Τεχνικές Εξόρυξης Δεδομένων: Χειμερινό Εξάμηνο 2011/2012

Δεύτερη Άσκηση (ομάδες έως 2 ατόμων)

Ημερομηνία Παράδοσης: 12/2/2012

Στην άσκηση αυτή θα υλοποιήσετε τον αλγόριθμο **COP-KMEANS** που κάνει ομαδοποίηση σημείων όταν υπαρχουν περιορισμοί της μορφής: «Το σημείο Α πρέπει να είναι στο ίδιο cluster με το σημείο Β», ή «Το σημείο Γ δεν μπορεί να είναι στο ίδιο cluster με το σημείο 6».

Η περιγραφή του αλγορίθμου υπάρχει σην εργασία:
http://www.litech.org/~wkiri/Papers/wagstaff-kmeans-01.pdf

1. Σωστή υλοποίηση του αλγορίθμου: **50%**
   Μπορείτε να χρησιμοποιήσετε C, C++, ή Java.
   Κώδικας που δεν τρέχει δεν θα βαθμολογηθεί καθόλου.

   Λεπτομέρειες για το σύνολο δεδομένων που θα χρησιμοποιήσετε θα σας δωθούν σύντομα. Στη γενική περίπτωση θα έχετε ένα **σύνολο σημείων** και ένα **σύνολο περιορισμών** που πρέπει να ικανοποιούν τα clusters που θα δημιουργήσει ο αλγόριθμός σας. Οι περιορισμοί θα είναι τύπου **must link** και **cannot link**.

   Π.χ.

   N (αριθμός σημείων)
   x_1 y_1 (double, double)
   .
   .

   x_N y_N
   M_1 (αριθμός **must link** περιορισμών)
   A_1 B_1 (integer , integer – το A_1 ειναι ο αριθμός του σημείου στην παραπάνω σειρά εισόδου των σημείων)
   .

A_{M1} B_{M1}
M_2 (αριθμός **cannot link** περιορισμών)
$Γ_1$ $Δ_1$
.
.
$Γ_{M2}$ 6_{M2}

2. Documentation της υλοποίησης: **5%**

3. Γραφική αναπαράσταση των αποτελεσμάτων (ποια σημεία ανήκουν σε ποια clusters, ποιοι περιορισμοί ικανοποιούνται και ποιοι όχι). **15%**

4. Υπάρχει πιθανότητα ο αλγόριθμος να μην τερματίζει? Δώστε αντιπαραδείγματα ή αποδείξτε ότι τερματίζει πάντα. **10%**

5. Είναι δυνατό ο αλγόριθμος να επιστρέψει ότι δεν υπαρχει λύση, ενώ υπάρχει λύση? Δώστε αντιπαραδείγματα ή αποδείξτε ότι αυτό δεν μπορεί να συμβεί. **10%**

6. Πώς επηρεάζονται τα αποτελέσματα του αλγορίθμου αν η σειρά με την οποία εξετάζονται τα σημεία και οι περιορισμοί αλλάξει? **10%**

E-mail επικοινωνίας: Δημήτρης Κωτσάκος (dimkots [at] di.uoa.gr).