ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΥΕ041 - ΠΛΕ081: Διαχείριση Σύνθετων Δεδομένων

(ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2019-20)

ΕΡΓΑΣΙΑ 4 – Ανεστραμμένα αρχεία και γεω-κειμενική αναζήτηση

Προθεσμία: 30 Ιουνίου 2020, 9μ.μ.

Στο ecourse βρείτε και κατεβάστε το αρχείο Restaurants_London_England.tsv, το οποίο περιέχει πληροφορίες για εστιατόρια στο Λονδίνο. Σε κάθε γραμμή υπάρχουν α) ένα όνομα αρχείου το οποίο έχει μία περιγραφή και reviews για το εστιατόριο (τα αρχεία αυτά δεν δίνονται, αλλά υπάρχουν στον Ιστό για όποιον ενδιαφέρεται), β) οι συντεταγμένες της θέσης του εστιατορίου, και γ) μία ακολουθία από ετικέτες (tags) οι οποίες περιγράφουν το εστιατόριο.

Στόχος της εργασίας είναι να φτιάξετε ένα κειμενικό και ένα απλό χωρικό ευρετήριο για τα δεδομένα και να τα χρησιμοποιήσετε για αναζήτηση εστιατορίων με βάση (α) ετικέτες, (β) χωρική περιοχή, (γ) ετικέτες και χωρική περιοχή.

Μέρος 1: Ανεστραμμένο αρχείο και αναζήτηση με λέξεις-κλειδιά

Διαβάστε τα δεδομένα από το αρχείο Restaurants_London_England.tsv σε έναν πίνακα (ή λίστα) όπου η κάθε εγγραφή είναι μία γραμμή του αρχείου. Καθώς διαβάζετε τα δεδομένα, δημιουργήστε ένα ανεστραμμένο αρχείο (inverted file) στην κύρια μνήμη, όπου για κάθε ετικέτα καταγράψτε **σε αύξουσα σειρά** τους αριθμούς των γραμμών (εγγραφών) που περιέχουν αυτή την ετικέτα. Για παράδειγμα, η γραμμή 0 περιέχει τις ετικέτες chinese και thai, άρα οι ανεστραμμένες λίστες των chinese και thai θα πρέπει να περιέχουν (μεταξύ άλλων) και τον αριθμό 0. Μέσω των γραμμών που έχουμε γράψει στις λίστες θα μπορούμε να ανακτήσουμε τις εγγραφές των εστιατορίων που περιέχουν τις αντίστοιχες ετικέτες, κάνοντας **merge-join** τις λίστες οι οποίες **πρέπει να είναι ήδη ταξινομημένες**.

Τυπώστε τον αριθμό των διακριτών keywords (tags) που εμφανίζονται στα δεδομένα και τη συχνότητα αυτών σε αύξουσα σειρά. Προσοχή: υπάρχουν και tags τα οποία αποτελούνται από δύο λέξεις (π.χ. “late night”). Αυτά αντιμετωπίζονται σαν ένα keyword/tag.

Υλοποιήστε μια συνάρτηση kwSearchIF, η οποία παίρνει σαν όρισμα μία ακολουθία από query keywords και χρησιμοποιεί το ανεστραμμένο αρχείο για να βρει τις εγγραφές που περιέχουν όλα τα query keywords.

Υλοποιήστε επίσης μια συνάρτηση kwSearchRaw, η οποία παίρνει σαν όρισμα μία λίστα από query keywords, διαβάζει όλες τις εγγραφές και υπολογίζει και επιστρέφει εκείνες που περιέχουν τα query keywords (χωρίς τη χρήση του ανεστραμμένου αρχείου).

Παραδοτέο: Ένα πρόγραμμα το οποίο παίρνει σαν command-line arguments τα query keywords και τυπώνει τις γραμμές του αρχείου που αντιστοιχούν σε εστιατόρια τα οποία **περιέχουν όλα τα query keywords** στα tags τους. Το πρόγραμμα θα εκτελεί τις kwSearchIF και kwSearchRaw και **θα μετράει και θα τυπώνει τα αποτελέσματα** και των δύο (με αυτό τον τρόπο μπορείτε να επιβεβαιώσετε την ορθότητα των συναρτήσεών σας), καθώς και **το χρόνο εκτέλεσής τους**. Προσοχή: ένα biword εκφράζεται στο command-line με μονά εισαγωγικά (π.χ. 'late night').

Παράδειγμα:

```
./part1 greek bar

number of keywords: 168
frequencies: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2,
2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5,
5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 10, 10, 12, 12, 13, 13, 13, 13,
14, 15, 16, 16, 16, 18, 19, 20, 21, 24, 25, 25, 25, 25, 26, 27, 27, 28,
28, 31, 32, 33, 35, 35, 37, 38, 39, 40, 40, 41, 44, 46, 46, 47, 47, 49,
50, 53, 56, 59, 59, 60, 64, 66, 68, 68, 69, 70, 73, 75, 86, 88, 92, 93,
94, 95, 95, 105, 112, 127, 146, 147, 149, 150, 160, 164, 165, 166, 185,
205, 240, 256, 261, 276, 298, 315, 327, 333, 337, 338, 339, 344, 353,
353, 355, 368, 380, 395, 404, 466, 499, 499, 602, 714, 826, 936, 972,
1346, 1380, 1386, 2062, 3012]

kwSearchRaw: 1 results, cost = 0.009675979614257812 seconds
Restaurant_Review-g186338-d1015433-Reviews-Costas_Grill-
London_England.html  location: 51.508144,-0.197733 tags:
american,greek,bar,grill

kwSearchIF: 1 results, cost = 0.0005671977996826172 seconds
Restaurant_Review-g186338-d1015433-Reviews-Costas_Grill-
London_England.html   location: 51.508144,-0.197733    tags:
american,greek,bar,grill
```

Μέρος 2: Χωρικό ευρετήριο και χωρική αναζήτηση

Διαβάστε τα δεδομένα από το αρχείο Restaurants_London_England.tsv σε έναν πίνακα (ή λίστα) όπου η κάθε εγγραφή είναι μία γραμμή του αρχείου. Φτιάξτε ένα απλό χωρικό ευρετήριο βασισμένο σε σχάρα (grid) από τα δεδομένα του αρχείου. Η σχάρα χωρίζει το χώρο που καλύπτουν τα σημεία σε 50*50 = 2500 ισομεγέθη ορθογώνια (**κελιά**). Για να δημιουργήσετε το grid θα πρέπει να διαβάσετε τις θέσεις των εστιατορίων και να βρείτε τη μικρότερη και τη μεγαλύτερη τιμή σε κάθε συντεταγμένη (x και y). Κατόπιν, χωρίστε το εύρος τιμών σε κάθε συντεταγμένη σε 50 ίσα διαστήματα τιμών. Π.χ. τα 50 διαστήματα στον χ άξονα, όπου το εύρος τιμών είναι [50.546856, 55.95297] πρέπει να είναι [50.546856, 50.65497828), [50.65497828, 50.76310056) ,…, [55.84484772,55.95297]. Δημιουργήστε μία

δισδιάστατη δομή (π.χ. πίνακα) όπου η θέση [i][j] αντιστοιχεί στο κελί (i,j). Σε κάθε θέση του πίνακα-grid αποθηκεύστε τους αριθμούς των γραμμών (εγγραφών) των εστιατορίων που πέφτουν μέσα στο αντίστοιχο διάστημα τιμών. Π.χ. το grid[5][36] πρέπει να έχει τα εστιατόρια [3918, 4902, 5786, 7265, 8858], στα οποία οι συντεταγμένες είναι μέσα στο κελλί με όρια [51.0874674, 51.19558968] στον χ άξονα και [-0.18886772, -0.07623624] στον y άξονα.

Τυπώστε για κάθε διάσταση την μικρότερη και τη μεγαλύτερη τιμή των συντεταγμένων των εστιατορίων και το εύρος τιμών σε κάθε διάσταση. Αφού κατασκευάσετε το grid, **τυπώστε** για κάθε κελί που δεν είναι άδειο τον αριθμό των εστιατορίων σε αυτό.

Κατόπιν υλοποιήστε μία συνάρτηση spaSearchGrid η οποία παίρνει σαν όρισμα μια δισδιάστατη ορθογώνια περιοχή (range query) και υπολογίζει και επιστρέφει τα εστιατόρια μέσα σε αυτή χρησιμοποιώντας το grid. Η περιοχή ορίζεται από το κάτω και το πάνω όριο σε κάθε διάσταση. Υλοποιήστε επίσης μία συνάρτηση spaSearchRaw, η οποία παίρνει σαν όρισμα μία range query και υπολογίζει και επιστρέφει τα εστιατόρια μέσα σε αυτή απ' ευθείας πάνω στα δεδομένα του πίνακα, δηλαδή χωρίς τη χρήση του grid.

Παραδοτέο: Ένα πρόγραμμα το οποίο παίρνει σαν command-line arguments τα όρια του query range και τυπώνει τις γραμμές του αρχείου που αντιστοιχούν σε εστιατόρια τα οποία **περιέχονται στο query range**. Το πρόγραμμα θα εκτελεί τις spaSearchGrid και spaSearchRaw και **θα μετράει και θα τυπώνει τα αποτελέσματα** και των δύο (με αυτό τον τρόπο μπορείτε να επιβεβαιώσετε την ορθότητα των συναρτήσεών σας), καθώς και **το χρόνο εκτέλεσής τους**.

Παράδειγμα:

```
./part2.py 51 51.20 -0.5 0
bounds: 50.546856 55.95297 -4.243601 1.387973
widths: 5.406114000000002 5.631574
0 2 1
4 29 1
5 36 5
6 32 1
6 34 1
6 39 1
6 49 1
...

spaSearchRaw: 5 results, cost = 0.0134301186 seconds
Restaurant_Review-g186338-d3318593-Reviews-Jamie_Oliver_s_Restaurant-
London_England.html  location: 51.159298,-0.172481 tags: international
Restaurant_Review-g186338-d3381895-Reviews-Joes_Kitchen_Coffee_House-
London_England.html  location: 51.161,-0.172959    tags: central european
Restaurant_Review-g186338-d2343541-Reviews-Garfunkel_s_at_Gatwick_Aiport-
London_England.html  location: 51.160873,-0.179375 tags:
european,breakfast/brunch,late night
Restaurant_Review-g186338-d4046898-Reviews-Jamie_s_italian-
London_England.html  location: 51.16089,-0.176477  tags: italian
Restaurant_Review-g186338-d2375022-Reviews-Garfunkel_s-
London_England.html  location: 51.160877,-0.17403  tags: english
```

```
spaSearchGrid: 5 results, cost = 0.0000488758 seconds
Restaurant_Review-g186338-d3318593-Reviews-Jamie_Oliver_s_Restaurant-
London_England.html  location: 51.159298,-0.172481 tags: international
Restaurant_Review-g186338-d3381895-Reviews-Joes_Kitchen_Coffee_House-
London_England.html  location: 51.161,-0.172959   tags: central european
Restaurant_Review-g186338-d2343541-Reviews-Garfunkel_s_at_Gatwick_Aiport-
London_England.html  location: 51.160873,-0.179375 tags:
european,breakfast/brunch,late night
Restaurant_Review-g186338-d4046898-Reviews-Jamie_s_italian-
London_England.html  location: 51.16089,-0.176477  tags: italian
Restaurant_Review-g186338-d2375022-Reviews-Garfunkel_s-
London_England.html   location: 51.160877,-0.17403    tags: english
```

Μέρος 3: Χωρο-κειμενική αναζήτηση

Γράψτε ένα πρόγραμμα, το οποίο θα διαβάζει τα δεδομένα και θα φτιάχνει και το ανεστραμμένο αρχείο και το grid που υλοποιήσατε στα Μέρη 1 και 2.

Υλοποιήστε μία συνάρτηση kwSpaSearchIF η οποία παίρνει σαν όρισμα ένα query range και μία λίστα από query keywords και υπολογίζει και επιστρέφει τα εστιατόρια που περιέχονται στο query range και περιέχουν όλα τα query keywords στα tags τους. Για το σκοπό αυτό χρησιμοποιεί το ανεστραμμένο αρχείο να βρει τα εστιατόρια που περιέχουν τα query keywords και για το καθένα από αυτά επαληθεύει αν αυτό είναι μέσα στο query range.

Υλοποιήστε μία συνάρτηση kwSpaSearchGrid η οποία παίρνει σαν όρισμα ένα query range και μία λίστα από query keywords και υπολογίζει και επιστρέφει τα εστιατόρια που περιέχονται στο query range και περιέχουν όλα τα query keywords στα tags τους. Για το σκοπό αυτό χρησιμοποιεί το grid να βρει τα εστιατόρια που περιέχονται στο query range και για το καθένα από αυτά επαληθεύει αν αυτό περιέχει όλα τα query keywords στα tags του.

Υλοποιήστε μία συνάρτηση kwSpaSearchRaw η οποία παίρνει σαν όρισμα ένα query range και μία λίστα από query keywords και υπολογίζει και επιστρέφει τα εστιατόρια που περιέχονται στο query range και περιέχουν όλα τα query keywords στα tags τους. Η συνάρτηση απλά εξετάζει ένα-ένα τα εστιατόρια και ελέγχει το αν βρίσκεται μέσα στο query range και το αν περιέχει όλα τα query keywords χωρίς τη χρήση του ανεστραμμένου αρχείου και χωρίς τη χρήση του grid.

Παραδοτέο: Ένα πρόγραμμα το οποίο παίρνει σαν command-line arguments τα όρια του query range και τουλάχιστον ένα query keyword και τυπώνει τις γραμμές του αρχείου που αντιστοιχούν σε εστιατόρια τα οποία **περιέχονται στο query range και περιέχουν όλα τα query keywords**. Το πρόγραμμα θα εκτελεί τις kwSpaSearchIF, kwSpaSearchGrid και kwSpaSearchRaw και **θα μετράει και θα τυπώνει τα αποτελέσματα** και των τριών (με αυτό τον τρόπο μπορείτε να επιβεβαιώσετε την ορθότητα των συναρτήσεών σας), καθώς και **το χρόνο εκτέλεσής τους**.

Παράδειγμα:

```
./part3.py 51 51.50 -0.5 0 british bar
```

```
kwSpaSearchRaw: 5 results, cost = 0.0097060204 seconds
Restaurant_Review-g186338-d944622-Reviews-Yacht-London_England.html
    location: 51.484703,-0.00394  tags: english,british,bar,grill
Restaurant_Review-g186338-d734073-Reviews-Castle-London_England.html
    location: 51.47279,-0.173312  tags: british,bar,grill
Restaurant_Review-g186338-d1017689-Reviews-Society_Bar_Restaurant-
London_England.html  location: 51.496468,-0.206702 tags:
contemporary,international,british,bar,reservations,private dining
Restaurant_Review-g186338-d806450-Reviews-THE_TERRACE_KITCHEN_BAR-
London_England.html  location: 51.412285,-0.12382  tags:
caribbean,vegetarian,british,bar,families with children,romance,outdoor
seating,breakfast/brunch
Restaurant_Review-g186338-d817313-Reviews-Barstory-London_England.html
    location: 51.46933,-0.070317  tags:
british,bar,bistro,romance,outdoor seating

kwSpaSearchIF: 5 results, cost = 0.0021228790 seconds
Restaurant_Review-g186338-d944622-Reviews-Yacht-London_England.html
    location: 51.484703,-0.00394  tags: english,british,bar,grill
Restaurant_Review-g186338-d734073-Reviews-Castle-London_England.html
    location: 51.47279,-0.173312  tags: british,bar,grill
Restaurant_Review-g186338-d1017689-Reviews-Society_Bar_Restaurant-
London_England.html  location: 51.496468,-0.206702 tags:
contemporary,international,british,bar,reservations,private dining
Restaurant_Review-g186338-d806450-Reviews-THE_TERRACE_KITCHEN_BAR-
London_England.html  location: 51.412285,-0.12382  tags:
caribbean,vegetarian,british,bar,families with children,romance,outdoor
seating,breakfast/brunch
Restaurant_Review-g186338-d817313-Reviews-Barstory-London_England.html
    location: 51.46933,-0.070317  tags:
british,bar,bistro,romance,outdoor seating

kwSpaSearchGrid: 5 results, cost = 0.0066959858 seconds
Restaurant_Review-g186338-d1017689-Reviews-Society_Bar_Restaurant-
London_England.html  location: 51.496468,-0.206702 tags:
contemporary,international,british,bar,reservations,private dining
Restaurant_Review-g186338-d734073-Reviews-Castle-London_England.html
    location: 51.47279,-0.173312  tags: british,bar,grill
Restaurant_Review-g186338-d806450-Reviews-THE_TERRACE_KITCHEN_BAR-
London_England.html  location: 51.412285,-0.12382  tags:
caribbean,vegetarian,british,bar,families with children,romance,outdoor
seating,breakfast/brunch
Restaurant_Review-g186338-d944622-Reviews-Yacht-London_England.html
    location: 51.484703,-0.00394  tags: english,british,bar,grill
Restaurant_Review-g186338-d817313-Reviews-Barstory-London_England.html
     location: 51.46933,-0.070317    tags:
british,bar,bistro,romance,outdoor seating
```

Μέρος 4: Γραπτό

Γράψτε μία αναφορά όπου εξηγείτε συνοπτικά τη λειτουργία των προγραμμάτων και συναρτήσεών σας. Αφού δοκιμάσετε να τρέξετε τις συναρτήσεις με αρκετές ερωτήσεις, γράψτε τα συμπεράσματά σας σχετικά με:

α) την απόδοση της kwSearchIF σε σχέση με την kwSearchRaw στο Μέρος 1

β) την απόδοση της spaSearchGrid σε σχέση με την spaSearchRaw στο Μέρος 2

γ) τη σχετική απόδοση των τριών συναρτήσεων στο Μέρος 3

Παραδοτέα: Κάντε turnin στο assignment4@mye041 τα προγράμματά σας και ένα PDF αρχείο για το Μέρος 4.

Οδηγίες για τις υποβολές:

1) Μπορείτε να χρησιμοποιήσετε δομές όπως priority queue ή heap από τις βιβλιοθήκες της γλώσσας προγραμματισμού (π.χ. το module heapq της Python).
2) Αν χρησιμοποιήσετε Java, το πρόγραμμά σας θα πρέπει να γίνεται compile και να τρέχει και εκτός Eclipse στους υπολογιστές του εργαστηρίου. **Μην χρησιμοποιείτε packages**.
3) Αν χρησιμοποιήσετε Python, μην χρησιμοποιήστε τη βιβλιοθήκη pandas και μην υποβάλετε κώδικα για interactive programming (π.χ. ipython)
4) Υποβάλετε τις εργασίες σας σε ένα **zip** αρχείο (**όχι rar**) το οποίο πρέπει να περιλαμβάνει όλους τους κώδικες καθώς και ένα αρχείο τεκμηρίωσης το οποίο να περιγράφει τη μεθοδολογία σας και να περιλαμβάνει το PDF αρχείο. **Μην υποβάλετε αρχεία δεδομένων**.
5) Μην ξεχνάτε να βάζετε το όνομά σας (σε greeklish) και το ΑΜ σε κάθε αρχείο που υποβάλετε.