

Εργασία στο μάθημα Πολυδιάστες Δομές Δεδομένων.

Δεσύλλας Περικλής 1059548
Παναγιωτόπουλος Παναγιώτης 1057602
Στεμιτσιώτης Χρήστος 1054375
Χριστόπουλος Παναγιώτης 1054409

1.K-D Tree

Τα K-D tree είναι μια δομή δεδομένων διαχωρισμού χώρου για την οργάνωση σημείων σε ένα χώρο k διαστάσεων. Τα δέντρα αυτά είναι δυαδικά και κάθε κόμβος είναι ένα σημείο k διαστάσεων. Για το συγκεκριμένο δέντρο δημιουργήσαμε τις εξής συναρτήσεις:

1. Συνάρτηση 'distance' η οποία υπολογίζει την απόσταση δυο σημείων.
2. Συνάρτηση 'closestNeighbour' η οποία δοθέντων 2 σημείων βρίσκει το κοντινότερο σημείο μέσω σύγκρισης των αποστάσεων
3. Συνάρτηση 'minDistance' η οποία μας επιστρέφει ποιο σημείο είναι πιο κοντά στο σημείο που ζητάμε.
4. Συνάρτηση 'checkLength' η οποία
5. Συνάρτηση 'kdTree' η οποία ορίζει το δέντρο
6. Συνάρτηση 'kdTreeClosest' η οποία μας επιστρέφει την καλύτερη δυνατή λύση

2.Quad Tree

Το Quad tree είναι μια δομής δεδομένων δέντρου στο οποίο κάθε εσωτερικός κόμβος έχει ακριβώς 4 παιδιά και χρησιμοποιούνται κυρίως για το χώρισμα ενός δισδιάστατου χώρου με αναδρομική υποδιαιρέση σε τέσσερις 'περιφέρειες'. Με άλλα λόγια χωρίζουμε το δισδιάστατο χώρο σε τέσσερα κουτιά. Στη συνέχεια αν το κουτί περιέχει ένα ή περισσότερα σημεία δημιουργούμε ένα 'παιδί αντικείμενο' που σε αυτό αποθηκεύεται ο δισδιάστατος χώρος του κουτιού. Αν το κουτί δεν περιέχει κάποιο σημείο τότε δεν δημιουργείται παιδί. Αυτό γίνεται για όλα τα παιδιά. Για το συγκεκριμένο δέντρο δημιουργήσαμε μια κλάση 'quadTree' στην οποία αρχικά θέσαμε τα min και max και στην συνέχεια βάλαμε όλα τα data σε μια λίστα. Στην συνέχεια δημιουργήσαμε συνάρτηση η οποία δημιουργεί από το αρχικό δέντρο τέσσερα υποδέντρα. Με την συνάρτηση 'findNearestNeighbors' επιστρέφουμε μια λίστα με τους κοντινότερους γείτονες για κάθε υποδέντρο.

3.Range Tree

Το range tree είναι μια δομή δεδομένων διατεταγμένου δέντρου με σκοπό να κρατά μια λίστα από σημεία. Δίνει την δυνατότητα σε όλα τα σημεία μέσα σε μια συγκεκριμένη εμβέλεια να αναφέρονται αποτελεσματικά. Κυρίως το range tree χρησιμοποιείται για δυο ή περισσότερες διαστάσεις. Ένα range tree για μονοδιάστατα σημεία είναι ένα διατεταγμένο δυαδικό δέντρο. Τα σημεία αποθηκεύονται στα φύλλα του δέντρου, ενώ κάθε εσωτερικός κόμβος αποθηκεύει την μεγαλύτερη τιμή στο αριστερό υποδέντρο. Ένα range tree για σημεία πολλών διαστάσεων είναι αναδρομικά ορισμένο δυαδικό δέντρο πολλών επιπέδων. Στο πρώτο επίπεδο είναι ένα δυαδικό δέντρο αναζήτησης στην πρώτη από τις d-συντεταγμένες. Για το συγκεκριμένο δέντρο δημιουργήσαμε μια κλάση 'quadTree' στην οποία αρχικά θέσαμε τα min και max και στην συνέχεια βάλαμε όλα τα data σε μια λίστα. Εν συνέχεια με την συνάρτηση 'rangeTreeDivision' υποδιαιρέσαμε το αρχικό δέντρο σε δυο υποδέντρα 'leftSubtree' και 'rightSubtree'. Με την

συνάρτηση ‘findNearestNeighbors’ επιστρέφουμε μια λίστα με τους κοντινότερους γείτονες για τα δυο υποδέντρα.

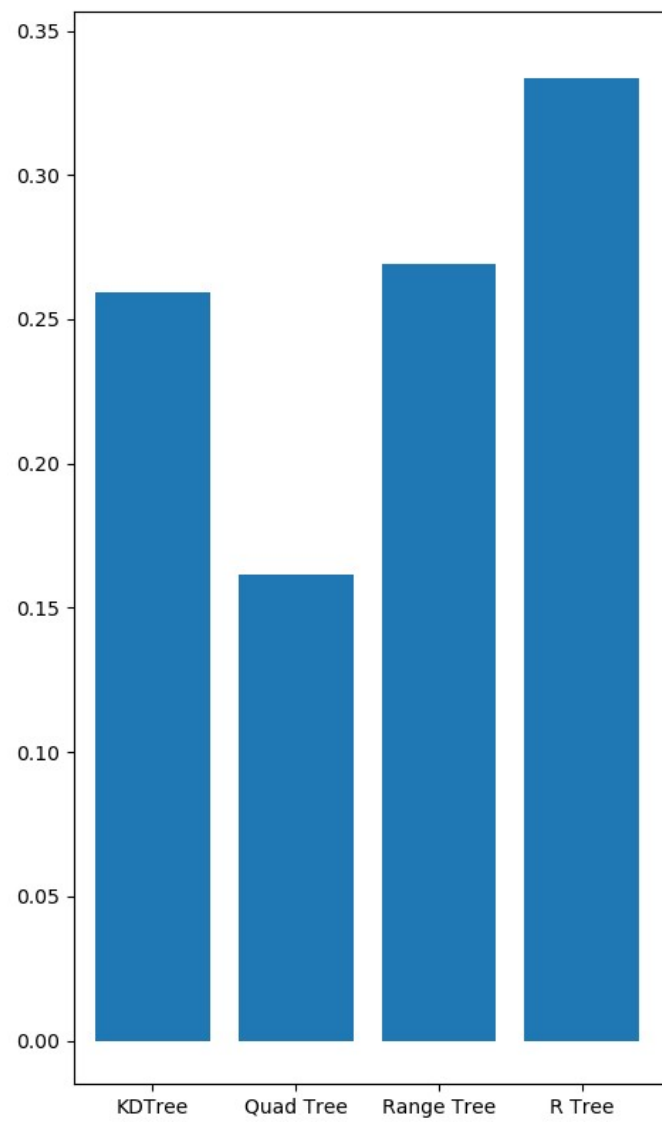
4.R – Tree

Τα R- tree είναι μια δομή δεδομένων δέντρου που χρησιμοποιείται για την αποθήκευση ευρετηρίου χωρικών δεδομένων με αποδοτικό τρόπο. Τα R – tree είναι πολύ βοηθητικά για ερωτήματα χωρικών δεδομένων και αποθήκευση αυτών. Χρησιμοποιούνται για πολυδιαστάτες πληροφορίες όπως γεωγραφικές συντεταγμένες, πολύγωνα ή ορθογώνια. Η βασική ιδέα είναι η ομαδοποίηση κοντινών αντικειμένων και η αναπαραστασή τους στο ελάχιστο οριοθετημένο ορθογώνιο στο ακριβώς επόμενο υψηλότερο επίπεδο του δέντρου. Στο επίπεδο των φύλλων κάθε ορθογώνιο είναι ένα αντικείμενο.

Αποτελέσματα Πειραμάτων:

```
see the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
articlesDataFrame.dropna(inplace=True)
Processing ..... Users Calculated: 40/40 Estimated Time Remaining: 0:00:00
Executing KD Tree, Quad Tree, Range Tree, R Tree functions using data from 31 articles
Data has been split to 27 train articles to be added to Trees and 4 articles for queries
..... KD Tree Demo.....
0.0004134178161621094 seconds elapsed to Insert 27 train articles to KD Tree. Average Insertion time: 1.5311770968967015e-05
Randomly selected article:
Han is the first North Korean player in the Serie A and was praised during his appearances during youth World cups.
The article's vector is [2530, 32, 82, 2544, 2562, 2572, 404, 3, 36, 2576]
Using KD Tree, its nearest neighbor is [1528, 187, 82, 1523, 1519, 1515, 480, 1510, 1506, 154]
Closest neighbor to use for category classification, according to KD Tree, is:
Israel says it will be dispatching a team of firefighters to assist Brazil with combatting the fires that have consumed large swaths of the country's rainforest
LSH Similarity: 0.25925925925925924
..... KD Tree Demo End.....
Press Enter for Quad Tree Demo...
.....Quad Tree Demo.....
0.004954338073730469 seconds elapsed to insert 27 train articles to KD Tree. Average insertion time: 0.0001834948027307581
Randomly selected article:
While rushing to strike an agreement with the Taliban ahead of the 2020 election, Washington is ignoring Afghan demands.
The article's vector is [2135, 230, 207, 2130, 2126, 2111, 32, 2107, 1964, 2516]
Using Quad Tree, its nearest neighbor list is: [[621, 647, 498, 3042, 35, 3037, 3031, 3025, 3020, 3014]]
Closest neighbors to use for category classification, according to Quad Tree, are:
Latest figures suggest Government is on course to hit its tax and spending targets
LSH Similarity: 0.16129832258064516
.....Quad Tree Demo End.....
Press Enter for Range Tree Demo...
.....Range Tree Demo.....
0.007526874542236320 seconds elapsed to Insert 27 train articles to KD Tree. Average insertion time: 0.0002787731311939381
Randomly selected article:
Private equity giant The Carlyle Group said on Tuesday it agreed to buy a majority stake in HireVue, a company that makes video and AI software for employers to screen job applicants. Carlyle can provide HireVue access to its technology advisers and introduce...
The article's vector is [958, 954, 951, 376, 943, 36, 54, 939, 935, 932]
Using Range Tree, its nearest neighbor list is: [[79, 288, 69, 354, 54, 64, 1220, 394, 901, 1212]]
Closest neighbors to use for category classification, according to Range Tree, are:
President Donald Trump suggested on Tuesday that a more than yearlong trade dispute with China could potentially last until 2020. He said that any agreement to defuse tensions would be 'tougher' if he were to win re-election. Any other candidate would allow t...
LSH Similarity: 0.2692387692307692
Press Enter for R Tree Demo...
.....R Tree Demo.....
0.002845287322998047 seconds elapsed to insert 27 train articles to KD Tree. Average insertion time: 0.00010538101196289062
Randomly selected article:
While rushing to strike an agreement with the Taliban ahead of the 2020 election, Washington is ignoring Afghan demands.
The article's vector is [2135, 230, 207, 2130, 2126, 2111, 32, 2107, 1964, 2516]
Using R Tree, its nearest neighbor list is: [[621, 647, 498, 3042, 35, 3037, 3031, 3025, 3020, 3014]]
Closest neighbors to use for category classification, according to Quad Tree, are:
Latest figures suggest Government is on course to hit its tax and spending targets
LSH Similarity: 0.3333333333333333
```

LSH Similarity



Insert Speed

