataset όπου τα «καλάθια» είναι τα emails και τα «στοιχεία» οι διευθύνσεις. Στα δεδομένα σας θα πρέπει να έχετε τουλάχιστον 1000 καλάθια, κατά προτίμηση παραπάνω.
3. Προτείνετε εσείς ένα dataset το οποίο πιστεύετε ότι είναι ενδιαφέρον να αναλύσετε. . Στα δεδομένα σας θα πρέπει να έχετε τουλάχιστον 1000 καλάθια, κατά προτίμηση παραπάνω. Επικοινωνήστε μαζί μου για την πρόταση σας.

Αφού αποφασίσετε για τα δεδομένα, 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. Θα πρέπει να μετατρέψετε τα δεδομένα στο format εισόδου που απαιτεί το κάθε πακέτο.

Τρέξτε το πακέτο για συχνά στοιχειοσύνολα χρησιμοποιώντας αρκετά υψηλό κατώφλι, ώστε να μην πάρετε ένα πολύ μεγάλο αριθμό από στοιχειοσύνολα. Αν το υποστηρίζει το πακέτο 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:

- Το προεπεξεργασμένο αρχείο εισόδου.
- Αρχείο με τα στοιχειοσύνολα
- Μια σύντομη αναφορά όπου θα περιγράφετε το αρχείο εισόδου, πώς κάνατε την προεπεξεργασία, την επιλογή για το κατώφλι υποστήριξης, και ένα σχολιασμό για το τι βρήκατε.

Τεχνικές Λεπτομέρειες

Για το pre-processing των δεδομένων, οι παρακάτω unix εντολές μπορεί να σας φανούν χρήσιμες:

1. cut: επιτρέπει να πάρουμε συγκεκριμένες κολώνες από ένα αρχείο με διαχωριζόμενες τιμές
2. sort: ταξινομεί τις γραμμές ενός αρχείου σε αλφαβητική σειρά . –n for αριθμητική σειρά
3. uniq: αφαιρεί συνεχόμενες γραμμές που είναι ίδιες.
4. grep: βρίσκει μια έκφραση μέσα σε ένα αρχείο.