

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ + ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ  
Εθνικόν και Καποδιστριακόν  
Πανεπιστήμιον Αθηνών  
—ΙΔΡΥΘΕΝ ΤΟ 1837—



Αναζήτηση και συσταδοποίηση  
διανυσμάτων στη C/C++

# Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα

— 1η Προγραμματιστική Εργασία

ΔΗΜΗΤΡΗΣ ΧΑΜΑΡΙΑΣ - 1115201600190

ΑΝΤΩΝΙΑ ΡΟΥΣΣΟΥ - 1115201600147

Χειμερινό εξάμηνο 2021-2022

# Κατάλογος αρχείων

## Δομή φακέλων / αρχείων

- Φάκελος **cluster** με τα αρχεία:
  - *cluster.conf* (Αρχείο ρύθμισης παραμέτρων)
  - *cluster.cpp* (Εκτέλεση αλγορίθμων για τη συσταδοποίηση διανυσμάτων)
  - *clusteringMethods.cpp* (Ορισμοί μεθόδων της κλάσης Cluster)
  - *clusteringMethods.hpp* (Ορισμός της κλάσης Cluster)
- Φάκελος **cube** με τα αρχεία:
  - *cube.cpp* (Εκτέλεση αλγορίθμων για την τυχαία προβολή στον υπερκύβο)
  - *cubeSearch.cpp* (Ορισμός αλγορίθμων Hypercube)
  - *cubeSearch.hpp* (Δηλώσεις αλγορίθμων(συναρτήσεων) Hypercube)
- Φάκελος **lsh** με τα αρχεία:
  - *lsh.cpp* (Εκτέλεση αλγορίθμων LSH)
  - *lshSearch.cpp* (Ορισμός αλγορίθμων LSH)
  - *lshSearch.hpp* (Δηλώσεις αλγορίθμων(συναρτήσεων) LSH)
- Φάκελος **Datasets** με τα αρχεία (σύνολο δεδομένων και σύνολο αναζήτησης):
  - *input\_b\_id*
  - *input\_small\_id*
  - *query\_b\_id*
  - *query\_small\_id*
- Φάκελος **utilities** με τα αρχεία:
  - *hash.cpp* (Ορισμός των μεθόδων HashTable)
  - *hash.hpp* (Ορισμός της κλάσης HashTable και της κλάσης Data)
  - *hypercube.cpp* (Ορισμός των μεθόδων Hypercube)
  - *hypercube.hpp* (Ορισμός της κλάσης Hypercube)
  - *metrics.cpp* (Ορισμός των μετρικών, ευκλείδεια απόσταση, manhattan και hamming )
  - *metrics.hpp* (Δηλώσεις μετρικών)
  - *PriorityQueue.cpp* (Ορισμός των μεθόδων PriorityQueue)
  - *PriorityQueue.hpp* (Ορισμός της κλάσης PriorityQueue)
  - *utilities.cpp* (Ορισμός κοινών συναρτήσεων για parsing και διάβασμα/εκτύπωση αρχείων)

- ο *utilities.hpp* (Ορισμός του struct Neighbor και δηλώσεις κοινών συναρτήσεων)

## Περιγραφή κλάσεων / structs

### Data

Αντιπροσωπεύει ένα διάνυσμα και έχει τις εξής πληροφορίες: το ίδιο το διάνυσμα **vec**, το id του **id**, το cluster στο οποίο ανήκει **cluster** και την απόσταση του πλησιέστερου cluster **minDist**.

### HashTable

Η βασική δομή ενός hash table που περιέχει τις εξής πληροφορίες: το μέγεθος του πίνακα **size**, τον ίδιο τον πίνακα **table** που υλοποιείται ως ένα array από λίστες(η κάθε λίστα περιέχει δείκτες σε Data), το πλήθος των στοιχείων του **containedItems** και μια λίστα από τις hash functions **hashFunctions**.

### Hypercube

Η δομή του υπερκύβου που κληρονομεί από την κλάση HashTable και περιέχει τις επιπλέον πληροφορίες: μια λίστα από unordered maps που περιέχουν για κάθε συνάρτηση h, ζευγάρια <τιμή h, 0 ή 1>.

### Neighbor

Αναπαριστά έναν γείτονα (πλησιέστερο σημείο) και αποτελείται από ένα **id** και την απόστασή του από ένα query **dist**.

### Centroid

Αντιπροσωπεύει ένα κεντροειδές και αποτελείται από τις εξής πληροφορίες: το κεντροειδές ως διάνυσμα **vec**, το άθροισμα των σημείων που περιέχει **vecSum**, τα indexes των σημείων που περιέχει **indexes** και το μέσο silhouette κάθε σημείου **silhouette**.

## PriorityQueue

Δομή μιας ουράς προτεραιότητας με αναπαράσταση vector που περιέχει Neighbor (*heap*).

## Περιγραφή προγράμματος

Τα τρία βασικά αρχεία **lsh.cpp**, **cube.cpp** και **cluster.cpp** ακολουθούν την ίδια λογική. Κάνουν parse τις παραμέτρους της γραμμής εντολών, δημιουργούν τις κατάλληλες δομές που χρειάζονται (Hashtable, Hypercube, Cluster), διαβάζουν τα input αρχεία και τέλος εκτελούν τους αντίστοιχους αλγορίθμους.

Το πρόγραμμα lsh και το cube αντίστοιχα έχουν τις δικές τους εκδοχές για τους αλγόριθμους nearest neighbor, τους οποίους χρησιμοποιεί μετά το cluster.

Το πρόγραμμα επεκτείνεται και σε διανυσματικούς χώρους με άλλη μετρική καθώς η κλήση των συναρτήσεων/αλγορίθμων πραγματοποιείται με χρήση function pointers.

## Οδηγίες χρήσης προγράμματος

### Μεταγλώττιση

- Όλο το πρόγραμμα: **make**
- Πρόγραμμα lsh: **make lsh**
- Πρόγραμμα cube : **make cube**
- Πρόγραμμα cluster: **make cluster**

### Clean

- **make clean**

### Εκτέλεση

- Πρόγραμμα lsh: **./build/lsh -i datasets/input\_small\_id -q test\_query\_data -o outputLSH.txt -N 3 -R 250**
- Πρόγραμμα cube : **./build/cube -i datasets/input\_small\_id -q test\_query\_data -k 3 -probes 1 -M 10 -o outputCube.txt -N 4 -R 100000**

- 
- Πρόγραμμα cluster: `./build/cluster -i datasets/input_small_id -c cluster/cluster.conf -o outputCluster.txt -m LSH`

Github repository: [https://github.com/DImiTrisXam/algo\\_project1](https://github.com/DImiTrisXam/algo_project1)

■ ■ ■