Τεχνικές Εξόρυξης Δεδομένων (ΥΣ 11)
Χειμερινό Εξάμηνο 2013 - 2014

2η Άσκηση, Ημερομηνία Παράδοσης: 25-03-2014
(Μπορείτε να παραδώσετε την εργασία έως 27-04-2014 με μειωμένη βαθμολογία 10%)

Ομαδική εργασία (2 ατόμων)

Σε αυτή την άσκηση θα μελετήσετε το πρόβλημα εύρεσης ομοιότητας μεταξύ τροχιών (trajectory similarity problem) το οποίο είναι ένα γνωστό πρόβλημα στον τομέα της Εξόρυξης Δεδομένων. Ο αλγόριθμος που καλείστε να υλοποιήσετε και επιλύει το παραπάνω πρόβλημα είναι ο LCSS (Longest Common Subsequence) [1]. Η κάθε τροχιά περιγράφεται από μια ακολουθία γεωγραφικών σημείων (latidute,longitude). Κάθε αρχείο του dataset που σας δίνεται αντιπροσωπεύει μια διαδρομή ενός ταξί που έχει πραγματοποιηθεί στην πόλη του Πεκίνου. Ένα ενδεικτικό παράδειγμα της μορφής των αρχείων είναι το εξής:

366,Mon Mar 03 00:05:59 EET 2014,39.90732,116.45353
366,Mon Mar 03 00:10:59 EET 2014,39.90729,116.45348
366,Mon Mar 03 00:15:59 EET 2014,39.90725,116.45334
366,Mon Mar 03 00:20:59 EET 2014,39.90722,116.4533
366,Mon Mar 03 00:25:59 EET 2014,39.90722,116.45327
366,Mon Mar 03 00:30:59 EET 2014,39.90725,116.4532
366,Mon Mar 03 00:35:59 EET 2014,39.9076,116.45309
366,Mon Mar 03 00:40:59 EET 2014,39.9077,116.453
366,Mon Mar 03 00:45:59 EET 2014,39.9076,116.45281
366,Mon Mar 03 00:50:59 EET 2014,39.90767,116.45271
366,Mon Mar 03 00:55:59 EET 2014,39.90771,116.45262

Η κάθε γραμμή του αρχείου περιέχει τα ακόλουθα χαρακτηριστικά χωρισμένα με κόμμα:
taxi id, timestamp, latitude, longitude
Στην παρούσα εργασία τα χαρακτηριστικά που θα χρησιμοποιήσετε είναι το latitude (γεωγραφικό πλάτος ενός σημείου) και το longitude (γεωγραφικό μήκος ενός σημείου)

Ζητούμενα:

1. Υλοποίηση του αλγόριθμου LCSS (40%)
    a. Τεκμηρίωση της υλοποίησή σας
    b. Υποστήριξη της εφαρμογής με παραθυρικό περιβάλλον χρησιμοποιώντας κάποιο αντίστοιχο framework (πχ Java Swing,Windows Forms/WPF, Qt).
2. Για κάθε φάκελο του dataset (1000Points,5000Points,10000Points) συγκρίνετε κάθε αρχείο-διαδρομή με τα υπόλοιπα του φακέλου. Έπειτα, για κάθε διαδρομή βρείτε την πιο όμοια από αυτές. Καταγράψτε το ποσοστό ομοιότητας μαζί με το όριο ταιριάσματος

(παράμετρος ε) που επιλέξατε και τους χρόνους από κάθε σύγκριση που εκτελέσατε στην αναφορά και σχολιάστε τα αποτελέσματα. (20%)

3. Τροποποιήστε το πρόγραμμά σας, έτσι ώστε έχοντας σαν είσοδο 2 τροχιές S, Q με μήκη L_s, L_q αντίστοιχα, όπου $L_s >> L_q$, να βρίσκει το υποσύνολο της τροχιάς S το οποίο παρουσιάζει την μεγαλύτερη ομοιότητα με την τροχιά σύγκρισης Q. Δηλαδή να επιστρέφει το τμήμα της S μήκους $L_q + \delta$ ($0 < \delta < L_q$) το οποίο έχει το μεγαλύτερο ποσοστό ομοιότητας με την Q. Συγκρίνετε τις διαδρομές που περιέχονται στον φάκελο 1000Points με κάθε μία από τις διαδρομές που βρίσκονται στο φάκελο 10000Points και σχολιάστε τα αποτελέσματα τους. (20%)
4. Αναζητήστε στην βιβλιογραφία **(**βιβλία**,** δημοσιεύσεις κ.α**.)** τις διαφορές μεταξύ του αλγόριθμου LCSS και του Dynamic Time Warping (10%)
5. Παρουσιάστε την τροχιά σύγκρισης, σε συνδυασμό με την ομοιότερη τροχιά που βρήκατε μέσω του LCSS πάνω στους άξονες x,y. (τις συνολικές τροχιές όχι μόνο το κοινό τους κομμάτι) (10%)
6. Η εμφάνιση σε χάρτη (πχ. Google Maps) της τροχιάς σύγκρισης, σε συνδυασμό με την ομοιότερη τροχιά που βρήκατε μέσω του LCSS θα βαθμολογηθεί με 10% bonus. (τις συνολικές τροχιές όχι μόνο το κοινό τους κομμάτι)

Διευκρινήσεις

Για κάθε σύγκριση που θα κάνετε με χρήση του αλγορίθμου καταγράψτε το ποσοστό ομοιότητας των τροχιών καθώς και την παραμετροποίηση που κάνατε στον αλγόριθμο.
Η αναφορά που καλείστε να παραδώσετε μαζί με τον κώδικα της εργασίας σας, θα πρέπει να περιλαμβάνει:
Τα αποτελέσματα των πειραμάτων που θα τρέξετε, τον σχολιασμό που θα κάνετε πάνω σε αυτά, μια σύντομη περιγραφή της υλοποίησης σας και οτιδήποτε πιστεύετε πως θα βοηθήσει τον διορθωτή να βαθμολογήσει την εργασία σας.
Εργασίες των οποίων ο κώδικας δεν μεταγλωττίζεται δεν θα βαθμολογηθούν. Είστε ελεύθεροι να χρησιμοποιείστε οποιαδήποτε γλώσσα προγραμματισμού επιθυμείτε.

Παραδοτέα

- Ένας φάκελος με το project σας (ο οποίος θα περιλαμβάνει τον πηγαίο σας κώδικα.
- Την αναφορά σε μορφή PDF.

Μπορείτε να επιλέξετε οποιαδήποτε γλώσσα προγραμματισμού και πλατφόρμα θέλετε. Καλείστε να έχετε καλύψει ακραίες περιπτώσεις στον κώδικά σας, όπως π.χ. μη αναμενόμενη είσοδο, όχι επαρκής διάθεση δυναμικής μνήμης κ.α.. Η αναφορά είναι αποδεκτή μόνο σε PDF format.

Στο cmitatakis@di.uoa.gr και nlarios@di.uoa.gr μπορείτε να επικοινωνείτε για απορίες και εκεί θα παραδώσετε και την εργασία σας.

Αναφορές

[1] http://www.cs.bu.edu/groups/dblab/pub_pdfs/icde02.pdf