Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα

Χειμερινό εξάμηνο 2016-17

3η Προγραμματιστική Εργασία

Η εργασία υλοποιήθηκε σε περιβάλλον linux.

ΟΝ/ΜΟ: Δουμιά Φωτεινή AM: 1115201300039

ΟΝ/ΜΟ: Κωνσταντάκης Δημήτριος AM: 1115201300079

Recommendation

Α. Μέθοδος NN-LSH Recommendation

Έχουν υλοποιηθεί όλες οι απαιτούμενες λειτουργίες σύμφωνα με τις διαφάνειες και τη θεωρία.

Παρατηρήσεις:

Στη μετρική hamming εκτελείται η αποκοπή (cut off), ενώ στις euclidean και cosine γίνεται κανονικοποιήση των δεδομένων. Ειδικότερα στην Euclidean τα δεδομένα κανονικοποιούνται και ύστερα εισάγονται στα Hashtables , ενώ στην μετρική cosine η κανονικοποίηση γίνεται μετά.

Similarity: Η ομοιότητα στον hamming υπολογίζεται αφαιρώντας από την μονάδα την απόσταση hamming των χρηστών διαιρεμένη από τον αριθμό των bits. Αντίθετα για τις διανυσματικές μετρικές χρησιμοποιείται η ομοιότητα cosine.

Closer Neighbors: Έχουν υλοποιηθεί και οι δύο μέθοδοι που εντοπίζουν τους P κοντινότερους γείτονες. Για λόγους ταχύτητας χρησιμοποιείται η μέθοδος που εντοπίζει έναν έναν τους γείτονες και όποιον βρίσκει τον αφαιρεί από την προσωρινή λίστα , ώστε να μην επιλεχτεί ξανά (σε πολυπλοκότητα μοιάζει με την αντίστοιχη sort συνάρτηση για τους πρώτους P γείτονες). Η άλλη μέθοδος παίνρει δύο ακτίνες και με βάση το θεώρημα Bolzano παίρνει υποδιπλάσιο δίαστημα στο οποίο υπάρχει η ακτίνα r όπου για αυτή έχουμε P γείτονες.

Validation: Χωρίζει τα ζεύγη σε 10 υποσύνολα και με βάση αυτά φτιάχνει τα νέα αντικείμενα. Αφού προβλέψει τις νέες τιμές τις συγκρίνει με τις πραγματικές και υπολογίζει το MAE.

Εκτέλεση:

make

./recommendation -d Files/yahoo_music_small.dat -o Files/Output.csv

./recommendation -d Files/yahoo_music_small.dat -o Files/Output.csv –validate

B. Μέθοδος συσταδοποίησης (Clustering) Recommendation

Παρόμοιο με το NN_LSH , με τη διαφορά ότι χειριζόμαστε clusters.

Εκτέλεση:

make

./recommendation -d Files/yahoo_music_small.dat -o Files/Output.csv

./recommendation -d Files/yahoo_music_small.dat -o Files/Output.csv –validate

(Σημείωση: για το yahoo_music_big.dat , αρκεί το small να γίνει big)

Συσταδοποίηση μοριακών διαμορφώσεων

A.

Έχουν υλοποιηθεί όλες οι απαιτήσεις . Χρησιμοποιήθηκε ως βάση σχεδιασμού η μετρική Euclidean και πάνω σε αυτό χτίστηκε η cRMSD. Η απόσταση υπολογίζεται με την απόσταση cRMSD (όπου καλείται η ευκλείδια γίνεται η αντικατάσταση της με την cRMSD)

Έγινε χρήση της βιβλιοθήκης γραμμικής άλγεβρας eigen. Υπάρχουν οι βιβλιοθήκες μέσα στο αρχείο και γίνονται include μέσω του makefile.

Εκτέλεση:

make

./medoids -d Files/bio_small_input.dat -o Files/Output.csv

(Σημείωση: για το bio_big_input.dat, αρκεί το small να γίνει big)

B.

Έχουν υλοποιηθεί όλες οι απαιτήσεις . Χρησιμοποιήθηκε ως βάση σχεδιασμού η μετρική Euclidean και πάνω σε αυτό χτίστηκε η dRMSD. Η απόσταση υπολογίζεται με την ευκλείδια απόσταση. Οι αποστάσεις επιλέγονται ως προς την τιμή ανάλογα την τιμή T (0 για min αποστάσεις, 1για max αποστάσεις, 2 ή άλλο για random αποστάσεις) και το πλήθος των αποστάσεων R (N, N^1.5, N*(N-1)/2). Οι τυχαίες αποστάσεις για λόγους ταχύτητας

λαμβάνονται σειριακά (απόσταση point 1 από 2, 1 από 3, 1 από 4, …, 1 από N, 2 από 3, …, N-1 από N).

Το πρόγραμμα έχει τρέξει για όλους τους συνδυασμούς (T, R, k) και το αποτέλεσμα υπάρχει στο αρχείο experim .Γενικά παρατηρήσαμε ότι οι καλύτερες μέσες τιμές για το Silhouette ήταν:

k = 5, R = N, T = min , Silhouette = 0.247475

k = 10, R = N, T = min , Silhouette = 0.240380

k = 15, R = N, T = min , Silhouette = 0.256753

k = 15, R = N, T = max , Silhouette = 0.255894

k = 15, R = N, T = random , Silhouette = 0.252201

k = 15, R = N^1.5, T = max , Silhouette = 0.246491

Εκτέλεση:

make

./medoids -d Files/bio_small_input.dat -o Files/Output.csv

(Σημείωση: για το bio_big_input.dat, αρκεί το small να γίνει big)