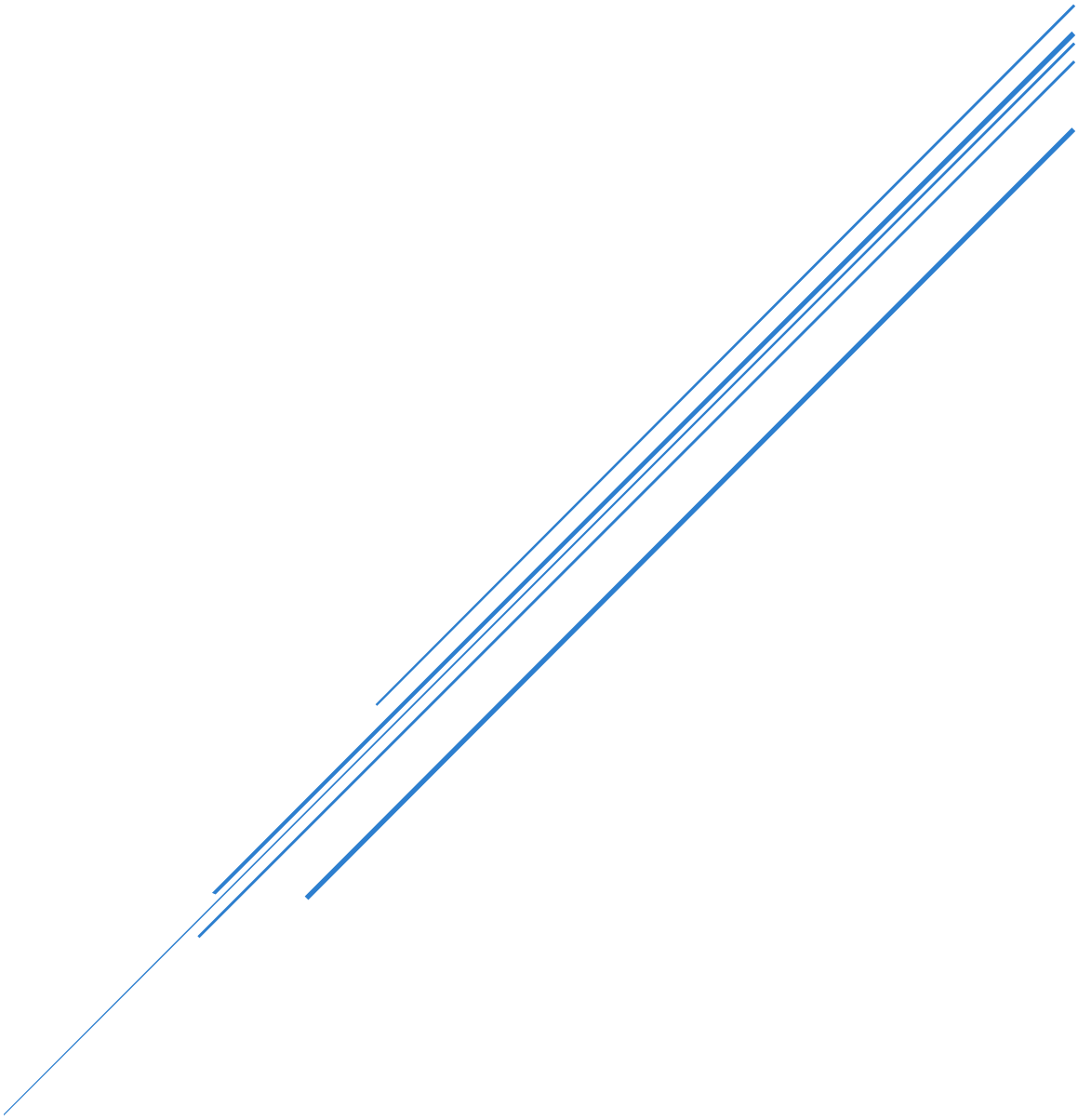
ΤΕΧΝΙΚΕΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΔΕΔΟΜΕΝΩΝ

Εργασία Μαθήματος



Περιεχόμενα

Τεχνική Αναφορά: Υλοποίηση Αλγορίθμου Top-K με Χρήση R-Tree για Επεξεργασία Χωρικών Δεδομένων

Σκοπός της Εργασίας

Ο σκοπός της συγκεκριμένης εργασίας είναι η υλοποίηση του αλγορίθμου Top-k με τη χρήση του χωρικού ευρετηρίου R-Tree. Ο στόχος είναι να εντοπιστούν τα Κ καλύτερα σημεία (στο παράδειγμά μας: ξενοδοχεία) από το dataset, βάσει κριτηρίων όπως η απόσταση, η βαθμολογία (rating) και ο αριθμός κριτικών (reviews), με δυνατότητα ορισμού βαρών στη γραμμική συνάρτηση (linear ή weighted sum) από τον χρήστη. Η συνάρτηση αυτή δηλώνει τις προτιμήσεις του χρήστη (preference function), ουσιαστικά, μία μεγάλη τιμή βάρους σε ένα γνώρισμα δηλώνει υψηλή προτίμηση για αυτό το γνώρισμα.

Η εργασία επεκτείνεται με τη δημιουργία διαδραστικού χάρτη που απεικονίζει τα αποτελέσματα χρησιμοποιώντας πραγματικά γεωγραφικά δεδομένα από το TripAdvisor (https://www.kaggle.com/datasets/justdataplease/tripadvisor-hotels-greece).

Περιγραφή του Προβλήματος

Το πρόβλημα που έχουμε είναι η εύρεση των Κ κορυφαίων σημείων ενδιαφέροντος (ξενοδοχείων) από ένα dataset που περιλαμβάνει γεωγραφικά δεδομένα, με βάση τα εξής:

1. **Γεωγραφική Απόσταση** από ένα σημείο αναφοράς.
2. **Βαθμολογία (Rating)**, που αντιπροσωπεύει την ποιότητα του ξενοδοχείου.
3. **Αριθμό Κριτικών (Number of Reviews)**, που υποδηλώνει τη δημοτικότητα του ξενοδοχείου.

Ο χρήστης μπορεί να ορίσει τα βάρη που αντιστοιχούν σε κάθε κριτήριο (π.χ. 50% απόσταση, 30% rating, 20% αριθμός κριτικών). Ο αλγόριθμος υπολογίζει ένα συνολικό σκορ για κάθε σημείο και επιστρέφει τα Κ καλύτερα βάσει αυτού.

Δομή και Υλοποίηση

Η εφαρμογή είναι γραμμένη σε Python και βασίζεται στις βιβλιοθήκες Pandas, GeoPandas, Rtree, Shapely και Folium.

Περιγραφή Εισόδου

Το πρόγραμμα χρησιμοποιεί ένα αρχείο CSV που περιέχει δεδομένα για ξενοδοχεία στην Ελλάδα. Κάθε εγγραφή περιλαμβάνει:

- Γεωγραφικό πλάτος (latitude) και μήκος (longitude).
- Όνομα ξενοδοχείου
- Βαθμολογία (rating)
- Αριθμό κριτικών (number of reviews)

Σημείωση: Το αρχείο CSV έχει όνομα tripadvisor_hotels_greece_202210.csv και βρίσκεται στον ίδιο φάκελο με τον πηγαίο κώδικα και το συγκεκριμένο report.

Περιγραφή Εξόδου

Το πρόγραμμα επιστρέφει:

1. Τα K κορυφαία ξενοδοχεία σε μορφή κειμένου, με πληροφορίες για το όνομα, τη βαθμολογία, τον αριθμό κριτικών, την απόσταση και το συνολικό σκορ.
2. Έναν διαδραστικό χάρτη που εμφανίζει:
    - Όλα τα ξενοδοχεία (πράσινα points)
    - Τα Top-K ξενοδοχεία (κόκκινα points)
    - Το σημείο αναφοράς του χρήστη (μπλε points)

Περιγραφή Μεθόδων

Δημιουργία Γεωμετρίας και R-Tree

- **Κατασκευή R-Tree:** Το πρόγραμμα χρησιμοποιεί τη δομή R-Tree για την αποδοτική αποθήκευση και αναζήτηση των γεωγραφικών σημείων. Τα δεδομένα εισάγονται στο R-Tree με συντεταγμένες σε μορφή (x_min, y_min, x_max, y_max) για κάθε σημείο.
- **Μετατροπή σε GeoDataFrame:** Η γεωμετρία κάθε ξενοδοχείου (γεωγραφικό σημείο) δημιουργείται μέσω της βιβλιοθήκης Shapely και αποθηκεύεται σε GeoDataFrame για περαιτέρω χωρική ανάλυση.

Υπολογισμός Απόστασης

Υπολογίζεται η Ευκλείδεια απόσταση μεταξύ δύο σημείων.

Κανονικοποίηση Τιμών

Οι τιμές για την απόσταση, τη βαθμολογία και τον αριθμό κριτικών κανονικοποιούνται σε κλίμακα [0, 1] για να είναι συγκρίσιμες. Ακόμη, έχει γίνει τροποποίηση και μικρότερες αποστάσεις λαμβάνουν μεγαλύτερη βαρύτητα(1 – distance/max distance).

Υπολογισμός Σκορ

Το συνολικό σκορ ενός σημείου υπολογίζεται με βάση τη παρακάτω συνάρτηση:

$$score = w_d * (1 - normalized\ distance) + w_r * normalized\ rating + w_n * normalized,$$

οπου w_d, w_r, και w_n είναι τα βάρη για την απόσταση, τη βαθμολογία, και τον αριθμό κριτικών αντίστοιχα. Οι τιμές κανονικοποιούνται, και εφαρμόζεται ποινή στις μεγάλες αποστάσεις (penalized_distance = (1 - norm_distance) ** 2).

Αλγόριθμος Top-K

Βήματα:

1. Αναζήτηση υποψήφιων σημείων από το R-Tree.
2. Υπολογισμός της απόστασης και κανονικοποίηση των χαρακτηριστικών για κάθε σημείο.
3. Υπολογισμός του συνολικού σκορ βάσει των βαρών.
4. Ταξινόμηση σημείων κατά φθίνουσα σειρά σκορ και επιλογή των πρώτων K.

Δημιουργία Χάρτη

Ο χάρτης δημιουργείται με τη χρήση της βιβλιοθήκης Folium.

Κάθε σημείο εμφανίζεται ως δείκτης με αναδυόμενο παράθυρο που περιλαμβάνει πληροφορίες για το ξενοδοχείο.

Χρήση του Προγράμματος

Είσοδος Χρήστη

Το πρόγραμμα ζητά από το χρήστη:

1. Συντεταγμένες σημείου αναφοράς (γεωγραφικό μήκος και πλάτος).
2. Αριθμό αποτελεσμάτων K.
3. Βάρη για την απόσταση, τη βαθμολογία και τον αριθμό κριτικών (ως ποσοστά που αθροίζουν στο 100%).

Εκτέλεση Προγράμματος

Τα αποτελέσματα εμφανίζονται στη console και στον χάρτη. Ο χάρτης αποθηκεύεται ως αρχείο HTML που μπορεί να ανοίξει σε οποιονδήποτε browser.

Συμπεράσματα

Το πρόγραμμα υλοποιεί έναν αποδοτικό αλγόριθμο Top-K χρησιμοποιώντας τη δομή R-Tree, επιτρέποντας τη γρήγορη αναζήτηση χωρικών δεδομένων με βάση πολλαπλά κριτήρια. Η χρήση πραγματικών δεδομένων και η δημιουργία χάρτη ενισχύουν την πρακτική αξία της εφαρμογής.

Αναφορές

Dataset

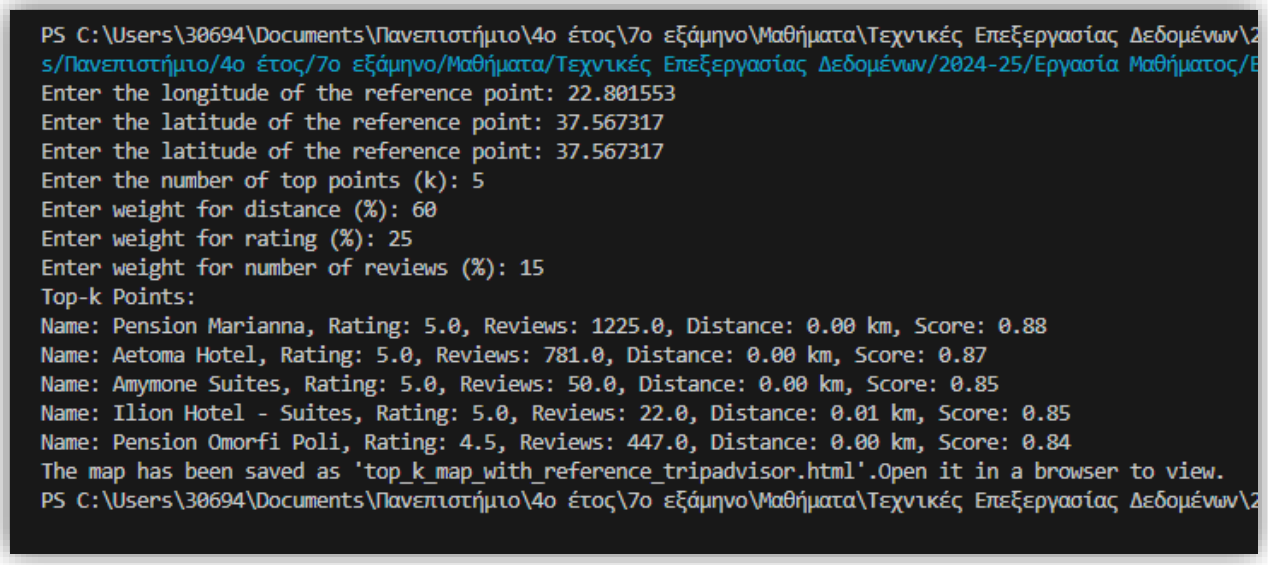
TripAdvisor Dataset (Kaggle):
https://www.kaggle.com/datasets/justdataplease/tripadvisor-hotels-greece

Βιβλιοθήκες

- Pandas: Διαχείριση δεδομένων (https://pypi.org/project/pandas/)
- GeoPandas: Επεξεργασία γεωχωρικών δεδομένων (https://github.com/geopandas/geopandas)
- Shapely: Δημιουργία γεωμετρικών αντικειμένων (https://github.com/shapely/shapely)
- Rtree: Δομή ευρετηρίου για αποδοτική αναζήτηση (https://github.com/Toblerity/rtree, https://rtree.readthedocs.io/en/latest/class.html)
- Folium: Δημιουργία χαρτών (https://www.geeksforgeeks.org/python-plotting-google-map-using-folium-package/)

Παράδειγμα Χρήσης του Προγράμματος

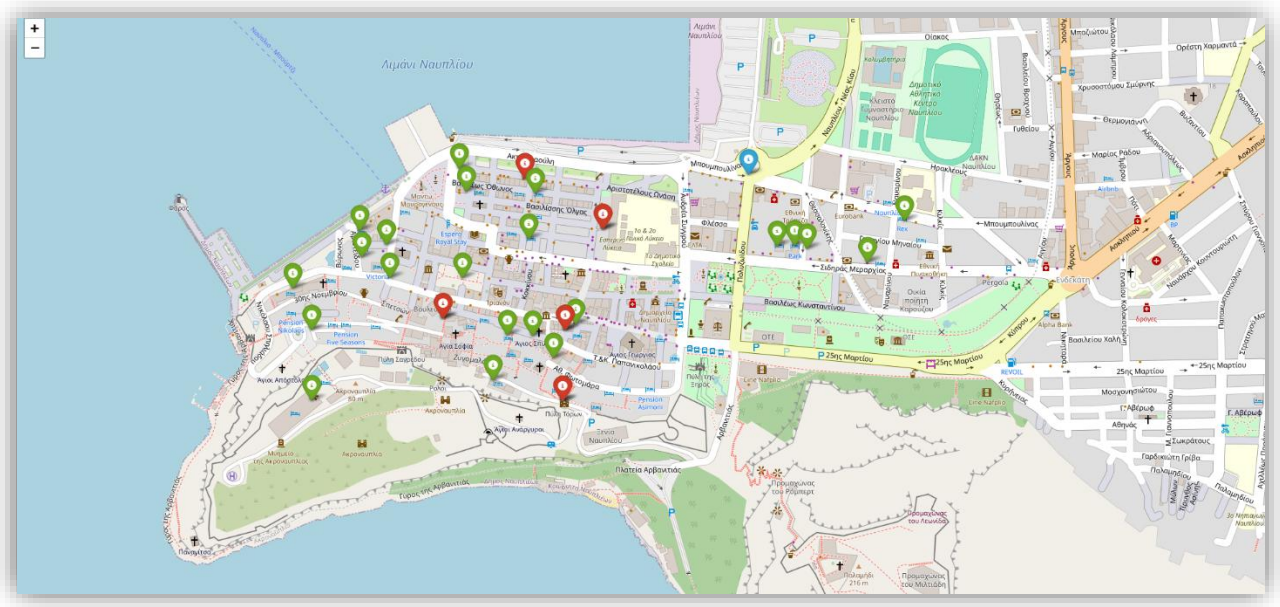
Στο παρακάτω παράδειγμα, χρησιμοποιήθηκαν συντεταγμένες που αντιστοιχούν στο Ναύπλιο (γεωγραφικό πλάτος: 37.567317, γεωγραφικό μήκος: 22.801553) ως σημείο αναφοράς. Στόχος ήταν να εντοπιστούν τα 5 καλύτερα ξενοδοχεία βάσει της απόστασης, της βαθμολογίας και του αριθμού των κριτικών, με διαφορετικά βάρη που ορίστηκαν από τον χρήστη.



Παράμετροι

- Αριθμός ξενοδοχείων (k): 5
- Βάρη:
    - Απόσταση: 60%
    - Βαθμολογία: 25%
    - Αριθμός κριτικών: 15%

Ανοίγοντας το αρχείο top_k_map_with_reference_tripadvisor.html στο browser μας, έχουμε:



<u>Παρατήρηση:</u> Λόγω των πολλών ξενοδοχείων (δεδομένα του dataset) ενδέχεται ο χάρτης να καθυστερεί να φορτώσει.