Κ23γ: Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα - Χειμερινό εξάμηνο 2017-18
3η Προγραμματιστική Εργασία

Η εργασία έχει 2 σκέλη. Πρέπει να υλοποιηθεί σε σύστημα Linux και να υποβληθεί στις Εργασίες του e-class το αργότερο την Παρασκευή 12/01/2018 στις 23.59.

1. **Συσταδοποίηση μοριακών διαμορφώσεων**

Δίνεται σύνολο μοριακών διαμορφώσεων (conformations). Κάθε διαμόρφωση αποτελείται από μία συγκεκριμένη ακολουθία N σημείων στον τριδιάστατο Ευκλείδειο χώρο.

Η είσοδος είναι αρχείο κειμένου input.dat. Οι δύο πρώτες γραμμές περιέχουν το πλήθος διαμορφώσεων και το N. Το υπόλοιπο αρχείο είναι διαχωρισμένο με στηλοθέτες σε 3 στήλες που περιέχουν τις N τριάδες συντεταγμένων: x y z της 1ης διαμόρφωσης, έπειτα N τριάδες της 2ης διαμόρφωσης κοκ, δηλ. συνολικά numConform*N γραμμές μετά τις 2 πρώτες:

```
numConform: <Integer>
N: <Integer>
-32.5       91.2        11.7
12.8        -18.3       79.1
...
```

Α. Ζητείται (1) η υλοποίηση της συνάρτησης απόστασης c-RMSD, με την χρήση εξωτερικής βιβλιοθήκης γραμμικής άλγεβρας, π.χ. LAPACKE / GNU Scientific Library / Eigen, (2) η συσταδοποίηση σε k συστάδες (clusters), για δοσμένο k, με αλγόριθμο της επιλογής σας τύπου k-medoids και χρήση της c-RMSD, (3) η εύρεση του κατάλληλου k μέσω του μέτρου αξιολόγησης silhouette.

Η έξοδος είναι αρχείο κειμένου crmsd.dat, το οποίο εκφράζει τη συσταδοποίηση με την βέλτιστη τιμή silhouette, διαχωρισμένο με στηλοθέτες με την ακόλουθη γραμμογράφηση:

```
k: <Integer>
s: <real in [-1,1]>
1     9     11    12 ….
2     3     17    [η 2η συστάδα περιέχει 3 στοιχεία]
```

όπου η 1η γραμμή περιέχει το k, η 2η την τιμή του silhouette, και οι επόμενες k γραμμές περιέχουν τους δείκτες των στοιχείων των αντίστοιχων συστάδων σε αύξουσα σειρά.

B. Ζητείται:

(1) Η συσταδοποίηση με την προσέγγιση k-means με χρήση της απόστασης Discrete Frechét. Πριν τον υπολογισμό της απόστασης οι διαμορφώσεις μετατοπίζονται και περιστρέφονται με κριτήριο την ελαχιστοποίηση της απόστασης c-RMSD,
(2) Η εύρεση του βέλτιστου k μέσω του μέτρου silhouette.

Η έξοδος είναι αρχείο κειμένου `frechet.dat`, το οποίο εκφράζει τη συσταδοποίηση με την βέλτιστη τιμή silhouette, διαχωρισμένο με στηλοθέτες με την ακόλουθη γραμμογράφηση:

```
k: <Integer>
s: <real in [-1,1]>
1     9     11    12 ….
2     3     17    [η 2η συστάδα περιέχει 3 στοιχεία]
```

όπου η 1η γραμμή περιέχει το k, η 2η την τιμή του silhouette, και οι επόμενες k γραμμές περιέχουν τους δείκτες των στοιχείων των αντίστοιχων συστάδων σε αύξουσα σειρά.

2. Συσταδοποίηση οδικών τμημάτων

Ζητείται:

(1) Η λήψη των χαρτογραφικών δεδομένων της πλατφόρμας Open Street Map για την περιοχή της Αθήνας από τη διεύθυνση https://s3.amazonaws.com/metro-extracts.mapzen.com/athens_greece.osm.bz2 και η εξαγωγή των δεδομένων που αφορούν οδούς (ways) που φέρουν την ετικέτα (tag) highway με τις ακόλουθες τιμές: service, motorway, trunk, primary, secondary, tertiary, unclassified, residential, καθώς και τους κόμβους (nodes) που περιλαμβάνουν οι οδοί, σε αρχείο athens.csv με την ακόλουθη γραμμογράφηση:

```
way_ID, type, node1_lat, node1_lon, . . ., last_node_lat, last_node_lon
4349655, residential, 38.0572542, 23.5304318, . . ., 38.0597047, 23.5308991
…    …    …    …    …    …    …    …    …    …    …    …
```

(2) Η κατάτμηση των οδών σε τμήματα (segments). Οι οδοί κατατέμνονται στα σημεία που ακολουθούν διασταυρώσεις (όταν δύο κόμβοι ανήκουν σε περισσότερες από μία οδούς) ή όταν αλλάζει η καμπυλότητά τους πάνω από ένα κατώφλι. Η καμπυλότητα μίας ακμής σε μια πολυγωνική καμπύλη ορίζεται ως η ελάχιστη καμπυλότητα των 2 τριγώνων στα οποία ανήκουν οι κορυφές της ακμής. Λαμβάνεται πρόνοια ώστε να μην προκύπτουν πολύ μικρά ή πολύ μεγάλα οδικά τμήματα: σε περίπτωση διαδοχικών διασταυρώσεων η κατάτμηση γίνεται στο σημείο μετά το πέρας όλων των διασταυρώσεων, επιλέγεται κατάλληλα το κατώφλι αλλαγής καμπυλότητας για την κατάτμηση μέσω πειραματισμού, πολύ μεγάλα οδικά τμήματα κατατέμνονται βάσει κατάλληλου ευριστικού κανόνα (λ.χ. μέγιστο μήκος οδικού τμήματος). Τα οδικά τμήματα αποθηκεύονται σε αρχείο segment.csv με την ακόλουθη γραμμογράφηση:

no, wayID, num_of_segs, node1_lat, node1_lon, . . ., last_node_lat, last_node_lon
0, 15918864, 2, 40.6157885, 22.9575427, 40.6163788, 22.9574743

... ...

(3) Η συσταδοποίηση σε k συστάδες (clusters), για δοσμένο k, με αλγόριθμο k-means / k-medoids (επιλέγετε έναν από όσους αναπτύχθηκαν στην δεύτερη εργασία) και χρήση των αποστάσεων Discrete Frechét και DTW (για ομάδες 2 ατόμων). Πριν τον υπολογισμό της απόστασης οι διαμορφώσεις μετατοπίζονται και περιστρέφονται με κριτήριο την ελαχιστοποίηση της απόστασης c-RMSD. Ζητείται επίσης η εύρεση του βέλτιστου k μέσω του μέτρου αξιολόγησης silhouette.

Η έξοδος είναι αρχείο κειμένου `kmeans_ways_frechet.dat ή kmeans_ways_dtw.dat` (για ομάδες 2 ατόμων) το οποίο εκφράζει τη συσταδοποίηση με την βέλτιστη τιμή silhouette, διαχωρισμένο με στηλοθέτες με την ακόλουθη γραμμογράφηση:

```
k: <Integer>
s: <real in [-1,1]>
clustering_time: <number of msec>
1     9     11    12 ….
2     3     17    [η 2η συστάδα περιέχει 3 στοιχεία]
```

όπου η 1η γραμμή περιέχει το k, η 2η την τιμή του silhouette, και οι επόμενες k γραμμές περιέχουν τους δείκτες (no) των οδικών τμημάτων που ανήκουν στις αντίστοιχες συστάδες σε αύξουσα σειρά.

(4 – bonus για ατομικές εργασίες / υποχρεωτικό για ομάδες 2 ατόμων) Η συσταδοποίηση με τη χρήση του αλγορίθμου LSH για πολυγωνικές καμπύλες και της απόστασης Discrete Frechét. Οι συστάδες αντιστοιχούν στους κάδους (buckets) του πίνακα κατακερματισμού. Εκτελούνται πειράματα με αλλαγή του αριθμού κ των καμπυλών πλέγματος ώστε να επιλεχθεί αυτός που βελτιστοποιεί το μέτρο αξιολόγησης silhouette. Ο αριθμός L των πινάκων κατακερματισμού είναι ίσος με 1.

Η έξοδος είναι αρχείο κειμένου `lsh_ways_clustering.dat` (για ομάδες 2 ατόμων) το οποίο εκφράζει τη συσταδοποίηση με την βέλτιστη τιμή silhouette, διαχωρισμένο με στηλοθέτες με την ακόλουθη γραμμογράφηση:

```
k: <Integer>
clustering_time: <number of msec>
s: <real in [-1,1]>
1     9     11    12 ….
2     3     17    [η 2η συστάδα περιέχει 3 στοιχεία]
```

όπου η 1η γραμμή περιέχει τον αριθμό των συστάδων που σε αυτή την περίπτωση δεν είναι προκαθορισμένος, η 2η την τιμή του silhouette, και οι επόμενες k γραμμές περιέχουν τους δείκτες (no) των οδικών τμημάτων που ανήκουν στις αντίστοιχες συστάδες σε αύξουσα σειρά.

Στην αναφορά θα παρουσιάζεται η απόδοση και οι επιδόσεις της συσταδοποίησης των οδικών τμημάτων στα ερωτήματα (3) και (4) και τα συμπεράσματα που εξάγονται (για ομάδες 2 ατόμων).

Επιπρόσθετες απαιτήσεις

Όπως στις προηγούμενες εργασίες.