Πρόγραμμα " Σχέδιο Συμφωνίας Συμβιβασμού μεταξύ της Ελληνικής Δημοκρατίας και των εταιριών SiemensAG και Siemens A.E"

Έργο: Αλυσίδες Αξίας Δεδομένων σε Βιομηχανικά και Αστικά Περιβάλλοντα

Υποέργο: CitySense: Δυναμική, Διαδραστική και Πληθοποριστική Αστική Ανάλυση και Βιώσιμη Κινητικότητα

Αστική υποδομή δεδομένων: Αλγόριθμοι ολοκλήρωσης και μοντέλα αποθήκευσης

Παραδοτέο Π1.1

Συγγραφείς

ΑλέξανδροςΕφεντάκης, Χάρης Νάκος, Δανάη Πλα-Καρύδη,

Γιάννης Σταύρακας, Θοδωρής Δαλαμάγκας

Ημερομηνία: 31/03/2016

Επικαιροποιημένο την: 30/09/2016

«Αθηνά» - Ερευνητικό Κέντρο Καινοτομίας στις Τεχνολογίες της Πληροφορίας, των Επικοινωνιών και της Γνώσης

Πίνακας περιεχομένων

Περίληψη

Ένας από τους βασικούς στόχους του έργου Citysense και της εφαρμογής που θα υλοποιηθεί στα πλαίσια του συγκεκριμένου έργου, είναι η ρεαλιστική απεικόνιση του "ίχνους" μιας αστικής περιοχής σε πολλαπλά επίπεδα. Δηλαδή, η χρήση διαφόρων κατηγοριών δεδομένων προκειμένου να συντεθεί η εικόνα μιας πόλης για κάθε μια από τις περιοχές της: οι δραστηριότητες σε διαφορετικές χρονικές περιόδους, η κίνηση και επικοινωνία, οι υποδομές και η χρήση τους, το κλίμα και οι συνθήκες διαβίωσης, οι απόψεις του πληθυσμού. Βασικός πυλώνας του στόχου αυτού είναι η εκμετάλλευση όσο το δυνατόν περισσότερων δεδομένων που αφορούν τη συγκεκριμένη περιοχή, είτε αυτά προέρχονται από επίσημες πηγές (δήμους, περιφέρειες, κρατικές υπηρεσίες), είτε από δικτυακές πηγές μέσω API (π.χ., GooglePlacesAPI), είτε χρησιμοποιώντας περιεχόμενο που προέρχεται από χρήστες μέσων κοινωνικής δικτύωσης (Twitter, Flickr). Ως περιοχή πιλότος του έργου, επιλέχτηκε η πόλη του Chicago, κυρίως λόγω του πλούτου των "επίσημων" δεδομένων που υπάρχουν διαθέσιμα και προέρχονται από την αντίστοιχη κρατική υπηρεσία της συγκεκριμένης πόλης. Στο παρόν παραδοτέο θα περιγράψουμε λοιπόν δύο κυρίως αντικείμενα:

α) Την εφαρμογή που αναπτύχθηκε (και στην οποία θα αναφερόμαστε εφεξής ως "Area-Profiler") προκειμένου να αντλήσουμε δεδομένα για μια γεωγραφική περιοχή που προέρχονται από διαδικτυακά APIs (π.χ. GooglePlacesAPI, FoursquareAPI, Twitter API) και

β) τα ανοικτά δεδομένα που υπάρχουν για τη συγκεκριμένη περιοχή και τα οποία μπορούμε να κατεβάσουμε στην ολότητά τους για τοπική επεξεργασία, είτε αυτά προέρχονται από επίσημες πηγές (την κρατική υπηρεσία της πόλης του Chicago), είτε από διαδικτυακούς τόπους (δεδομένα καιρού από το WeatherUnderground[1]), όπως και δεδομένα που προέρχονται από τη γνωστή υπηρεσία πληθοπορισμού χαρτών OpenStreetMaps[2].

[1]https://www.wunderground.com/
[2] http://www.openstreetmap.org

1 Area Profiler

1.1 Εισαγωγή

Για τις ανάγκες του CitySense χρειαζόμαστε ένα εργαλείο που να δίνει τη δυνατότητα λήψης όλων των σημείων ενδιαφέροντος μίας μεγάλης αστικής περιοχής που να καλύπτει μία ολόκληρη πόλη. Τοσυγκεκριμένο εργαλείο που αναπτύξαμε, το αποκαλούμεAreaProfiler. Η βασική λειτουργία του AreaProfiler είναι η λήψη όλων των σημείων ενδιαφέροντος (POIs) μίας μεγάλης περιοχής, τα οποία είναι διαθέσιμα μέσω των παρεχόμενων APIs του GooglePlaces και του Foursquare, καθώς και του περιεχομένου που προέρχεται από χρήστες του Twitter και έχει γεωχωρική πληροφορία, το οποίο είναι διαθέσιμο μέσω του TwitterAPI. Επιπλέον η λειτουργία του AreaProfiler περιλαμβάνει τη μόνιμη αποθήκευση αυτών σε βάση δεδομένων, ώστε να είναι διαθέσιμα για προβολή και επεξεργασία. Ο AreaProfilereίναι ένα webservice και επομένως χρησιμοποιεί ένα webserver όπως είναι ο ApacheTomcatπροκειμένου να γίνει deploy. Επίσης χρησιμοποιεί το σύστημα διαχείρισης βάσης δεδομένωνPostgreSQL, όπου αποθηκεύονται συστηματικά τα δεδομένα συλλογής καθώς και οι παράμετροι του συστήματος.

1.2 Λεξιλόγιο

Για τη διευκόλυνση της ανάγνωσης της ενότητας, παραθέτουμε ορισμούς για τις βασικότερες έννοιες του AreaProfiler.

Σημείο Ενδιαφέροντος (POI): Περιγράφει μια τοποθεσία στην οποία κάποιος μπορεί να βρει ένα τόπο ενδιαφέροντος, ένα προϊόν ή μια υπηρεσία. Τυπικά σε ένα σημείο ενδιαφέροντος έχει αποδοθεί ένα όνομα και συνήθως ανήκει σε κάποια κατηγορία[3]. Παραδείγματα σημείων ενδιαφέροντος που μπορεί πρωτίστως να ενδιαφέρουν τους σκοπούς του έργου είναι η τοποθεσία πάρκων, σχολείων, θέσεων στάθμευσης, υπηρεσιών υγείας κ.ά.

Crawler: Στο παρόν έγγραφο ο όρος crawler χρησιμοποιείται για την περιγραφή μονάδων (modules) τα οποία χρησιμοποιούνται για την συλλογή δεδομένων από διάφορες πηγές (υπηρεσίες) στο διαδίκτυο. Για τους σκοπούς του έργου, μας ενδιαφέρουν οι crawlers με τους οποίους γίνεται η συλλογή σημείων ενδιαφέροντος.

Campaign: Ο συγκεκριμένος όρος περιγράφει μία αυτοτελή συλλογή δεδομένων για την οποία έχουν χρησιμοποιηθεί συγκεκριμένες παράμετροι. Οι παράμετροι συλλογής παραμένουν αναλλοίωτες για όλη την διάρκεια ζωής του campaign. Οι παράμετροι αυτές αφορούν κυρίως τους crawlers που χρησιμοποιήθηκαν, τις κατηγορίες των σημείων ενδιαφέροντος, την τοποθεσία συλλογής και το χρονικό διάστημα της συλλογήςτων σημείων ενδιαφέροντος.

Crawl: Ο συγκεκριμένος όρος περιγράφει μια συλλογή δεδομένων που πραγματοποιείται μέσα από ένα συγκεκριμένο campaign και η οποία ξεκινάει σε προκαθορισμένη χρονική στιγμή και αφορά μια συγκεκριμένη πηγή δεδομένων. Στην πραγματικότητα όλα τα δεδομένα που συλλέγονται σε ένακαμπαιγν πραγματοποιούνται μέσα από ένα ή περισσότερα crawls. Επομένως όλα τακαμπαιγνς έχουν τουλάχιστον ένα crawl. Η χρήση των crawls επιβάλλεται λόγω της διαφορετικής συχνότητας συλλογής που έχουν τα διάφορα δεδομένα που θέλουμε να συλλέξουμε.

[3] Ο ορισμός που δίνει το W3C υπάρχει στο http://www.w3.org/2010/POI/documents/Core/core-20111216.html

1.3 APIs πηγών σημείων ενδιαφέροντος

Σε αυτή την ενότητα θα περιγράψουμε τα APIs των πηγών σημείων ενδιαφέροντος που χρησιμοποιούνται από τον AreaProfiler, δηλαδή τα APIs του GooglePlaces, του Foursquarekαι του Twitter. Καθώς οι δυνατότητες κάθε API είναι πολυπληθείς, περιγράφεται μόνο η λειτουργικότητα που έχει άμεση εφαρμογή στον AreaProfiler. Για κάθε πηγή μας ενδιαφέρουν δύο μέθοδοι.Στην περίπτωση της συλλογής σημείων ενδιαφέροντος, η (γενική) μέθοδος απαιτεί τον προσδιορισμό μίας γεωγραφικής περιοχής και μιας κατηγορίας σημείων ενδιαφέροντος και επιστρέφει μία λίστα από σημεία ενδιαφέροντος που περιέχονται στην περιοχή και ανήκουν στησυγκεκριμένη κατηγορία. Η δεύτερη (ειδική) μέθοδος απαιτεί τον προσδιορισμό, μέσω ενός μοναδικού αναγνωριστικού, ενός σημείου ενδιαφέροντος και επιστρέφει επιπλέον πληροφορίες για το σημείο ενδιαφέροντος από αυτές που επιστρέφει η γενική μέθοδος. Σε κάθε περίπτωση, ηγενική μέθοδος επιστρέφει το μοναδικό αναγνωριστικό κάθε σημείου ενδιαφέροντος που περιέχεται στη λίστα της απάντησης. Το αναγνωριστικό αυτό θα μπορεί να χρησιμοποιηθεί από τη δεύτερη (ειδική)μέθοδο για τη λήψη περισσότερων πληροφοριών για το συγκεκριμένο σημείο. Το τμήμα των πληροφοριών που λαμβάνονται από τη δεύτερη μέθοδο αποθηκεύονται στη αντίστοιχη βάση δεδομένων του AreaProfiler. Στην περίπτωση της συλλογής γεωχωρικών tweets, η μέθοδος απαιτεί τον προσδιορισμό μίας γεωγραφικής περιοχής και επιστρέφει μία λίστα από γεωχωρικά προσδιορισμένα tweets, των οποίων η δημοσίευση εμπίπτει στην περιοχή. Το τμήμα των πληροφοριών που λαμβάνονται από το TwitterAPI αποθηκεύονται σε μορφή json σε directory structure.

1.3.1 Google Places

Το API του GooglePlaces που έχει εφαρμογή στον AreaProfiler αποτελείται από τις μεθόδους "PlaceSearch - RadarSearch" (γενική) και "PlaceDetails"(ειδική). Οι συγκεκριμένες μέθοδοι καλύπτουν τη λειτουργικότητα που αναφέρθηκε παραπάνω και αναλύονται παρακάτω.

1.3.1.1 Place Search - Radar Search

Με τη μέθοδο RadarSearch είναι δυνατή η εύρεση έως και 200 σημείων ενδιαφέροντος του GooglePlaces με ένα αίτημα. Χρησιμοποιείται για τον προσδιορισμό σημείων ενδιαφέροντος που ανήκουν σε κάποια κατηγορία και βρίσκονται μέσα σε μία γεωγραφική περιοχή.

Ένα αίτημα RadarSearch περιέχει τις εξής παραμέτρους:

- key – Το κλειδί της εφαρμογής, το οποίο παρέχεται από το Google και ταυτοποιεί την εφαρμογή για λογαριασμό της οποίας γίνεται το αίτημα.
- location – Το σημείο (γεωγραφικό πλάτος, γεωγραφικό μήκος) γύρω από το οποίο θα γίνει η αναζήτηση.
- radius – Η ακτίνα σε μέτρα μέσα στην οποία θα βρίσκονται τα σημεία ενδιαφέροντος που θα επιστραφούν. Η μέγιστη ακτίνα είναι 50.000 μέτρα.
- type – Ο τύπος των σημείων ενδιαφέροντος που θα αναζητηθούν και θα επιστραφούν. Οι τύποι αναπαριστώνται από συγκεκριμένες λέξεις κλειδιά του Google Places, οι οποίες ορίζουν τα place types[4].

[4] https://developers.google.com/places/supported_types

Ένα ενδεικτικό αίτημα για την αναζήτηση καφετεριών, ή, διαφορετικά, στην ορολογία των placetypes, σημείων ενδιαφέροντος τύπου "cafe", σε μία ακτίνα 5.000 μέτρων γύρω από ένα σημείο είναι το εξής:

https://maps.googleapis.com/maps/api/place/radarsearch/json?location=48.859294,2.347589&radius=5000&type=cafe&key=*YOUR_API_KEY*

Η μορφή της απόκρισης φαίνεται στην παρακάτω εικόνα.


```
    "place_id": "ChIJyWEHuEmuEmsRm9hTkapTCrk",
    "reference": "CoQBdQAAAFSiijw5-cAV68xdf2O18pKIZ0seJh03u9h9wk_lEdG-cP1dWvp_QGS4SNCBMk_fB06YRsfMrNkINtPez22p5lRIlj5ty_HmcNw
}, {
    "geometry": {
        "location": {
            "lat": -33.866891,
            "lng": 151.200814
        }
    },
    "place_id": "ChIJqwS6fjiuEmsRJAMiOY9MSms",
    "reference": "CpQBhgAAAFN27qR_t5oSDKPUzjQIeQa3lrRpFTm5alW3ZYbMFm8k10ETbISfK9S1nwcJVfrP-bjra7NSPuhaRulxoonSPQklDyB-xGvcJnc
}, {
    "geometry": {
        "location": {
            "lat": -33.870943,
            "lng": 151.190311
        }
    },
    "place_id": "ChIJLfySpTOuEmsRsc_JfJtljdc",
    "reference": "CoQBdQAAANQSThnTekt-UokiTiX3oUFT6YDfdQJIG0ljlQnkLfWefcKmjxax0xmUpWjmpWdOsScl9zSyBNImmrTO9AE9DnWTdQ2hY7n-OOL
}, {
```


Εικόνα1. Google Places API radar response

Η απόκριση περιέχει μία λίστα από σημεία ενδιαφέροντος. Κάθε εγγραφή σημείου ενδιαφέροντος περιέχει τα εξής απαραίτητα πεδία για τον AreaProfiler:

- geometry – Το γεωγραφικό σημείο (γεωγραφικό πλάτος, γεωγραφικό μήκος) στο οποίο τοποθετείται χωρικά το σημείο ενδιαφέροντος.
- place_id – Το αναγνωριστικό του σημείου ενδιαφέροντος, απαραίτητο για την ταυτοποίηση του σημείου ενδιαφέροντος. Χρησιμοποιείται από την επόμενη (ειδική) μέθοδο για τη λήψη επιπλέον στοιχείων για το συγκεκριμένο σημείο.

1.3.1.2 Place Details

Έχοντας ένα place_id από την προηγούμενη μέθοδο, είναι η δυνατή η λήψη περισσότερων πληροφοριών για ένα συγκεκριμένο σημείο ενδιαφέροντος με τη χρήση της μεθόδου PlaceDetails (ειδική).

Ένα αίτημα PlaceDetails περιέχει τις εξής παραμέτρους:

- key – Το κλειδί της εφαρμογής, το οποίο παρέχεται από το Google και ταυτοποιεί την εφαρμογή για λογαριασμό της οποίας γίνεται το αίτημα.
- place_id – Το αναγνωριστικό του σημείου ενδιαφέροντος για το οποίο θα ζητηθούν περισσότερες πληροφορίες.

Ένα ενδεικτικό αίτημα για πληροφορίες για το σημείο ενδιαφέροντος με αναγνωριστικό το "ChIJN1t_tDeuEmsRUsoyG83frY4" είναι το εξής:

https://maps.googleapis.com/maps/api/place/details/json?placeid=ChIJN1t_tDeuEmsRUsoyG83frY4&key=YOUR_API_KEY

Η μορφή της απόκρισης φαίνεται στην παρακάτω εικόνα.

```
{
   "html_attributions": [],
   "result": {
      "address_components": [{
         "long_name": "48",
         "short_name": "48",
         "types": ["street_number"]
      }, {
         "long_name": "Pirrama Road",
         "short_name": "Pirrama Road",
         "types": ["route"]
      }, {
         "long_name": "Pyrmont",
         "short_name": "Pyrmont",
         "types": ["locality", "political"]
      }, {
         "long_name": "NSW",
         "short_name": "NSW",
         "types": ["administrative_area_level_1", "political"]
      }, {
         "long_name": "AU",
         "short_name": "AU",
         "types": ["country", "political"]
      }, {
         "long_name": "2009",
         "short_name": "2009",
         "types": ["postal_code"]
      }],
      "formatted_address": "48 Pirrama Road, Pyrmont NSW, Australia",
      "formatted_phone_number": "(02) 9374 4000",
      "geometry": {
         "location": {
            "lat": -33.8669710,
            "lng": 151.1958750
         }
      },
      "icon": "http://maps.gstatic.com/mapfiles/place_api/icons/generic_business-71.png",
      "id": "4f89212bf76dde31f092cfc14d7506555d85b5c7",
      "international_phone_number": "+61 2 9374 4000",
      "name": "Google Sydney",
      "place_id": "ChIJN1t_tDeuEmsRUsoyG83frY4",
      "scope": "GOOGLE",
      "alt_ids": [{
```


Εικόνα2. Google Places API place details response

Η απόκριση περιέχει αναλυτικές πληροφορίες για το σημείο ενδιαφέροντος. Συγκεκριμένα περιέχει τα εξής απαραίτητα πεδία για τον AreaProfiler:

- formatted_address – Η διεύθυνση του σημείου ενδιαφέροντος.
- formatted_phone_number – Το τηλέφωνο του σημείου ενδιαφέροντος.
- geometry – Το γεωγραφικό σημείο (γεωγραφικό πλάτος, γεωγραφικό μήκος) στο οποίο τοποθετείται χωρικά το σημείο ενδιαφέροντος.
- name – Το όνομα του σημείου ενδιαφέροντος.
- opening_hours – Οι ώρες λειτουργίας του σημείου ενδιαφέροντος.
- place_id – Το αναγνωριστικό του σημείου ενδιαφέροντος.
- rating – Η βαθμολογία του σημείου ενδιαφέροντος βάσει των κριτικών για αυτό.
- types – Οι τύποι, place types, που χαρακτηρίζουν το σημείο ενδιαφέροντος.
- url – Η ηλεκτρονική διεύθυνση του σημείου ενδιαφέροντος.

1.3.2 Foursquare

Το API του Foursquare που έχει εφαρμογή στον AreaProfiler αποτελείται από τις μεθόδους "SearchVenues" (γενική) και "VenueDetails" (ειδική). Οι συγκεκριμένες μέθοδοι αναλύονται παρακάτω.

1.3.2.1 Search Venues

Με τη μέθοδο SearchVenues είναι δυνατή η εύρεση έως και 50 σημείων ενδιαφέροντος του Foursquare με ένα αίτημα. Χρησιμοποιείται για τον προσδιορισμό σημείων ενδιαφέροντος που ανήκουν σε κάποια κατηγορία και περιέχονται μέσα σε μία γεωγραφική περιοχή.

Ένα αίτημα SearchVenues περιέχει τις εξής παραμέτρους:

- oauth_token – Το κλειδί της εφαρμογής, το οποίο παρέχεται από το Foursquare και ταυτοποιεί την εφαρμογή για λογαριασμό της οποίας γίνεται το αίτημα.
- ll – Το σημείο (γεωγραφικό πλάτος, γεωγραφικό μήκος) γύρω από το οποίο θα γίνει η αναζήτηση.
- intent – Ο σκοπός της αναζήτησης. Για την αναζήτηση σημείων ενδιαφέροντος μέσα σε μία περιοχή ενδιαφέροντος χρησιμοποιούμε την τιμή "browse".
- radius – Η ακτίνα σε μέτρα μέσα στην οποία θα βρίσκονται τα σημεία ενδιαφέροντος που θα επιστραφούν. Η μέγιστη ακτίνα είναι 100.000 μέτρα.
- categoryId – Η κατηγορία (ή οι κατηγορίες, χωρισμένες με κόμμα) των σημείων ενδιαφέροντος που θα αναζητηθούν και θα επιστραφούν. Οι κατηγορίες αναπαριστώνται από τα αναγνωριστικά τους στην ιεραρχία κατηγοριών του Foursquare[5].

Ένα ενδεικτικό αίτημα για την αναζήτηση μουσείων σε μία ακτίνα 10000 μέτρων γύρω από ένα σημείο είναι το εξής:

https://api.foursquare.com/v2/venues/search?ll=40.7,74&intent=browse&radius=10000&categoryId=4bf58dd8d48988d181941735&oauth_token=YOUR_OAUTH_TOKEN

Η μορφή της απόκρισης φαίνεται στην παρακάτω εικόνα.

```
],
response: {
   venues: [
      {
         id: "427c0500f964a52097211fe3",
         name: "The Metropolitan Museum of Art",
         contact: {
            phone: "+121253577710",
            formattedPhone: "+1 212-535-7710",
            twitter: "metmuseum",
            facebook: "6296252634",
            facebookUsername: "metmuseum",
            facebookName: "The Metropolitan Museum of Art, New York"
         },
         location: {
            address: "1000 5th Ave",
            crossStreet: "btwn E 80th & E 84th St",
            lat: 40.778936659294864,
            lng: -73.96229820007625,
            distance: 9344,
            postalCode: "10028",
            mayNotNeedAddress: false,
            cc: "US",
            city: "New York",
            state: "NY",
            country: "United States",
            formattedAddress: [
               "1000 5th Ave (btwn E 80th & E 84th St)",
               "New York, NY 10028",
               "United States"
```


Εικόνα3. FoursquareAPIsearch-venuesresponse

[5] https://developer.foursquare.com/categorytree

Η απόκριση περιέχει μία λίστα από σημεία ενδιαφέροντος. Κάθε εγγραφή σημείου ενδιαφέροντος περιέχει τα εξής απαραίτητα πεδία για τον AreaProfiler:

- location – Το γεωγραφικό σημείο (γεωγραφικό πλάτος, γεωγραφικό μήκος) στο οποίο τοποθετείται χωρικά το σημείο ενδιαφέροντος.
- id – Το αναγνωριστικό του σημείου ενδιαφέροντος, απαραίτητο για την ταυτοποίηση του σημείου ενδιαφέροντος. Χρησιμοποιείται από την επόμενη (ειδική)μέθοδο για τη λήψη περισσότερων πληροφοριών για το συγκεκριμένο σημείο ενδιαφέροντος.

1.3.2.2 Venue Details

Έχοντας ένα id από την προηγούμενη γενική μέθοδο είναι δυνατή η λήψη περισσότερων πληροφοριών για ένα συγκεκριμένο σημείο ενδιαφέροντος του Foursquare με τη χρήση της μεθόδου VenueDetails (ειδική).

Ένα αίτημα VenueDetails περιέχει τις εξής παραμέτρους:

- oauth_token – Το κλειδί της εφαρμογής, το οποίο παρέχεται από το Foursquare και ταυτοποιεί την εφαρμογή για λογαριασμό της οποίας γίνεται το αίτημα.
- id – Το αναγνωριστικό του σημείου ενδιαφέροντος του Foursquare για το οποίο θα ζητηθούν περισσότερες πληροφορίες.

Ένα ενδεικτικό αίτημα για λήψη λεπτομερειών για το σημείο ενδιαφέροντος με id το "427c0500f964a52097211fe3" είναι το εξής:

https://api.foursquare.com/v2/venues/427c0500f964a52097211fe3?oauth_token=YOUR_OAUTH_TOKEN

Η μορφή της απόκρισης φαίνεται στην παρακάτω εικόνα.


```
    }
],
response: {
    venue: {
        id: "427c0500f964a52097211fe3",
        name: "The Metropolitan Museum of Art",
        contact: {
            phone: "+12125357710",
            formattedPhone: "+1 212-535-7710",
            twitter: "metmuseum",
            facebook: "6296252634",
            facebookUsername: "metmuseum",
            facebookName: "The Metropolitan Museum of Art, New York"
        },
        location: {
            address: "1000 5th Ave",
            crossStreet: "btwn E 80th & E 84th St",
            lat: 40.778936659294864,
            lng: -73.96229820007625,
            postalCode: "10028",
            mayNotNeedAddress: false,
            cc: "US",
            city: "New York",
            state: "NY",
            country: "United States",
            formattedAddress: [
                "1000 5th Ave (btwn E 80th & E 84th St)",
                "New York, NY 10028",
                "United States"
            ]
        },
        canonicalUrl: "https://foursquare.com/v/the-metropolitan-museum-of-art
```


Εικόνα4. Foursquare API venue-details response

Η απόκριση περιλαμβάνει αναλυτικές πληροφορίες για το σημείο ενδιαφέροντος. Συγκεκριμένα περιέχει τα εξής απαραίτητα πεδία για τον AreaProfiler:

- location – Η γεωγραφική θέση (γεωγραφικό πλάτος, γεωγραφικό μήκος) στο οποίο τοποθετείται χωρικά το σημείο ενδιαφέροντος, καθώς και η διεύθυνση του σημείου ενδιαφέροντος.
- contact – Πληροφορίες επικοινωνίας, που περιλαμβάνουν το τηλέφωνο του σημείου ενδιαφέροντος.
- name – Το όνομα του σημείου ενδιαφέροντος.
- hours – Οι ώρες λειτουργίας του σημείου ενδιαφέροντος.
- id – Το αναγνωριστικό του σημείου ενδιαφέροντος.
- rating – Η βαθμολογία του σημείου ενδιαφέροντος βάσει των κριτικών χρηστών.
- categories – Οι κατηγορίες της ιεραρχίας κατηγοριών του Foursquare στις οποίες ανήκει το σημείο ενδιαφέροντος.
- tags – Οι ετικέτες που χαρακτηρίζουν το σημείο ενδιαφέροντος.
- url – Η ηλεκτρονική διεύθυνση του σημείου ενδιαφέροντος.

1.3.3 Twitter

Το API του Twitter που χρησιμοποιήθηκε είναι toStreaming API(PublicStreamsAPI), το οποίο εφόσον συνδεθεί στο endpoint (POST statuses/filter) προσφέρει δείγμα των δημόσιων δεδομένων, χωρίς τους περιορισμούς χρόνου και όγκου δεδομένων του RESTAPI. Το API που χρησιμοποιήθηκε επιστρέφει τα δημόσια tweets που ταιριάζουν σε ένα ή περισσότερα κριτήρια φιλτραρίσματος, από τα οποία εμείς χρησιμοποιούμε μόνο αυτό της τοποθεσίας.Το StreamingAPI του Twitter που έχει εφαρμογή στον Area Profiler αποτελείται από τη μέθοδο "Twitter Stream Crawler" που παρουσιάζεται αναλυτικά στην ενότητα 1.8. ToStreaming API του Twitter λειτουργεί σε πραγματικό χρόνο, δηλαδή συλλέγει τα tweets που δημοσιεύονται δυναμικά και όχι στατικά. Στα πλαίσια του AreaProfiler αυτό επιτυγχάνεται με τη χρήση του Streaming API σε κυλιόμενο παράθυρο χρόνου. Η μέθοδος που περιγράφεται συλλέγει γεωχωρικά tweets, βασισμένη στον προσδιορισμό μίας γεωγραφικής περιοχής και επιστρέφει μία λίστα από γεωχωρικά προσδιορισμένα tweets και check-ins των οποίων η δημοσίευση εμπίπτει στην περιοχή. Το τμήμα των πληροφοριών που λαμβάνονται από το Streaming API αποθηκεύονται σε directorystructuresε μορφή json.

1.4 Απαιτήσεις και ζητούμενα

Ο AreaProfiler υποστηρίζει την αποθήκευση διαφορετικών μορφών δεδομένων από τις πηγές σημείων ενδιαφέροντος, ικανοποιώντας ένα σύνολο λειτουργικών απαιτήσεων, που συνοπτικά περιγράφονται παρακάτω.

1.4.1 Παράλληλα Campaigns

Μια βασική απαίτηση είναι το σύστημα να μπορεί να υποστηρίξει την διενέργεια πολλαπλών campaigns ταυτόχρονα, καθώς κάθε campaign ενδέχεται να έχει διάρκεια εβδομάδων ή μηνών.

1.4.2 Διαχείριση Διαθέσιμων Πόρων

Το σύστημα χρησιμοποιεί πολλαπλούς, ανεξάρτητους crawlers ειδικού σκοπού με τους οποίους γίνεται η συλλογή των δεδομένων. Καθώς όμως κάθε πηγή θέτει τα δικά της όρια

σχετικά με τον αριθμό των αιτήσεων που μπορούν να πραγματοποιηθούν σε καθορισμένο χρονικό διάστημα, έπρεπε να αναπτύξουμε έναν μηχανισμό που να εγγυάται την τήρηση των υφιστάμενων περιορισμών. Καθώς οι περιορισμοί αυτοί ενδέχεται να αλλάζουν, διαβάζονται από εξωτερικά αρχεία ως παράμετροι για κάθε crawler. Επίσης, καθώς το σύστημα επιτρέπει την παράλληλη συλλογή, κάποιος crawler μπορεί να χρησιμοποιείται ταυτόχρονα από περισσότερα από έναcampaign. Αποτέλεσμα της παράλληλης συλλογής ήταν η ανάγκη για ανάπτυξη ενός μηχανισμού με τον οποίο να πραγματοποιείται η δίκαιη κατανομή των διαθέσιμων πόρων κάθε crawleraπό όλα τα campaigns που τον χρησιμοποιούν.

1.4.3 Καθολική ΜορφήΔεδομένων

Κάθε πηγή σημείων ενδιαφέροντος χρησιμοποιεί διαφορετική μορφή για την αναπαράσταση των σημείων ενδιαφέροντος. Μία από τις απαιτήσεις που έχουμε θέσει είναι η δυνατότητα προσθήκης crawlers για νέες πηγές δεδομένων. Η αποθήκευση όμως των συγκεκριμένων δεδομένων θα απαιτούσε και την ανάγκη για δημιουργία ενός νέου μοντέλου για την αποθήκευση τους. Επίσης, οι αλλαγές στην μορφή δεδομένων που χρησιμοποιούν οι πηγές θα απαιτούσε ενημέρωση εκτός από τον crawler και του μοντέλου για την αποθήκευση τους. Για να αποφύγουμετο συγκεκριμένο περιορισμό, αποφασίσαμε να χρησιμοποιήσουμε ενιαία μορφή για την αποθήκευση των δεδομένων. Το όφελος από την παρούσα απαίτηση είναι ότι πλέον χρειαζόμαστε να υλοποιήσουμε ένα μοντέλο αποθήκευσης το οποίο είναι ανεξάρτητο από τις πηγές και οποιαδήποτε μεταβολή στην μορφή των δεδομένων τους. Όμως με αυτόν τον τρόπο κάθε crawler εκτός από την συλλογή των δεδομένων απαιτείται να διαθέτει ένα μηχανισμό για την μετατροπή των δεδομένων που συλλέγει στην καθολική μορφή. Εκτός, όμως, από τα δεδομένα που μετατρέπονται ώστε να υπακούν στην καθολική μορφή, αποθηκεύονται και συγκεκριμένα δεδομένα ιδιαίτερα για κάθε τύπο crawler, τα οποία είναι αναγκαία για μελλοντική ανάλυση των αποθηκευμένων σημείων ενδιαφέροντος.

1.4.4 Παραμετροποίηση campaigns

Εξίσου επιτακτική απαίτηση είναι ο χρήστης να μπορεί να παραμετροποιήσει πλήρως τα campaigns που ορίζει. Οι παράμετροι που θα μπορεί να ορίσει είναι οι ακόλουθες:

- Όνομα και περιγραφή
- Διάρκεια εκτέλεσης
- Επιλογή crawlers
- Τοποθεσία συλλογής
- Επιλογή κατηγοριών
- Επιλογή συχνότητας συλλογής

1.4.4.1 Όνομα και περιγραφή

Ο χρήστης μπορεί να ορίσει το όνομα και την περιγραφή που επιθυμεί για τοcampaign που ορίζει. Η επιλογή του ονόματος και η περιγραφή δεν υπόκειται σε κανένα περιορισμό και επίσης περισσότερα από έναcampaigns δύναται να έχουν ίδιο όνομα και περιγραφή.

1.4.4.2 Διάρκεια εκτέλεσης

Ο χρήστης είναι σε θέση να ορίσει για πόσο χρονικό διάστημα θα εκτελεστεί ένα campaign. Ο χρήστης ορίζει την διάρκεια ενός campaign με το να θέσει την ημερομηνία τερματισμού ενός campaign, καθώς η έναρξη του campaign είναι πάντα η στιγμή που ορίζεται. Ο μοναδικός περιορισμός που υφίσταται είναι ότι η ημερομηνία τερματισμού πρέπει να αφορά μια μελλοντική ημερομηνία.

1.4.4.3 Τοποθεσία συλλογής

Για τιςcampaigns που πραγματοποιούν συλλογή σημείων ενδιαφέροντος ο ορισμός της τοποθεσίας είναι υποχρεωτικός. Μία ακόμα απαίτηση που θέσαμε είναι το σύστημα να εξασφαλίζει την συλλογή όλων των διαθέσιμων σημείων ενδιαφέροντος στην περιοχή συλλογής. Καθώς οι πηγές θέτουν όρια στον μέγιστο αριθμό των σημείων ενδιαφέροντος που επιστρέφουν θα πρέπει να ληφθεί ιδιαίτερη μνεία για την συγκεκριμένη απαίτηση.

1.4.4.4 Επιλογή crawlers

Ο χρήστης πρέπει να έχει την επιλογή να επιλέξει τους crawlers για την συλλογή σημείων ενδιαφέροντος που θα χρησιμοποιήσει στα campaigns που ορίζει.

1.4.4.5 Επιλογή κατηγοριών

Ο χρήστης πρέπει να μπορεί να επιλέξει τις κατηγορίες των σημείων ενδιαφέροντος που τον ενδιαφέρουν και επίσης να ορίσει τις λέξεις κλειδιά με τις οποίες θέλει να συλλεχθούν οι κατηγορίες που ορίζει στο campaign. Η πρώτη απαίτηση δίνει στον χρήστη μια ελευθερία ώστε να επιλέξει τις κατηγορίες που τον ενδιαφέρουν ώστε να αξιοποιήσει με τον καλύτερο δυνατό τρόπο τους πόρους που έχει στην διάθεση του. Η δεύτερη απαίτηση έχει ως σκοπό να μπορεί ο χρήστης να συλλέξει αποδοτικά τις κατηγορίες που τον ενδιαφέρουν σε οποιαδήποτε περιοχή ανεξαρτήτως των γλωσσικών ή πολιτισμικών υπόβαθρων. Για παράδειγμα η συλλογή των σημείων ενδιαφέροντος «Βενζινάδικο» στην Ελλάδα θα πρέπει να πραγματοποιηθεί με την λέξη «βενζινάδικο» ενώ στις ΗΠΑ που χρησιμοποιείται διαφορετική γλώσσα ομιλίας θα ήταν προτιμότερο να χρησιμοποιήσουμε τους όρους "fillingstation" και "gasstation". Αντίθετα στο Ηνωμένο Βασίλειο, παρά την χρήση της αγγλικής γλώσσας όπως και στις ΗΠΑ, λόγω διαφορετικού πολιτισμικού υπόβαθρου θα ήταν προτιμότερο να χρησιμοποιήσουμε τον όρο "patrolstation"[6]. Παρέχεται, επίσης, η δυνατότητα στο χρήστη να σαρώσει μία περιοχή για όλα ανεξαιρέτως τα σημεία ενδιαφέροντος που παρέχονται από τους υποστηριζόμενους crawlers.

1.4.4.6 Συχνότητα συλλογής

Μία από τις απαιτήσεις ήταν η δυνατότητα παρακολούθησης της εξέλιξης των δεδομένων που συλλέγονται στην πάροδο του χρόνου. Συνεπώς, το σύστημα πρέπει να είναι σε θέση να μπορεί να πραγματοποιεί campaigns με μεγάλο χρονικό ορίζοντα, ακόμα και μηνών. Επιπλέον κατά την διάρκεια ενόςcampaign, θα πρέπει να μπορούν να πραγματοποιηθούν πολλαπλές συλλογές χρησιμοποιώντας τα ίδια κριτήρια. Αυτό επιτρέπει να γίνει σύγκριση των δεδομένων που συλλέχτηκαν σε διαφορετικές χρονικές στιγμές. Επιπλέον, θα πρέπει να ορίζεται η συχνότητα συλλογής δεδομένων για κάθε μία κατηγορία σημείων ενδιαφέροντος μεμονωμένα. Ο λόγος για τη συγκεκριμένη απαίτηση είναι ότι η συχνότητα

[6] http://en.wikipedia.org/wiki/Filling_station

αλλαγής χρήσης ενός σημείου ενδιαφέροντος έχει σχέση με την κατηγορία του. Για παράδειγμα αναμένουμε η αλλαγή χρήσης ενός νοσοκομείου να είναι πολύ πιο σπάνια σε σχέση με τα σημεία ενδιαφέροντος της κατηγορίας μπαρ. Η συχνότητα συλλογής με βάση τον crawler απορρίφθηκε, καθώς στη γενική περίπτωση όλες οι πηγές περιέχουν όλες τις κατηγορίες σημείων ενδιαφέροντος.

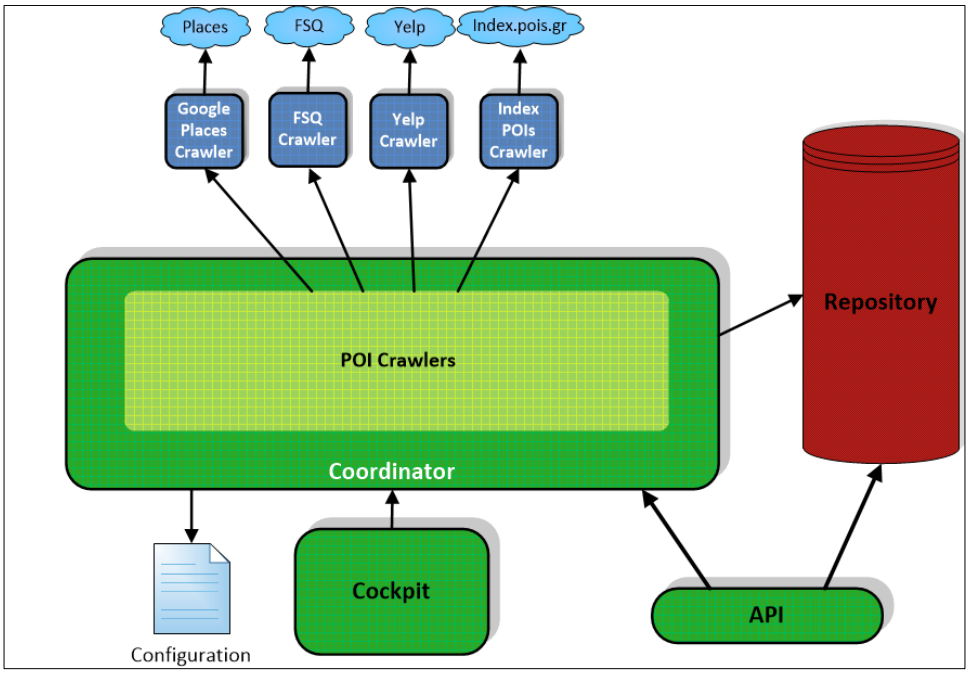
1.5 Σχεδιασμός του Area Profiler

Η βασική λειτουργία που πρέπει να επιτελεί το Επίπεδο Συλλογής (CrawlingLevel) είναι να προσφέρει μια διεπαφή (API) με την οποία (α) μπορούν να οριστούν νέα campaigns με συγκεκριμένες παραμέτρους και (β) μπορούν να ανακτηθούν τα δεδομένα που έχουν ήδη συλλεχθεί. Η δευτερεύουσα λειτουργία αφορά την παροχή ενός γραφικού περιβάλλον χρήστη (GUI) με την οποία μπορεί να οριστεί ένα νέο campaign με συγκεκριμένες παραμέτρους. Η βασική λειτουργία χρησιμοποιείται από το Επίπεδο Ολοκλήρωσης (*IntegrationLevel)* ενώ η βοηθητική από κάποιον χρήστη.

1.5.1 Αρχιτεκτονική του Επιπέδου Συλλογής

Η αρχιτεκτονική του υποσυστήματος συλλογής αποτελείται από τρία βασικά τμήματα. Το πρώτο τμήμα περιλαμβάνει τους crawlers και μια μονάδα υπεύθυνη για τον συντονισμό τους (coordinator), το δεύτερο τμήμα την βάση δεδομένων που αποθηκεύονται όλα τα δεδομένα που συλλέγονται μέσω των crawlers και το τρίτο τμήμα αφορά την διεπαφή (API) που παρέχει την δυνατότητα διαχείρισης των crawlers και την δυνατότητα ανάκτησης των συλλεγμένων δεδομένων.

Η αρχιτεκτονική του υποσυστήματος συλλογής παρουσιάζεται στο παρακάτω σχήμα:



Εικόνα 5. Αρχιτεκτονική Υποσυστήματος Συλλογής

Crawlers: Το σύστημα περιλαμβάνει ένα μεγάλο πλήθος μονάδων με crawlers για τη συλλογή των σημείων ενδιαφέροντος. Καθώς οι μονάδες αυτές είναι ανεξάρτητες, η δημιουργία και η ενημέρωση των μονάδων είναι εύκολη διαδικασία, καθώς δεν επηρεάζει κανένα άλλο τμήμα του υποσυστήματος.

Coordinator: Η συγκεκριμένη μονάδα είναι υπεύθυνη για τον συντονισμό των crawlers και την διαχείριση των campaigns. Μία ακόμα λειτουργία που επιτελεί είναι η διάσπαση των περιοχών συλλογής δεδομένων σε μικρότερες περιοχές, ώστε να εξασφαλίζει την καθολική συλλογή των σημείων ενδιαφέροντος. Επίσης η συγκεκριμένη μονάδα διαχειρίζεται τους πόρους για τις πηγές, ώστε να θέτει όρια στον αριθμό των αιτημάτων καινα διαχειρίζεται τον τρόπο με τον οποίο γίνεται η αναζήτηση των κατηγοριών για κάθε crawler. Αναλυτικότερα, η συγκεκριμένη μονάδα είναι υπεύθυνη για την εγκυρότητα των παραμέτρων των campaigns και την εκτέλεση των campaigns. Αν οι παράμετροι είναι μη έγκυρες, τότε αποτρέπει την εκκίνηση των campaigns και επιστρέφει το κατάλληλο μήνυμα λάθους. Αντίθετα αν οι παράμετροι είναι έγκυρες τότε η μονάδα αναλαμβάνει την παράλληλη εκτέλεση των campaigns, έτσι ώστε κάθε campaign να εκτελείται σε δικό του ανεξάρτητο νήμα (thread).

Επίσης ο coordinator αναλαμβάνει για κάθε campaign, σύμφωνα με τις παραμέτρους του, να συντονίσει τους κατάλληλους crawlers και να θέσει τις παραμέτρους συλλογής όπως την περιοχή, την κατηγορία δεδομένων κ.α.

Μία ακόμα λειτουργία που επιτελεί ο coordinatoreίναι η διάσπαση των περιοχών συλλογής δεδομένων. Η λειτουργία αυτή είναι επιβεβλημένη καθώς οι πηγές δεδομένων θέτουν όρια στον μέγιστο αριθμό των σημείων ενδιαφέροντος που επιστρέφουν. Λόγω της απαίτησης για καθολική συλλογή των σημείων ενδιαφέροντος στις περιοχές που ορίζονται από τον χρήστη, υλοποιήθηκε ένας μηχανισμός ώστε να συλλέγει όλα τα σημεία ενδιαφέροντος. Ο μηχανισμός αυτός λειτουργεί με την διάσπαση των περιοχών συλλογής σε περισσότερα μικρά τμήματα, ώστε να συλλέγεται το σύνολο των σημείων ενδιαφέροντος που περιλαμβάνονται. Αυτό πραγματοποιείται με την αναδρομική διάσπαση της περιοχής συλλογής,ώστε σε κάθε περιοχή επιστρέφεται ο μέγιστος αριθμός των σημείων ενδιαφέροντος που επιτρέπει η εκάστοτε πηγή.

Repository: Η μονάδα αυτή είναι υπεύθυνη για την αποθήκευση και ανάκτηση των δεδομένων που συλλέγονται.

API: Με τη συγκεκριμένημονάδα παρέχεται μια διεπαφή με την οποία πραγματοποιείται η διαχείριση των campaigns και η ανάκτηση των δεδομένων. Η μονάδα αυτή διαχειρίζεται τις αιτήσεις και επιστρέφει τις κατάλληλες αποκρίσεις χρησιμοποιώντας ένα προσυμφωνημένο πρωτόκολλο σε μορφή JSON.

Cockpit: Η συγκεκριμένημονάδα παρέχει ένα ακόμα τρόπο διαχείρισης της μονάδας συντονισμού μέσω ενός γραφικού περιβάλλοντος εργασίας (GUI). Σε αντίθεση με την διεπαφή που προορίζεται για προγραμματιστική χρήση, με τη συγκεκριμένη μονάδα ένας χρήστης, μέσα από ένα φιλικό περιβάλλον, μπορεί να χρησιμοποιήσει τις παρεχόμενες εντολές διαχείρισης.

1.5.2 Αρχειοθέτηση ανακτηθέντων δεδομένων

Για την αρχειοθέτηση των ανακτηθέντων δεδομένων χρησιμοποιείται μία σχεσιακή βάση δεδομένων με σκοπό την ορθή οργάνωση δεδομένων που συλλέγουμε με τους crawlers.

Με την χρήση της βάσης είναι εφικτή η άμεση ανάκτηση της πληροφορίας με διάφορα κριτήρια σύμφωνα με τις απαιτήσεις της εφαρμογής.

1.5.3 Ανάλυση απαιτήσεων αρχειοθέτησης

Οι απαιτήσεις που πρέπει να ικανοποιεί η βάση δεδομένων αφορούν την συγκέντρωση και την οργάνωση πληροφοριών για τα *σημεία ενδιαφέροντος* και τις *μεταπληροφορίες* των campaigns.

Για τα σημεία ενδιαφέροντος απαιτείται η αποθήκευση των χαρακτηριστικών που τα συνοδεύουν, όπως το όνομα, η διεύθυνση, ώρες λειτουργίας κ.α.

Πέρα από τα δεδομένα που συλλέγονται από τις πηγές απαιτείται και η αποθήκευση μεταπληροφοριών για κάθε campaign, όπως πληροφορίες σχετικά με το όνομα, την περιγραφή, την ημερομηνία έναρξης και λήξης, καθώς και τις παραμέτρους συλλογής. Οι παράμετροι αφορούν τους crawlers που χρησιμοποιήθηκαν, την γεωγραφική περιοχή αναζήτησης, πληροφορίες σχετικά με τα crawls, τις κατηγορίες των σημείων ενδιαφέροντος που χρησιμοποιήθηκαν ,μαζί με τους όρους αναζήτησης και τον χρόνο ενημέρωσης κάθε κατηγορίας. Οι παραπάνω απαιτήσεις που θέσαμε σχετικά με την αρχειοθέτηση των παραμέτρων, υπερκαλύπτουν τις ανάγκες της εφαρμογής ώστε να αποτρέψουμε την ανάγκη για επανασχεδιασμό της βάσης, σε περίπτωση που έχουμε μικρές αλλαγές στις απαιτήσεις της εφαρμογής.

1.5.4 Συνοπτική παρουσίαση εντολών

Εκτός όμως από την συγκέντρωση πληροφοριών η βάση θα πρέπει να υποστηρίζει και λειτουργίες, όπως η χρήση εντολών διαχείρισης και σύνθετων εντολών ανάκτησης δεδομένων. Ακολουθεί η παράθεση του συνόλου των εντολών που υποστηρίζονται, χωρισμένες στις δύο παραπάνω κατηγορίες.

Εντολές διαχείρισης

- Εισαγωγή, διαγραφή και τροποποίηση των crawlers για τα σημεία ενδιαφέροντος που υποστηρίζει το σύστημα.
- Εισαγωγή, διαγραφή και τροποποίηση των κατηγοριών για τα σημεία ενδιαφέροντος που υποστηρίζονται από το σύστημα.
- Εισαγωγή και διαγραφή των campaigns και των παραμέτρων που συνοδεύουν κάθε μία μεμονωμένα. Οι παράμετροι είναι οι ακόλουθες:
    - Οι crawlers για τα σημεία ενδιαφέροντος που χρησιμοποιούνται στην συγκεκριμένη campaign
    - Οι κατηγορίες που χρησιμοποιούνται στην συγκεκριμένη campaign
    - Η τοποθεσία αναζήτησης
    - Εισαγωγή και διαγραφή των παραμέτρων συλλογής που αφορούν τους crawlers για τα σημεία ενδιαφέροντος.
    - Εισαγωγή και διαγραφή των παραμέτρων συλλογής που αφορούν την τοποθεσία της συλλογής.
    - Εισαγωγή και διαγραφή των παραμέτρων συλλογής που αφορούν τις κατηγορίες των σημείων ενδιαφέροντος.

- Εισαγωγή και διαγραφή των crawls για κάθε συλλογή.

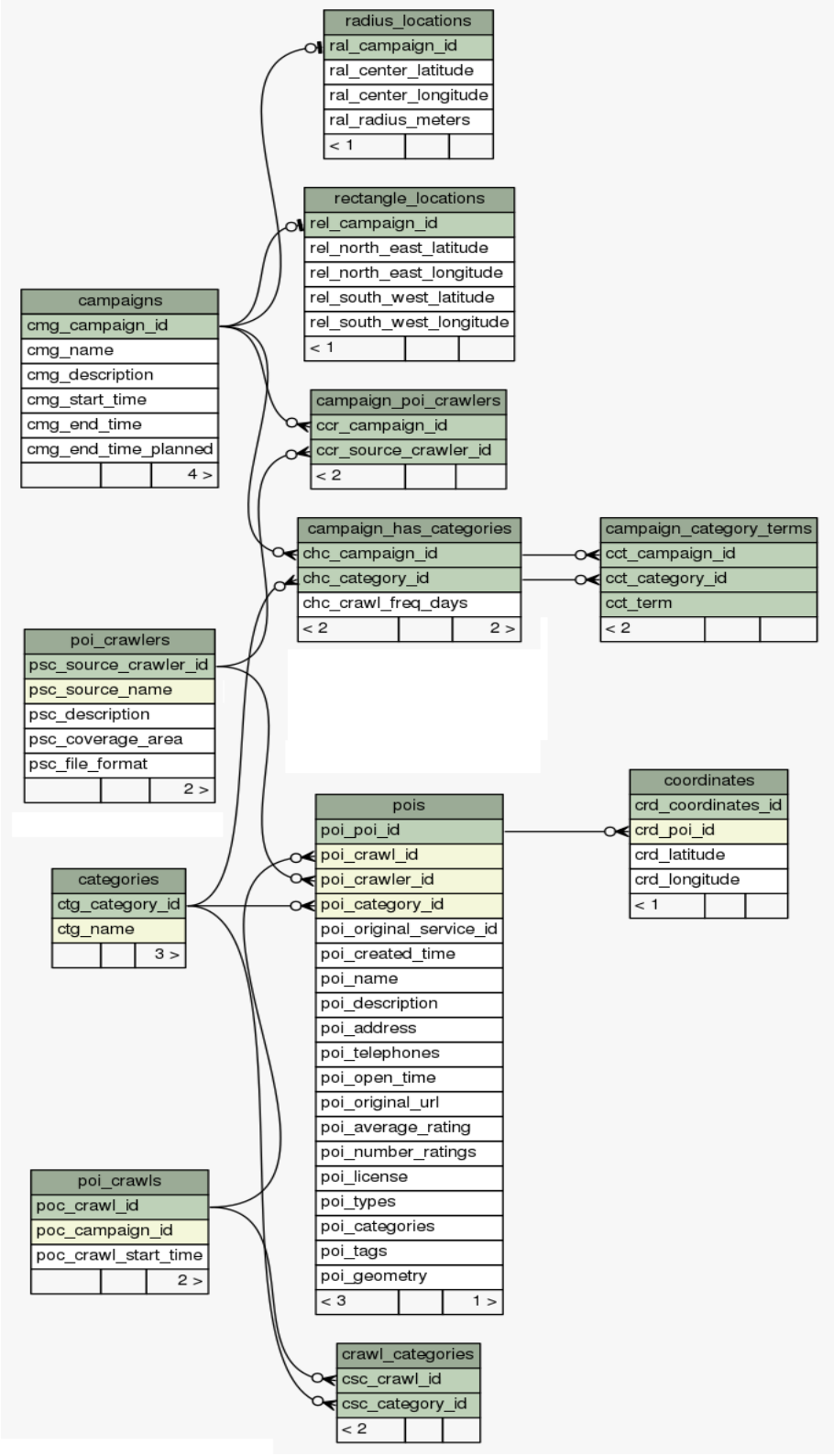
- Εισαγωγή και διαγραφή των σημείων ενδιαφέροντος.

Εντολές ανάκτησης

- Ανάκτηση των σημείων ενδιαφέροντος με χρήση μεμονωμένων ή σε συνδυασμό των ακόλουθων κριτηρίων:
    - μοναδικό αναγνωριστικό του σημείου ενδιαφέροντος
    - μοναδικό αναγνωριστικό της συλλογής στην οποία συλλέχτηκε το σημείο ενδιαφέροντος
    - μοναδικό αναγνωριστικό των κατηγοριών βάσει των οποίων έχει συλλεχθεί το σημείο ενδιαφέροντος
    - μοναδικό αναγνωριστικό του crawler με τον οποίο έγινε η συλλογή του σημείου ενδιαφέροντος

1.5.5 Σχεδιασμός της βάσης δεδομένων

Η βάση δεδομένων που παρουσιάζεται εδώ δεν αποτελεί την τελική βάση του CitySense, αλλα ένα αποθετήριο, στο οποίο ο AreaProfiler αποθηκεύει τα δεδομένα που συλλέγει σε πρώτη φάση. Η βάσηαυτή είναι σχεδιασμένη ώστε να υποστηρίζει το σύνολο των απαιτήσεων που τέθηκαν και επιπλέον να είναι εύκολα προσαρμόσιμη σε μελλοντικές απαιτήσεις και εισαγωγή νέων δεδομένων που ενδεχομένως να προκύψουν. Τοαντίστοιχο ER διάγραμμα της βάσης δεδομένωναπεικονίζεται στο παρακάτω σχήμα.



Εικόνα 6. Διάγραμμα Βάσης Δεδομένων

Στη συνέχεια ακολουθεί μια αναλυτική περιγραφή της χρήσης και των δεδομένων για όλους τους πίνακες της βάσης δεδομένων.

poi_crawlers: Ο πίνακας αυτός περιέχει όλους τους crawlers των σημείων ενδιαφέροντος που υποστηρίζει το σύστημα. Αναλυτικότερα, περιέχει το μοναδικό αναγνωριστικό, το

όνομα, μια σύντομη περιγραφή, την περιοχή κάλυψης και τη μορφή δεδομένων που υποστηρίζει.

categories: Ο πίνακας περιέχει τις κατηγορίες των σημείων ενδιαφέροντος που υποστηρίζει το σύστημα. Στον πίνακα αποθηκεύεται το μοναδικό αναγνωριστικό και το όνομα κάθε κατηγορίας.

campaigns: Ο συγκεκριμένος πίνακας παρέχει πληροφορίες για τις συλλογές που έχουν πραγματοποιηθεί από το σύστημα. Για κάθε συλλογή αποθηκεύεται το μοναδικό αναγνωριστικό, το όνομα, η περιγραφή και η ημερομηνία έναρξης και λήξης της συλλογής.

campaign_poi_crawlers: Ο συγκεκριμένος πίνακας χρησιμοποιείται για να προσδιορίζει τους crawlers των σημείων ενδιαφέροντος που χρησιμοποιήθηκαν σε κάθε συλλογή. Οι πληροφορίες που αποθηκεύονται είναι το μοναδικό αναγνωριστικό της συλλογής και το μοναδικό αναγνωριστικό των crawlers.

campaign_has_categories: Ο συγκεκριμένος πίνακας χρησιμοποιείται για να προσδιορίζει τις κατηγορίες που χρησιμοποιήθηκαν σε κάθε συλλογή και τις παραμέτρους συλλογής για κάθε κατηγορία, όπως αυτές αποτυπώνονται από την περίοδο και το περιεχόμενο συλλογής. Οι πληροφορίες που αποθηκεύονται είναι τα μοναδικά αναγνωριστικά της συλλογής και της κατηγορίας, καθώς και η περίοδος συλλογής για κάθε μία κατηγορία.

radius_locations&rectangle_locations: Οι συγκεκριμένοι πίνακες χρησιμοποιούνται για να προσδιορίζουν την περιοχή αναζήτησης για tacampaigns. Υπενθυμίζουμε ότι κάθε campaign υποχρεούται να περιλαμβάνει ακριβώς μία περιοχή αναζήτησης και συνεπώς κάθε συλλογή θα πρέπει να περιέχει την περιοχή αναζήτησης σε έναν από αυτούς τους δύο πίνακες ανάλογα με το ορισμό της περιοχής.

poi_crawls: Ο συγκεκριμένος πίνακας προσδιορίζει τα crawls για τα σημεία ενδιαφέροντος για tacampaigns. Οι πληροφορίες που αποθηκεύονται αφορούν τα μοναδικά αναγνωριστικά των crawls και των campaigns και τη χρονική στιγμή έναρξης των crawls.

crawl_poi_crawls: Ο συγκεκριμένος πίνακας χρησιμοποιείται για να προσδιορίζονται οι crawlers των σημείων ενδιαφέροντος που χρησιμοποιήθηκαν για κάθε crawl. Οι πληροφορίες που αποθηκεύονται στον πίνακα αφορούν τα μοναδικά αναγνωριστικά των crawls και των crawlers.

pois: Οσυγκεκριμένος πίνακας χρησιμοποιείται για την αποθήκευση των σημείων ενδιαφέροντος. Για να είναι το σύστημα ικανό να απαντάει σε πολύπλοκα ερωτήματα, εκτός από την πληροφορία που αφορά τα ίδια τα σημεία ενδιαφέροντος, αποθηκεύεται και πληροφορία που αφορά τον crawler με τον οποίο έγινε η συλλογή, η κατηγορία στην οποία ανήκει το σημείο ενδιαφέροντος και το crawl. Ακόμα αποθηκεύεται το μοναδικό αναγνωριστικό που χρησιμοποιεί η πηγή, ένα σημείο που προσδιορίζει χωρικά το σημείο ενδιαφέροντος, η ημερομηνία δημιουργίας που προέρχεται από την πηγή, το όνομα, η περιγραφή, η διεύθυνση, το τηλέφωνο, οι ώρες λειτουργίας, το URL, το είδος της άδειας χρήσης, η μέση βαθμολογία και ο αριθμός αξιολογήσεων από τις οποίες προήλθε η μέση βαθμολογία. Τέλος, για το GooglePlaces αποθηκεύεται η λίστα των placetypes που χαρακτηρίζουν το σημείο ενδιαφέροντος και για το Foursquare αποθηκεύονται οι κατηγορίες της Foursquare κατηγοριοποίησης (ιεραρχίας κατηγοριών) στις οποίες ανήκει το σημείο ενδιαφέροντος, καθώς και οι ετικέτες (tags) που το χαρακτηρίζουν.

coordinates: Ο συγκεκριμένος πίνακας χρησιμοποιείται για την αποθήκευση όλων των συντεταγμένων των σημείων ενδιαφέροντος. Η πληροφορία αυτή δεν θα μπορούσε να αποθηκευτεί στον πίνακα **pois**που αποθηκεύει τα σημεία ενδιαφέροντος, καθώς ένα

σημείο ενδιαφέροντος μπορεί να προσδιορίζεται από περισσότερες από μια τοποθεσίες. Η πληροφορία που αποθηκεύεται αφορά το μοναδικό αναγνωριστικό του σημείου ενδιαφέροντος και τις συντεταγμένες για συγκεκριμένο σημείο.

1.5.6 Αποφάσεις κατά τη φάση σχεδιασμού της βάσης δεδομένων

Βασικό κριτήριο κατά τον σχεδιασμό της βάσης ήταν η λειτουργικότητα που θα πρέπει να επιτελεί η εφαρμογή. Στην συνέχεια θα αναφερθούμε σε αυτή,προκειμένου να δικαιολογήσουμε τις σχεδιαστικές επιλογές της βάσης δεδομένων.

Η κατάσταση του συστήματος εξαρτάται από τους διαθέσιμους crawlers σημείων ενδιαφέροντος και από τις κατηγορίες των σημείων ενδιαφέροντος. Καθώς τα δεδομένα αυτά παίζουν πολύ σημαντικό ρόλο για την εφαρμογή θα πρέπει να έχουμε εξασφαλίσει ότι η βάση δεδομένων είναι ενημερωμένη με τα δεδομένα της εφαρμογής. Αυτό γίνεται με την αρχικοποίηση της βάσης κατά την οποία οι αντίστοιχοι πίνακες που περιέχουν τα παραπάνω δεδομένα συμπληρώνονται και στην συνέχεια παραμένουν αμετάβλητοι καθ' όλη την διάρκεια εκτέλεσης της εφαρμογής. Οι πίνακες αυτοί θα πρέπει να αλλάξουν μόνο στην περίπτωση που αλλάξει η κατάσταση του συστήματος, η οποία πραγματοποιείται από τον προγραμματιστή της εφαρμογής. Ο συνήθης λόγος για την αλλαγή της κατάστασης είναι η προσθήκη νέων crawlers και κατηγοριών στο σύστημα.

Πρωταρχικό στοιχείο της εφαρμογής είναι τα campaigns μέσω των οποίων πραγματοποιείται η συλλογή των σημείων ενδιαφέροντος. Τα campaigns είναι τα αποκλειστικά μέσα για τη συλλογή των παραπάνω δεδομένων, ενώ πρέπει να διευκρινιστεί ότι η συλλογή των δεδομένων δεν πραγματοποιείται άμεσα από το campaign, αλλά με την βοήθεια των crawls. Αυτό σημαίνει ότι κάθε campaign έχει τουλάχιστον ένα crawl και αντίστροφα κάθε crawl ανήκει υποχρεωτικά σε μόνο ένα campaign. Η ιδιαιτερότητα αυτή αποτυπώνεται και στην βάση δεδομένων, όπου τα σημεία ενδιαφέροντος συνδέονται μόνο με τα crawls, ενώ τα crawls συνδέονται με τα campaigns.

Τα campaigns περιλαμβάνουν ένα αριθμό από παραμέτρους που αφορούν τους crawlers που χρησιμοποιούν, τις κατηγορίες των σημείων ενδιαφέροντος που θα συλλέξουν με την χρήση λέξεων κλειδιών και τέλος την τοποθεσία της συλλογής. Οι παραπάνω παράμετροι παραμένουν αναλλοίωτες στη διάρκεια μιας συγκεκριμένης συλλογής και επομένως μπορούν να αποτυπωθούν στην βάση δεδομένων με πίνακες που αποθηκεύουν όλες αυτές τις πληροφορίες για κάθε campaign. Ιδιαίτερη μνεία πρέπει να γίνει στο γεγονός ότι η βάση δεδομένων υποστηρίζει την αποθήκευση δύο μορφών σχετικά με την περιοχή αναζήτησης. Καθώς όμως σε κάθε campaign επιτρέπεται μόνο μία τοποθεσία, η βάση δεδομένων δεν επιτρέπει για μία συγκεκριμένη campaign να υπάρχουν και οι δύο μορφές της τοποθεσίας. Τέλος, καθώς η χρήση πολλαπλών campaigns με όμοιες παραμέτρους συλλογής είναι επιτρεπτή, αυτό αποτυπώνεται στη βάση με την χρήση μοναδικού αναγνωριστικού ανεξάρτητο των παραμέτρων.

Για να διατηρηθεί η ακεραιότητα των δεδομένων που είναι αποθηκευμένα στην βάση δεδομένων κάθε χρονική στιγμή, η βάση διέπεται από ένα πλήθος κανόνων κατά την διαγραφή των στοιχείων της. Μέσω αυτών των κανόνων δεν επιτρέπεται η διαγραφή των crawlers και των κατηγοριών που χρησιμοποιούνται από campaigns, ώστε να μην είναι εφικτό να υπάρχουν αποθηκευμένα σημεία ενδιαφέροντος και ανοικτά δεδομένα που δεν συνδέονται με τους crawlers και τις κατηγορίες με τους οποίους πραγματοποιήθηκε η συλλογή. Για να διαγραφούν τα παραπάνω στοιχεία απαιτείται πρώτα η διαγραφή όλων των campaigns που κάνουν χρήση αυτών των στοιχείων. Ένας ακόμα κανόνας που διέπει την βάση είναι ότι τα σημεία ενδιαφέροντος θα πρέπει πάντα να συνδέονται με το

campaign για το οποίο συλλέχθηκαν. Επομένως, η διαγραφή campaign επιφέρει την διαγραφή όλων των δεδομένων που έχουν συλλεχθεί από αυτήν, καθώς και όλων των παραμέτρων με τις οποίες πραγματοποιήθηκε.

1.6 Τεχνολογίες ανάπτυξης

Για τον AreaProfiler βασιστήκαμε κυρίως στις ακόλουθες τρεις τεχνολογίες. Η πρώτη τεχνολογία είναι η Java, μια γλώσσα προγραμματισμού γενικού σκοπού, η δεύτερη τεχνολογία είναι ο ApacheTomcat [7] (WebServer)και τέλος το σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων PostgreSQL[8]με την επέκταση PostGIS[9] για την υποστήριξη χωρικών δεδομένων.

1.6.1 Java

Για την ανάπτυξη με την γλώσσα προγραμματισμού Java στην έκδοση 7, χρησιμοποιήσαμε το ανοικτού κώδικα γραφικό περιβάλλον ανάπτυξης Eclipse[10] με το οποίο διευκολύνεται η ανάπτυξη κώδικα, ενώ χρησιμοποιείται και σαν βάση για μια σειρά από πρόσθετα εργαλεία που χρησιμοποιήσαμε.

Για την δημιουργία κώδικα ελέγχου (testcases) χρησιμοποιήσαμε το εργαλείο JUnit[11] με την χρήση του οποίο απλοποιείται και μειώνεται ο χρόνος ανάπτυξης του κατάλληλου κώδικα. Παράλληλα χρησιμοποιήσαμε και το εργαλείο Clover[12] με την βοήθεια του οποίου ελέγξαμε ποια τμήματα του κώδικα της εφαρμογής έχουν καλυφθεί από κώδικα ελέγχου.

Καθώς η εφαρμογή απαιτούσε την χρήση πολλών βιβλιοθηκών χρησιμοποιήσαμε το εργαλείο Maven [13] το οποίο αναλαμβάνει την διαχείριση όλων των βιβλιοθηκώνπου χρησιμοποιούνται από την εφαρμογή. Με τη χρήση του απλοποιήθηκε σημαντικά η μεταφορά του έργου σε διαφορετικά μηχανήματα.

Εκτός από τις standard βιβλιοθήκες της Java για την ανάπτυξη του υποσυστήματος χρησιμοποιήσαμε μια σειρά από εξωτερικέςβιβλιοθήκες. Για την μετατροπή της μορφής JSON σε αντικείμενα Java και αντίστροφα χρησιμοποιήσαμε την βιβλιοθήκη "google-json"[14]. Για την επεξεργασία των εντολών που δίνονται σαν ορίσματα στην κύρια μέθοδο μέσα από την γραμμή εντολών χρησιμοποιήσαμε την βιβλιοθήκη "commonsCLI" [15] ενώ για την διαχείριση αρχείων χρησιμοποιήσαμε την βιβλιοθήκη "commonsIO"[16]. Για τη σύνδεση της βάσης δεδομένων με την Java χρησιμοποιήσαμε την βιβλιοθήκη "PostgreSQLJDBCDriver"[17] της PostgreSQL.

---

7 http://tomcat.apache.org/
8 http://www.postgresql.org/
9 http://www.postgis.net/
10 https://www.eclipse.org/
11 http://junit.org/
12 https://www.atlassian.com/software/clover/overview
13 http://maven.apache.org/
14 https://code.google.com/p/google-gson/
15 http://commons.apache.org/proper/commons-cli/
16 http://commons.apache.org/proper/commons-io/
17 https://jdbc.postgresql.org/

1.6.2 ApacheTomcat

Ο ApacheTomcat είναι ένας webserverπου παρέχει τις απαιτούμενες βιβλιοθήκες, ώστε να υποστηρίζει το πρωτόκολλο επικοινωνίας HTTP. Με τη χρήση του συγκεκριμένου πρωτοκόλλου οι clients που υποστηρίζουν το πρωτόκολλο μπορούν να συνδέονται στο server, με σκοπό να πραγματοποιούν αιτήσεις και να λαμβάνουν πίσω αποκρίσεις.

1.6.3 PostgreSQL/PostGIS

Η τρίτη τεχνολογία στην οποία βασιστήκαμε είναι τοσύστημα διαχείρισης σχεσιακών βάσεων δεδομένων PostgreSQLπου επιτρέπει την αποθήκευση των δεδομένων σε σχεσιακή μορφή.Χρησιμοποιήσαμε επίσης, την επέκταση PostGIS της PostgreSQL που δίνει τη δυνατότητα χρήσης χωρικών δεδομένων (σημεία, πολύγωνα) ως στήλες σε πίνακες της βάσης δεδομένων.

1.7 Υλοποίηση του Area Profiler

ToAreaProfiler εργαλείο που αναπτύξαμε στα πλαίσια του έργου Citysense στηρίχθηκε σε ένα παλαιότερο, πρώιμο πρωτότυπο ενός παλαιότερου έργου του Ινστιτούτου Πληροφοριακών Συστημάτων. Το συγκεκριμένο όμως εργαλείο, έπρεπε να ανανεωθεί ριζικά, προκειμένου να ανταπεξέλθει στις αυξημένες απαιτήσεις του Citysense. Κάποιες από τις κυριότερες αλλαγές θα περιγραφούν με συντομία στην παρούσα ενότητα.

1.7.1 Χρήση βάσης δεδομένων PostgreSQL/PostGIS

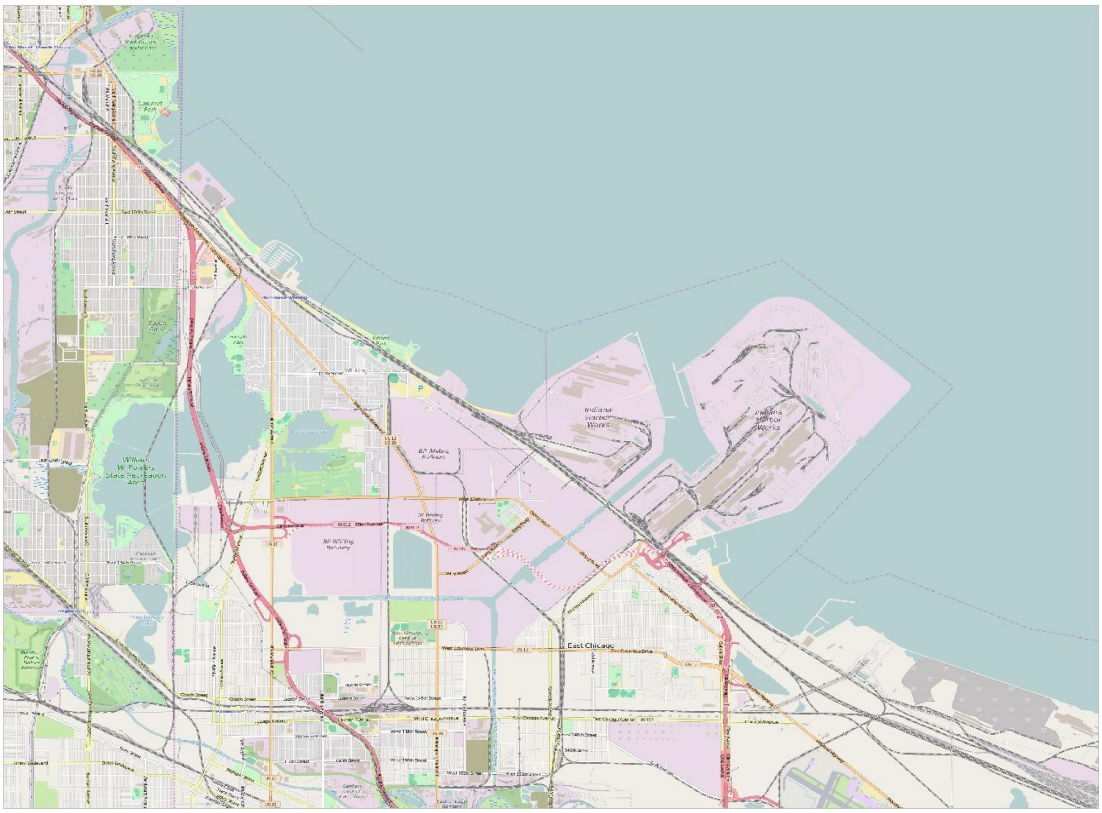
Η αρχική έκδοση του AreaProfiler χρησιμοποιούσε ως βάση δεδομένων τη MySQL. Τα χωρικά δεδομένα, όπως η τοποθεσία των σημείων ενδιαφέροντος (γεωγραφικό μήκος, γεωγραφικό πλάτος), αποθηκεύονταν ως βασικοί τύποι δεδομένων της MySQL, με αποτέλεσμα να μην υποστηρίζονται χωρικά ερωτήματα σε μορφή και απόδοση απαραίτητες για την εφαρμογή CitySense. Συνεπώς κρίθηκε αναγκαίο, να τροποποιηθεί ο AreaProfiler, ώστε να χρησιμοποιεί πλέον τη βάση δεδομένων PostgreSQL με την επέκταση PostGIS και να χρησιμοποιεί χωρικούς τύπους δεδομένων εγγενείς στην PostgreSQL/PostGIS για την αποθήκευση χωρικών δεδομένων που θα συμμετέχουν σε χωρικές σχέσεις και SQL ερωτήματα.

1.7.2 Τεμαχισμός πολύ μεγάλων περιοχών

Η αρχική έκδοση του AreaProfiler λειτουργούσε με τεμαχισμό μίας περιοχής και αναδρομική αναζήτηση σε κάθε κομμάτι ξεχωριστά, αφού έχει προηγηθεί μία αποτυχημένη αναζήτηση στην περιοχή, κατά την οποία προέκυπτε ότι η περιοχή περιέχει περισσότερα αποτελέσματα από αυτά που επιστράφηκαν στην αναζήτηση. Σε μεγάλες αστικές περιοχές, όπως μια ολόκληρη πόλη(που είναι η τυπικήπερίπτωση στο CitySense), η συγκεκριμένη μέθοδος κατανάλωνε περιττούς πόρους (επαναλαμβανόμενες κλήσεις στα API των GooglePlaces και Foursquare), καθώς η αναδρομή θα έπρεπε να φτάσει σε μεγάλο βάθος μέχρι να αρχίσουν να έρχονται αποτελέσματα, γεγονός που σημαίνει ότι θα είχε προηγηθεί μεγάλος αριθμός περιττών, αποτυχημένων αναζητήσεων. Αυτή η μέθοδος αναδρομικής αναζήτησης, σε συνδυασμό με το ότι ο AreaProfiler εκτελούσε αναζήτηση για κάθε κατηγορία σημείων ενδιαφέροντος ξεχωριστά (οπότε και η απόδοση του συστήματος επηρεαζόταν πολλαπλασιαστικά από το πλήθος των κατηγοριών κάθε αναζήτησης),

καθιστούσε μη πρακτική τη χρήση της συγκεκριμένης αναδρομικής αναζήτησης για την αναζήτηση όλων των σημείων ενδιαφέροντος μίας μεγάλης περιοχής.

Η αναγκαία τροποποίηση στο Citysenseγια τον AreaProfiler λαμβάνει υπόψη της το αρχικό μέγεθος της περιοχής, πριν την εκτέλεση της αναζήτησης. Σε περίπτωση που η περιοχή αυτή είναι μεγαλύτερη από 0.06 μοίρες, είτε κατά το γεωγραφικό μήκος, είτε κατά το γεωγραφικό πλάτος, τότε γίνεται τεμαχισμός της περιοχής σε τετραγωνικά κομμάτια πλευράς 0.03 μοιρών και η αναζήτηση εκτελείται σε κάθε κομμάτι ξεχωριστά. Σε διαφορετική περίπτωση, αν, δηλαδή, η περιοχή είναι μικρότερη από τετράγωνο πλευράς 0.06 μοιρών, τότε δεν εκτελείται τεμαχισμός, καθώς μία τέτοια περιοχή θεωρείται αρκούντως μικρή. Στα μικρά κομμάτια, πλέον, αν η πρώτη αναζήτηση δεν επιστρέψει όλα τα σημεία ενδιαφέροντος, τότε το κομμάτι τεμαχίζεται περαιτέρω και η αναζήτηση εκτελείται αναδρομικά στα ακόμα μικρότερα κομμάτια. Οι συντεταγμένες των γωνιών των τετραγωνικών κομματιών τεμαχισμού προκύπτουν εκ κατασκευής να είναι πολλαπλάσια των 0.03 μοιρών, με σκοπό το πλέγμα που σχηματίζουν να μην εξαρτάται από τις ακριβείς συντεταγμένες της αρχικής περιοχής αναζήτησης. Αυτό έχει ως αποτέλεσμα ο αριθμός των κομματιών να είναι ελαφρώς μεγαλύτερος από τον ελάχιστο απαιτούμενο, αλλά σε πολύ μεγάλες περιοχές η επίδραση αυτής της σχεδιαστικής επιλογής στην απόδοση είναι αμελητέα. Έστω η περιοχή που απεικονίζεται στο παρακάτω σχήμα. Η περιοχή αυτή αποτελεί ορθογώνιο με πλευρές περίπου 0.18 μοίρες κατά γεωγραφικό μήκος και 0.12 μοίρες κατά γεωγραφικό πλάτος.

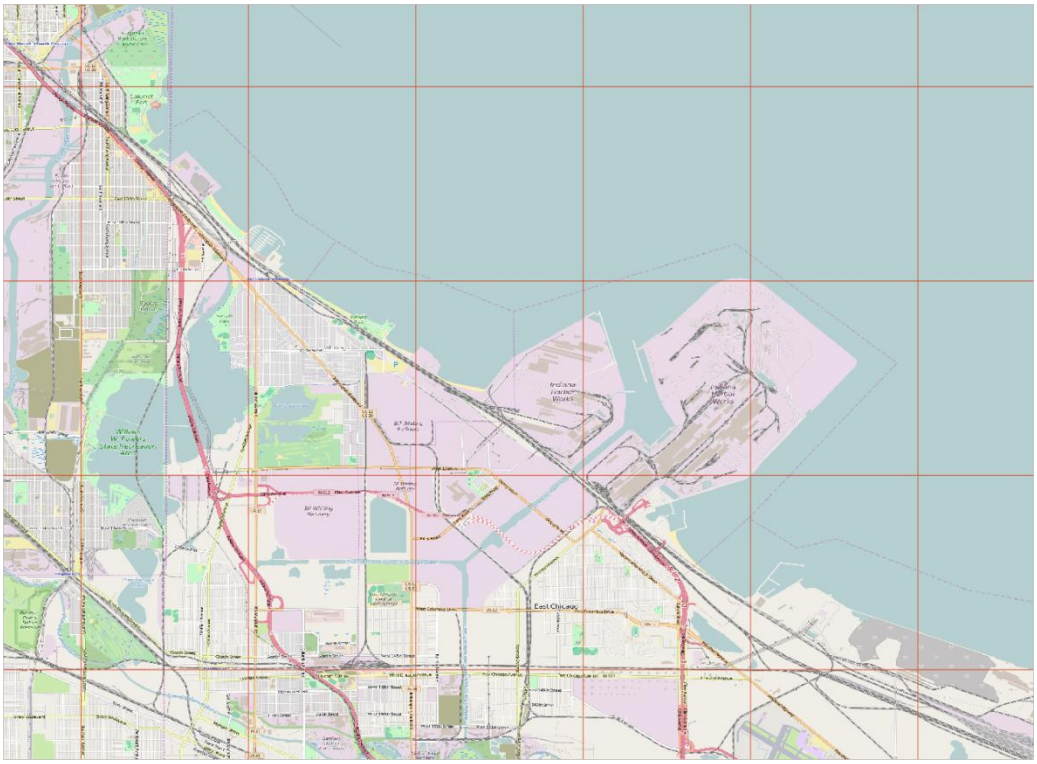


Εικόνα 7. Η περιοχή ενδιαφέροντος

Καθώς η περιοχή αυτή ικανοποιεί την προϋπόθεση για τεμαχισμό, ο τεμαχισμός της γίνεται όπως παρουσιάζεται στο παρακάτω σχήμα.

Τα 35 τετραγωνικά κομμάτια που εμφανίζονται (ολικώς ή μερικώς) έχουν πλάτος 0.03 μοίρες το καθένα. Σε αυτά ξεχωριστά θα εκτελεστεί η αναζήτηση. Σε οποιαδήποτε από αυτά χρειαστεί η αναζήτηση να πάει σε μεγαλύτερο βάθος, λόγω αποτυχημένης

αναζήτησης στα ίδια (που συνεπάγεται ανάγκη για περαιτέρω τεμαχισμό), αυτό θα γίνει αναδρομικά ανά κομμάτι.



Εικόνα 8. Η περιοχή ενδιαφέροντος τεμαχισμένη

1.7.3 Ολική σάρωση και κατηγοριοποίηση αποτελεσμάτων

Η αρχική έκδοση του AreaProfiler βασιζόταν σε στοχευμένη αναζήτηση στις πηγές σημείων ενδιαφέροντος GooglePlaces και Foursquare. Για παράδειγμα, γινόταν μία αναζήτηση για "Εστιατόρια" και για κάθε αποτέλεσμα που επιστρεφόταν αποθηκευόταν ο στόχος αναζήτησης ως η κατηγορία του, δηλαδή η κατηγορία στην οποία θεωρούνταν ότι εν προκειμένω ανήκει το αποτέλεσμα αναζήτησης ήταν τα "Εστιατόρια". Αυτό είχε ως αποτέλεσμα άλλες κατηγορίες στις οποίες θα μπορούσε να ανήκει το σημείο ενδιαφέροντος να χάνονται εντελώς, λόγου χάρη ότι κάποιο αποτέλεσμα της συγκεκριμένης αναζήτησης θα μπορούσε να είναι επίσης "Καφέ" ή "Μπαρ". Για τις ανάγκες του CitySense, από μία περιοχή πρέπει να ληφθούν όλα τα σημεία ενδιαφέροντος με όλα τα χαρακτηριστικά των κατηγοριών στις οποίες ανήκουν. Οπότε ο AreaProfiler τροποποιήθηκε, ώστε να σαρώνει όλες τις κατηγορίες σημείων ενδιαφέροντος για GooglePlaces και Foursquare. Για την αποδοτικότερη διαχείριση πόρων χρησιμοποιήθηκε ομαδοποίηση, ακόμα και ετερογενής, των κατηγοριών κατά την αναζήτηση. Μαζί με το σημείο ενδιαφέροντος αποθηκεύονται πλέον και όλα τα στοιχεία για τις κατηγορίες όπου ανήκει στην υπηρεσία που το παρέχει. Για GooglePlaces αποθηκεύεται η λίστα των placetypes που χαρακτηρίζουν το σημείο ενδιαφέροντος και για Foursquare αποθηκεύονται οι κατηγορίες της Foursquare κατηγοριοποίησης (ιεραρχίας κατηγοριών) στις οποίες ανήκει το σημείο ενδιαφέροντος, καθώς και οι ετικέτες (tags) που χαρακτηρίζουν το σημείο ενδιαφέροντος.

1.8 ΣυλλογήStreamingSocial Media Data

Η ενότητα αυτή περιγράφει τη διαδικασία συλλογής γεωχωρικών δεδομένων από τα μέσα κοινωνικής δικτύωσης και συγκεκριμένα του Twitterκαι του Foursquare. Η ιδιαιτερότητα αυτής της συλλογής δεδομένων είναι ότι γίνεται με δυναμικό τρόπο σε πραγματικό χρόνο. Σε αντίθεση με τα σημεία ενδιαφέροντος που συλλέγονται σε τακτικά χρονικά διαστήματα, όπως αυτά ορίζονται από το campaign, τα tweetsκαι tacheck-ins συλλέγονται απευθείας από το TwitterStreamσε πραγματικό χρόνο. Συγκεκριμένα συλλέγονται όλα τα γεωχωρικά προσδιορισμένα tweetsπου διαθέτει το TwitterStreamingAPI σε μια ορισμένη γεωχωρικά περιοχή. Στη συνέχεια από αυτά φιλτράρονται όσα αποτελούν check-instης εφαρμογής SwarmApp (Foursquare) και αποθηκεύονται χωριστά μαζί με επιπλέον πληροφορίες για τα σημεία ενδιαφέροντος που αυτά αφορούν. Συγκεκριμένα υλοποιούνται δύο μέθοδοι: η "Twitter Stream Crawler" και η μέθοδος "Check-In Details".Οι συγκεκριμένες μέθοδοι αναλύονται παρακάτω.

1.8.1 Twitter Stream Crawler

Με τη μέθοδο Twitter Stream Crawler είναι δυνατή η ανάκτηση όλων των tweetsπου δημοσιεύονται κατά τη διάρκεια λειτουργίας του crawler. Χρησιμοποιείται για τη συλλογή όλων των tweetsπου δημοσιεύονται μέσα σε μια γεωγραφική περιοχή.

Ένα αίτημα TwitterStreamCrawler περιέχει τις εξής παραμέτρους:

- consumer_key – Το κλειδί της εφαρμογής, το οποίο παρέχεται από το Twitter και ταυτοποιεί την εφαρμογή για λογαριασμό της οποίας γίνεται το αίτημα.
- consumer_secret– Το password τηςεφαρμογής, το οποίο παρέχεται από το Twitter και ταυτοποιεί την εφαρμογή για λογαριασμό της οποίας γίνεται το αίτημα.
- access_token–Το αναγνωριστικό που αποδίδεται στην εφαρμογή αμέσως μετά την επιτυχή ταυτοποίησή της, που ορίζει τα προνόμια (δεδομένα στα οποία έχει πρόσβαση) της εφαρμογής.
- access_token_secret– Το passwordγια τον ορισμό των προνομίων της εφαρμογής για λογαριασμό της οποίας γίνεται το αίτημα.
- location– Το ζεύγος σημείων (γεωγραφικό μήκος, γεωγραφικό πλάτος) που καθορίζουν ένα boundingbox, με βάση το οποίο φιλτράρεται η ροή δημοσιεύσεων του Twitter. Θα περιληφθούν μόνο tweetsμε γεωχωρικό προσδιορισμό εντός του boundingbox. Στο ζεύγος σημείων πρώτο σημειώνεται αυτό που αναπαριστά τη νοτιότερη γωνία του bounding box.

Η μορφή της απόκρισης φαίνεται στην παρακάτω εικόνα.

```
"contributors": null,
"truncated": false,
"text": "TeeMinus24's Shirt of the Day is Palpatine/Vader '12. Support the Sith. Change you can't stop. http://t.
"in_reply_to_status_id": null,
"id": 175090352598945794,
"entities": {
    "user_mentions": [],
    "hashtags": [],
    "urls": [
        {
            "indices": [
                95,
                115
            ],
            "url": "http://t.co/wFh1cCep",
            "expanded_url": "http://fb.me/1isEdQJSq",
            "display_url": "fb.me/1isEdQJSq"
        }
    ]
},
"retweeted": false,
"coordinates": null,
"source": "<a href="\&quot;http://www.facebook.com/twitter\&quot;" rel="\&quot;nofollow\&quot;">Facebook</a>",
"in_reply_to_screen_name": null,
"id_str": "175090352598945794",
"retweet_count": 0,
"in_reply_to_user_id": null,
"favorited": false,
"user": {
    "follow_request_sent": null,
```


Εικόνα9. Twitter Streaming APIlocationresponse

Η απόκριση περιέχει μία λίστα (που συνεχώς ανανεώνεται) από tweetsπου δημοσιεύονται σε πραγματικό χρόνο από την περιοχή που ορίζεται από το boundingbox. Κάθε εγγραφή tweet περιέχει τα εξής απαραίτητα πεδία για τον AreaProfiler:

- tweet_id – Το αναγνωριστικό toutweet, απαραίτητο για την ταυτοποίηση του tweet.
- user_id –Το αναγνωριστικό του χρήστη που δημοσίευσε το tweet.
- content– Το κειμενικό περιεχόμενο του tweet.
- latitude– Το γεωγραφικό πλάτος του σημείου από όπου δημοσιεύτηκε το tweet.
- longtitude – Το γεωγραφικό μήκος του σημείου από όπου δημοσιεύτηκε το tweet.
- timestamp– Ο χρονικός προσδιορισμός της στιγμής κατά την οποία δημοσιεύτηκε το tweet.
- hashtags– Η λίστα που περιλαμβάνει όλα τα hashtags, τα οποία περιέχονται στο κείμενο toutweet.

1.8.2 Check-In DetailsCrawler

Στη συνέχεια όσα tweetsαποτελούν δημοσιευμένο foursquarecheck-inστο Twitter (περιέχουν δηλαδή σύνδεσμο που ξεκινά με το: «https://www.swarmapp.com/c/»), χρησιμοποιούνται για την εύρεση περισσότερων πληροφόριών για το σημείο ενδιαφέροντος που αφορούν. Αυτό επιτυγχάνεται χρησιμοποιώντας τογεωγραφικό μήκος και πλάτος του tweet, που περιέχει το check-inαυτό.

Ένα αίτημα Twitter Check-In Details περιέχει τις εξής παραμέτρους:

- client_id– Το αναγνωριστικό της εφαρμογής, το οποίο παρέχεται από το Foursquare και ταυτοποιεί την εφαρμογή για λογαριασμό της οποίας γίνεται το αίτημα.
- client_secret – Το password της εφαρμογής, το οποίο παρέχεται από το Foursquarekαι ταυτοποιεί την εφαρμογή για λογαριασμό της οποίας γίνεται το αίτημα.
- latitude– Το γεωγραφικό πλάτος του σημείου από όπου δημοσιεύτηκε το tweet που περιέχει το check-in.
- longtitude – Το γεωγραφικό μήκος του σημείου από όπου δημοσιεύτηκε το tweetπου περιέχει το check-in.

Ένα ενδεικτικό αίτημα για λήψη λεπτομερειών για το σημείο ενδιαφέροντος από όπου έγινε το check-in του tweet είναι το εξής:

https://api.foursquare.com/v2/venues/search?client_id=client_id&v=20130815%20&client_secret=client_secret &ll=latitude,longtitude

Η μορφή της απόκρισης φαίνεται στην παρακάτω εικόνα.


```
    }
],
response: {
    venue: {
        id: "427c0500f964a52097211fe3",
        name: "The Metropolitan Museum of Art",
        contact: {
            phone: "+12125357710",
            formattedPhone: "+1 212-535-7710",
            twitter: "metmuseum",
            facebook: "6296252634",
            facebookUsername: "metmuseum",
            facebookName: "The Metropolitan Museum of Art, New York"
        },
        location: {
            address: "1000 5th Ave",
            crossStreet: "btwn E 80th & E 84th St",
            lat: 40.778936659294864,
            lng: -73.96229820007625,
            postalCode: "10028",
            mayNotNeedAddress: false,
            cc: "US",
            city: "New York",
            state: "NY",
            country: "United States",
            formattedAddress: [
                "1000 5th Ave (btwn E 80th & E 84th St)",
                "New York, NY 10028",
                "United States"
            ]
        },
        canonicalUrl: "https://foursquare.com/v/the-metropolitan-museum-of-art
```


Εικόνα10. Foursquare API check-in details response

Η απόκριση περιλαμβάνει αναλυτικές πληροφορίες για το σημείο ενδιαφέροντος. Συγκεκριμένα περιέχει τα εξής απαραίτητα πεδία για τον AreaProfiler:

- venue_name – Το όνομα του σημείου ενδιαφέροντος.
- venue_id – Το αναγνωριστικό του σημείου ενδιαφέροντος.
- venue_categories – Οι κατηγορίες της ιεραρχίας κατηγοριών του Foursquare στις οποίες ανήκει το σημείο ενδιαφέροντος.

1.8.3 Αποθήκευση και ολοκλήρωση πληροφορίας από StreamingSocialMediaData

Από τις παραπάνω μεθόδους προκύπτει πληροφορία που αφορά τον γεωχωρικό, τον χρονικό προσδιορισμό των tweets, αλλά και από την πλευρά περιεχομένου περιέχει τα hashtags, αλλά και τα σημεία ενδιαφέροντος από τα οποία οι χρήστες του Twitterκαι του SwarmΑρέκαναν check-in.Παρακάτω παρουσιάζονται συνοπτικά τα attributes των δεδομένων, τα οποία μαζεύονται δυναμικά και αποθηκεύονται σε μορφή jsonσε directory structure.

ΑπότοresponsetηςμεθόδουTwitterStreamCrawlerκαι της μεθόδου Check-In DetailsCrawlerδημιουργείται το αρχείο jsonμε την εξής μορφή:

```
1  {"venuename":"","hashtags":"","latitude":"","venueID":"","longtitude":"","tweet":"@ethanwp_98 can decide if you or the dog is cuter??","userid":"3245648597","tweetid":"772785771954577408","venuecategory":"","timestamp":"Mon Sep 05 16:17:40 EEST 2016"}
2  {"venuename":"","hashtags":"","latitude":"","venueID":"","longtitude":"","tweet":"@DGodfatherMoody @CBSThisMorning @MartinTruex_Jr @colecuster00 @JHNemechek now you know how Indycar folks feel we had a race yesterday too!","userid":"21515621","tweetid":"772785776090288128","venuecategory":"","timestamp":"Mon Sep 05 16:17:41 EEST 2016"}
3  {"venuename":"","hashtags":"","latitude":"","venueID":"","longtitude":"","tweet":"@SaraSavoy @Stillagroup Thank U 4 sharing All These Pics!? I MUST ask, R those wristbands obtainable somewhere? Have a Great Day!???","userid":"2396194225","tweetid":"772785788291522560","venuecategory":"","timestamp":"Mon Sep 05 16:17:44 EEST 2016"}
4  {"venuename":"","hashtags":"\t\t|Frankfort\t\t|job\t\t|KellyJobs\t\t|KellyServices\t\t|Hiring","latitude":"41.4958665","venueID":"","longtitude":"-87.8486613","tweet":"See our latest #Frankfort, IL #job and click to apply: Production Operators - https:\/\/t.co\/SyqK9lJItU #KellyJobs #KellyServices #Hiring","userid":"20831281","tweetid":"772785789797228544","venuecategory":"","timestamp":"Mon Sep 05 16:17:44 EEST 2016"}
5  {"venuename":"","hashtags":"\t\t|endofsummer","latitude":"41.69866727","venueID":"","longtitude":"-88.44454841","tweet":"One last run #endofsummer @optrix @ Raging Waves Waterpark https:\/\/t.co\/0z1ecmYxZQ","userid":"50056498","tweetid":"772785797091188736","venuecategory":"","timestamp":"Mon Sep 05 16:17:46 EEST 2016"}
6  {"venuename":"","hashtags":"","latitude":"","venueID":"","longtitude":"","tweet":"@GennyGL @ChapStickThvg I wonder if they realize by that logic their mom\/sister is open for rape when alcohol is in play...","userid":"411962153","tweetid":"772785797292404736","venuecategory":"","timestamp":"Mon Sep 05 16:17:46 EEST 2016"}
7  {"venuename":"","hashtags":"","latitude":"","venueID":"","longtitude":"","tweet":"I just regret we lost respect ? never getting back","userid":"2645977620","tweetid":"772785801008705536","venuecategory":"","timestamp":"Mon Sep 05 16:17:47 EEST 2016"}
8  {"venuename":"","hashtags":"\t\t|Illmatic","latitude":"","venueID":"","longtitude":"","tweet":"Hey Nas, thanks for making #Illmatic . It's dope.","userid":"2225558184","tweetid":"772785808248012800","venuecategory":"","timestamp":"Mon Sep 05 16:17:49 EEST 2016"}
9  {"venuename":"","hashtags":"\t\t|LaborDayWeekend\t\t|LaborDay\t\t|LaborDay2016\t\t|WalbraWritings\t\t|Writing\t\t|Reading","latitude":"","venueID":"","longtitude":"","tweet":"Enjoying this day of reflection. #LaborDayWeekend #LaborDay #LaborDay2016 #WalbraWritings #Writing #Reading ???. https:\/\/t.co\/MSzscOLuhs","userid":"196333681","tweetid":"772785806452879360","venuecategory":"","timestamp":"Mon Sep 05 16:17:48 EEST 2016"}
10 {"venuename":"","hashtags":"","latitude":"","venueID":"","longtitude":"","tweet":"Dad's making me breakfast ?","userid":"714763195","tweetid":"772785810219360256","venuecategory":"","timestamp":"Mon Sep 05 16:17:49 EEST 2016"}
11 {"venuename":"","hashtags":"","latitude":"","venueID":"","longtitude":"","tweet":"@MadisonLintonn the world is a sad place","userid":"610741958","tweetid":"772785828288487424","venuecategory":"","timestamp":"Mon Sep 05 16:17:53 EEST 2016"}
12 {"venuename":"","hashtags":"\t\t|hiring\t\t|job\t\t|Healthcare\t\t|Trafalgar\t\t|CareerArc","latitude":"39.4161604","venueID":"","longtitude":"-86.150823","tweet":"We're #hiring! Read about our latest #job opening here: Pharmacy Technician - https:\/\/t.co\/yY7ObQCp7I #Healthcare #Trafalgar, IN #CareerArc","userid":"21725584","tweetid":"772785835217264640","venuecategory":"","timestamp":"Mon Sep 05 16:17:55 EEST 2016"}
13 {"venuename":"","hashtags":"","latitude":"41.8829","venueID":"","longtitude":"-87.6267","tweet":"60% Off Nail Services https:\/\/t.co\/kZKYr7g3HQ","userid":"533314954","tweetid":"772785836110782464","venuecategory":"","timestamp":"Mon Sep 05 16:17:55 EEST 2016"}
14 {"venuename":"","hashtags":"","latitude":"","venueID":"","longtitude":"","tweet":"@Ajzelner I saw him filming something in front of Lane once, but I couldn't find it","userid":"1491522756","tweetid":"772785837759135744","venuecategory":"","timestamp":"Mon
```

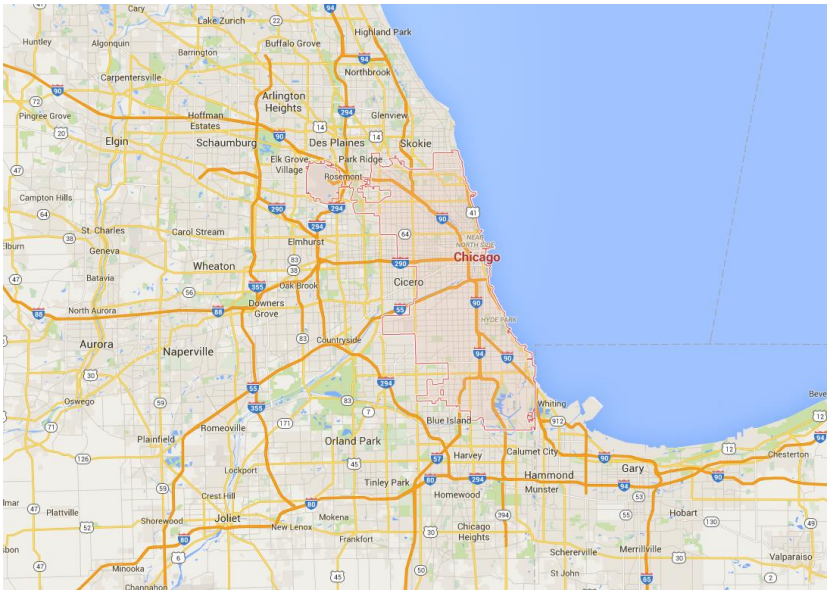

Εικόνα 11. Αρχείο αποθήκευσης TwitterStreamingData

Κάθε Json αντικείμενο αντιπροσωπεύει ένα tweetκαι κάθε name-valueζευγάρι χρησιμεύει στην αποθήκευση της πληροφορίας που περιέχεται στο tweetκαι χρησιμοποιεί η εφαρμογή.Συγκεκριμένα περιέχει τα εξής attributes:

- tweet_id – Το αναγνωριστικό τουtweet, απαραίτητο για την ταυτοποίηση του tweet, καθώς είναι μοναδικό για κάθε tweet.
- user_id –Το αναγνωριστικό του χρήστη που δημοσίευσε το tweet.
- hashtags– Η λίστα που περιλαμβάνει όλα τα hashtags, τα οποία περιέχονται στο κείμενο τουtweet.Εάν σε αυτό το tweet δενέχουν δημοσιευτεί hashtags, το πεδίο αυτό παραμένει κενό.
- tweet– Το κειμενικό περιεχόμενο του tweet.
- latitude– Το γεωγραφικό πλάτος του σημείου από όπου δημοσιεύτηκε το tweet.Εάν ο χρήστης έχει απενεργοποιήσει τη δημοσιευση της τοποθεσίας του, το πεδίο αυτό παραμένει κενό.
- longtitude – Το γεωγραφικό μήκος του σημείου από όπου δημοσιεύτηκε το tweet. Εάν ο χρήστης έχει απενεργοποιήσει τη δημοσιευση της τοποθεσίας του, το πεδίο αυτό παραμένει κενό.
- timestamp– Ο χρονικός προσδιορισμός της στιγμής κατά την οποία δημοσιεύτηκε το tweet, εκφρασμένος σε θερινή ώρα Ελλάδας.
- venue_name – Το όνομα του σημείου ενδιαφέροντος. Εάν σε αυτό το tweet δενέχει δημοσιευτεί κάποιο checkinσε σημείο ενδιαφέροντος, το πεδίο αυτό παραμένει κενό.
- venue_id – Το αναγνωριστικό του σημείου ενδιαφέροντος, το οποίο είναι μοναδικό για κάθε σημείο ενδιαφέροντος. Εάν σε αυτό το tweet δενέχει δημοσιευτεί κάποιο checkinσε σημείο ενδιαφέροντος, το πεδίο αυτό παραμένει κενό.
- venue_categories – Οι κατηγορίες της ιεραρχίας κατηγοριών του Foursquare στις οποίες ανήκει το σημείο ενδιαφέροντος. Εάν σε αυτό το tweet δενέχει δημοσιευτεί κάποιο checkinσε σημείο ενδιαφέροντος, το πεδίο αυτό παραμένει κενό.

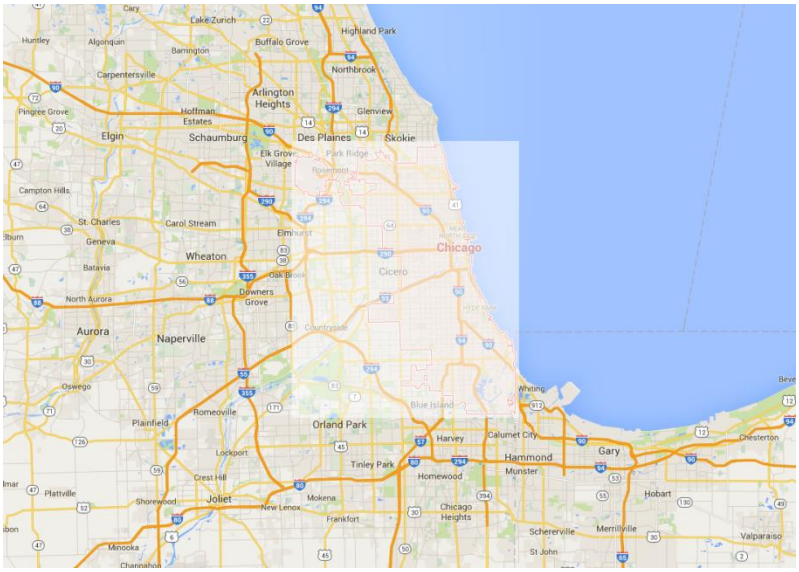
1.9 Σενάριο χρήσης του Area Profiler για το Chicago

Η ενότητα αυτή περιγράφει τρία σενάρια χρήσης της εφαρμογής AreaProfilerγια τις ανάγκες του CitySense. Το πρώτο σενάριο αφορά την έναρξη μίας καινούργιας campaign για συλλογή όλων των σημείων ενδιαφέροντος από το GooglePlaces για το Chicago. Το δεύτερο σενάριο αφορά την έναρξη μίας καινούργιας campaign για συλλογή όλων των σημείων ενδιαφέροντος από το Foursquare για το Chicago. Το τρίτο σενάριο αφοράτην έναρξη μίας καινούργιας campaign για συλλογή όλων τωντweetsκαι Check-insαπό το Twitterγια το Chicago.Η περιοχή ενδιαφέροντος για το Chicago, παρουσιάζεται στην παρακάτω εικόνα.



Εικόνα 12. Η ευρύτερη περιοχή του Chicago

Καθώς ο AreaProfiler απαιτεί τον καθορισμό μίας ορθογώνιαςή μίας κυκλικής περιοχής στο αίτημα για έναρξη ενός campaign, επιλέγουμε να προσεγγίσουμε εξωτερικά με ένα ορθογώνιο την παραπάνω περιοχή ενδιαφέροντος. Ως συντεταγμένες της νοτιοδυτικής γωνίας του ορθογωνίου μπορούμε να επιλέξουμε (γεωγραφικό μήκος, γεωγραφικό πλάτος) το (-87.949573, 41.627744) και ως συντεταγμένες της βορειοανατολικής γωνίας μπορούμε να επιλέξουμε το (-87.498048, 42.033020). Η επιλογή μας φαίνεται στην παρακάτω εικόνα.



Εικόνα 13. Η περιοχή του Chicago

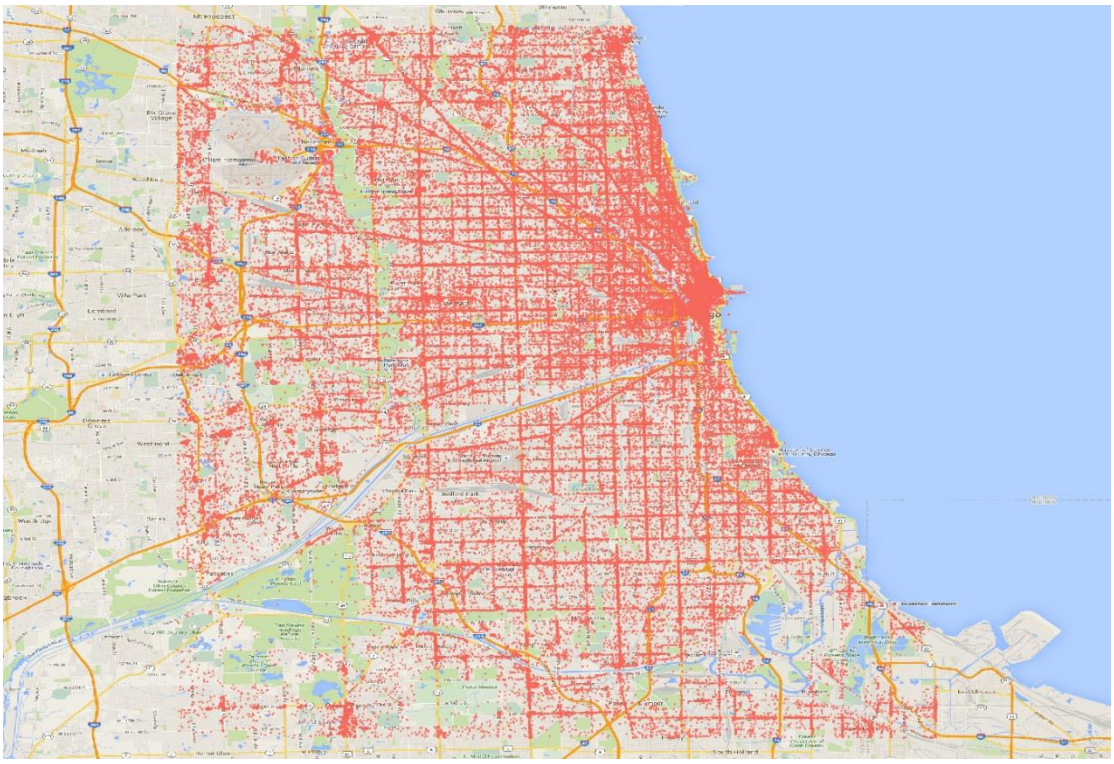
Πέρα από τον καθορισμό της περιοχής προς συλλογή σημείων ενδιαφέροντος, πρέπει να γίνει επίσης ο καθορισμός της συχνότητας συλλογής. Θέσαμε και για τα δύο σενάρια τη συχνότητα συλλογής στις 60 μέρες. Οι υπόλοιπες παράμετροι εξαρτώνται από το κάθε σενάριο και φαίνονται παρακάτω.

1.9.1 Συλλογή σημείων ενδιαφέροντος από Google Places

Στην τυπική εγκατάσταση του AreaProfiler, ο crawler για το GooglePlaces έχει το αναγνωριστικό "2". Στον ορισμό του campaign θα πρέπει να δώσουμε αυτήν την τιμή στο "poiCrawlers", το πεδίο που καθορίζει ποιοι crawlers θα χρησιμοποιηθούν στο νέο campaign. Επίσης, για να ζητήσουμε όλα ανεξαιρέτως τα σημεία ενδιαφέροντος που παρέχει το GooglePlaces, θα πρέπει να δώσουμε ως κατηγορία ("categoryName") το "google_places__all_place_types". Το αίτημα που θα πρέπει να αποσταλεί είναι το παρακάτω.

```
{
"request":{
"type":"startCampaign",
"name":"Chicago POIs from Google Places",
"description":"Chicago POIs from Google Places.",
"endTime":1435669471000,
"poiCrawlers":[
2
],
"categories":[
{
"categoryName":"google_places__all_place_types",
"crawlingPeriodInDays":60
}
],
"location":{
"type":"rectangleLocation",
"southWest":{
"latitude":41.627744,
"longitude":-87.949573
},
"northEast":{
"latitude":42.033020,
"longitude":-87.498048
}}}}
```

Μετά από 48 ώρες η συλλογή έχει ολοκληρωθεί και στη βάση έχουν αποθηκευτεί 184.392 σημεία ενδιαφέροντος τα οποία συλλέχθηκαν στο παραπάνω campaign.Στην παρακάτω εικόνα απεικονίζονται τα GooglePlacesπου έχουν συλλεχθεί για την ευρύτερη περιοχή του Chicago.



Εικόνα 14. Google Places στην ευρύτερη περιοχή του Chicago

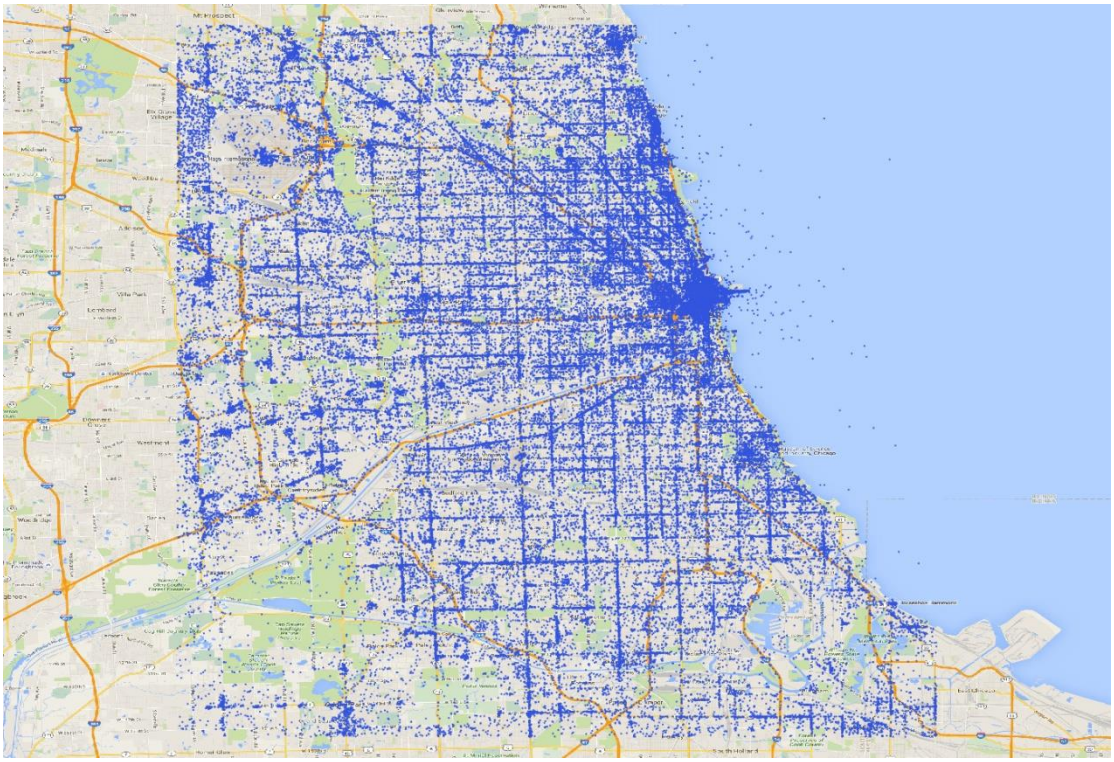
1.9.2 Συλλογή σημείων ενδιαφέροντος από Foursquare

Στην τυπική εγκατάσταση του AreaProfiler, ο crawler για το Foursquare έχει το αναγνωριστικό "1". Στον ορισμό του campaign θα πρέπει να δώσουμε αυτήν την τιμή στο "poiCrawlers", το πεδίο που καθορίζει ποιοι crawlers θα χρησιμοποιηθούν στο νέο campaign. Επίσης, για να αναζητήσουμε όλα ανεξαιρέτως τα σημεία ενδιαφέροντος που παρέχει το Foursquare, θα πρέπει να δώσουμε ως κατηγορία ("categoryName") το "foursquare__all_venue_categories". Το αίτημα που θα πρέπει να αποσταλεί είναι το παρακάτω.

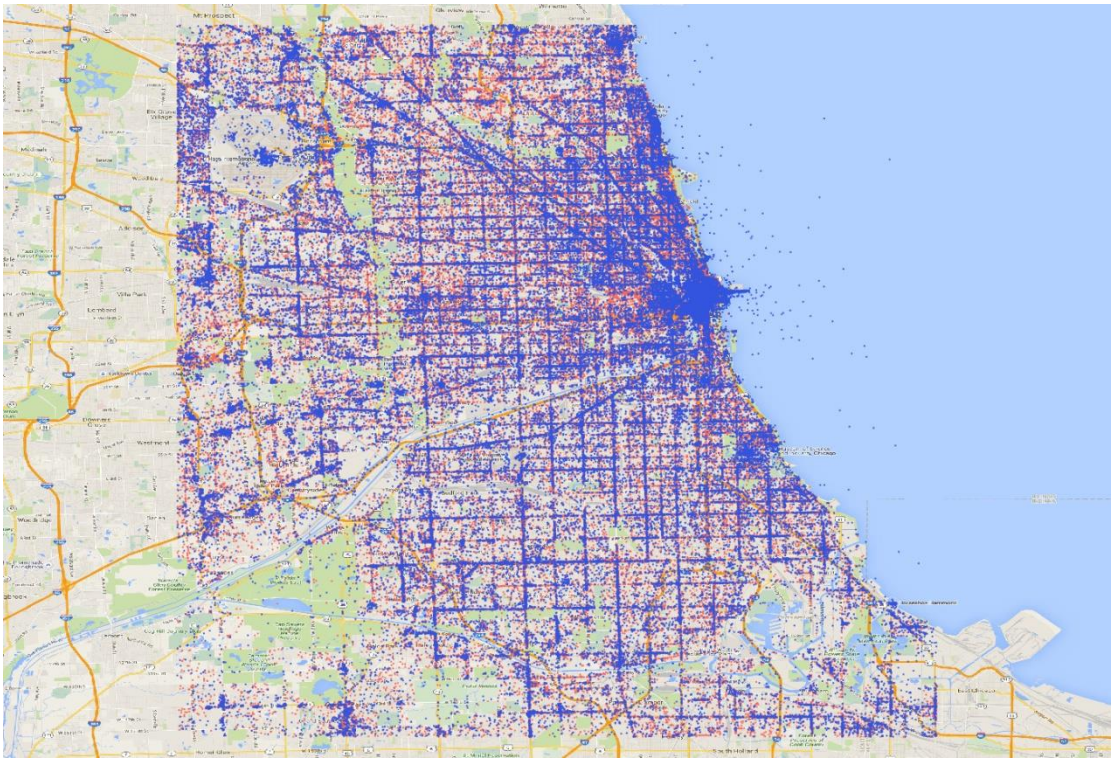
```
{
"request":{
"type":"startCampaign",
"name":"Chicago POIs from Foursquare",
"description":"Chicago POIs from Foursquare.",
"endTime":1435669471000,
"poiCrawlers":[
1
],
"categories":[
{
"categoryName":"foursquare__all_venue_categories",
"crawlingPeriodInDays":60
}
],
"location":{
"type":"rectangleLocation",
"southWest":{
"latitude":41.627744,
```

```
"longitude":-87.949573
},
"northEast":{
"latitude":42.033020,
"longitude":-87.498048
}}}}
```


Μετά από 48 ώρες, η συλλογή των σημείων ενδιαφέροντος έχει ολοκληρωθεί και στη βάση έχουν αποθηκευτεί 93.893 σημεία ενδιαφέροντος τα οποία συλλέχθηκαν για λογαριασμό της παραπάνω campaign.Στις παρακάτω εικόνες απεικονίζονται α) τα Foursquarevenuesπου έχουν συλλεχθεί για την περιοχή τουChicago β) Τα Foursquarevenues σε σύγκριση με τα αντίστοιχα GooglePlaces για την ευρύτερη περιοχή του Chicago.



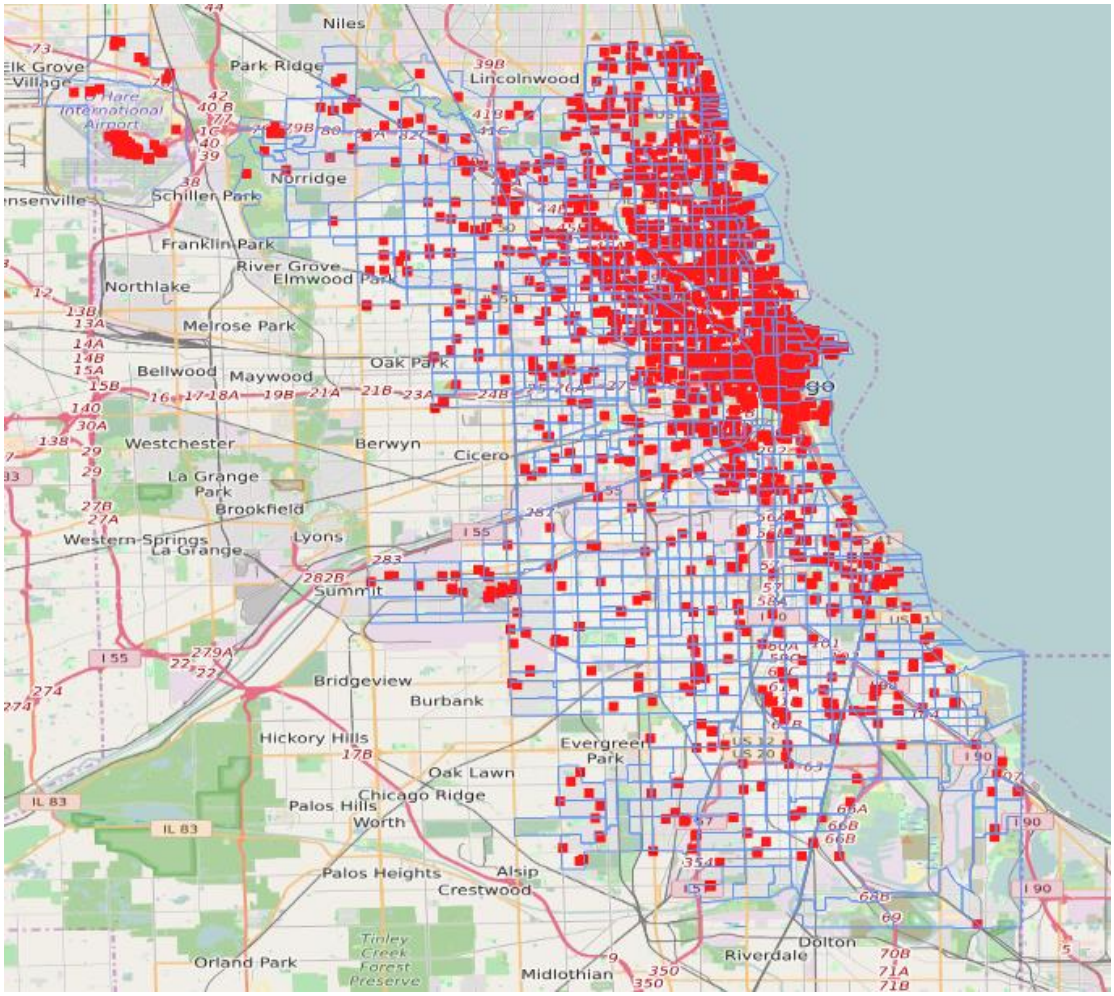
Εικόνα 15. Foursquarevenues στην ευρύτερη περιοχή του Chicago



Εικόνα 169. Σημεία ενδιαφέροντος από GooglePlaces και Foursquare στην ευρύτερη περιοχή του Chicago

1.9.3 Συλλογήstreaming social media data

Στο σενάριο αυτό στόχος είναι να συλλεχθούν γεωχωρικά προσδιορισμένα tweetsκαι checkinsγια έναχρονικό διάστημα δύο ημερών. OAreaProfilerχειρίζεται τη συλλογή streamingδεδομένων από τα κοινωνικά δίκτυα με διαφορετικό τρόπο από ότι για τα σημεία ενδιαφέροντος και συγκεκριμένα με τη χρήση του TwitterStreamCrawler και του Check-In Details Crawler σε κυλιόμενο παράθυρο χρόνου. Στο σενάριο αυτό χρησιμοποιούμε παράθυρο χρόνου δώδεκα ωρών και ως boudingboxτο ορθογώνιο που ορίστηκε στην αρχή της ενότητας και ορίζει την περιοχή του Σικάγο. Μετά από 48 ώρες, η συλλογή των δεδομένων από το Twitterέχει ολοκληρωθεί και έχουν αποθηκευτεί7750γεωχωρικά tweets, τα οποία περιείχαν 1256 μοναδικές αναφορές σε hashtagsκαι 747checkinsσε σημεία ενδιαφέροντος.Στην παρακάτω εικόνα φαίνονται τα γεωχωρικά tweets που συλλέχθηκαν στο σενάριο αυτό.



Εικόνα 1710. Σημεία δημοσίευσης γεωχωρικών tweets

1.10 Σύνοψη

Στην παρούσα ενότητα, περιγράψαμε το εργαλείο AreaProfiler. Κύριος στόχος του συγκεκριμένου εργαλείου είναι η ανάκτηση, συλλογή και αποθήκευση δεδομένων για σημεία ενδιαφέροντος (POIs) μιας περιοχής, που προέρχονται από διαδικτυακά APIs και γεωχωρικών tweetsαπό το Twitter. Η υλοποίηση του Area Profiler λαμβάνει δεδομένα από τα Google Places, Foursquare και TwitterStreamingAPIs.Παρά τις πολύ σημαντικές λειτουργίες του AreaProfiler που ήδη έχουν ολοκληρωθεί, στόχος μας είναι να εμπλουτίσουμε ακόμα περισσότερο το συγκεκριμένο εργαλείο, ώστε να μπορεί να συλλέγει δεδομένα, κυρίως πολυμεσικού περιεχομένου (φωτογραφίες) από επιπλέον πηγές, όπως το Flickr ή το GoogleStreetViewAPIs.

2 Ανοικτά δεδομένα

Στην προηγούμενη ενότητα περιγράψαμε συνοπτικά τον Areaprofiler και πως χρησιμοποιώντας το συγκεκριμένο εργαλείο, μπορέσαμε να συλλέξουμε δεδομένα για σημεία ενδιαφέροντος από τα GooglePlaces και FoursquareAPIs.Στηνπαρούσα ενότητα θα περιγράψουμε τα ανοικτά δεδομένα που θα χρησιμοποιήσουμε για την περιοχή-πιλότο (Chicago, USA). Πολλά από τα δεδομένα που θα χρησιμοποιήσουμε στο Citysense προέρχονται από το επίσημο portal της πόλης του Chicago[18]. Το συγκεκριμένο portal "θέλει να συμβάλει στην προώθηση της πρόσβασης σε κυβερνητικά δεδομένα και να ενθαρρύνει την ανάπτυξη δημιουργικών εργαλείων για την καλύτερη εξυπηρέτηση της κοινότητας του Σικάγο"[19] και φιλοξενεί πάνω από 200 σύνολα δεδομένων σε διάφορες μορφές φιλικές προς το χρήστη, όπως CSV και shapefiles (για χωρικά δεδομένα). Τα δεδομένα αυτά θα παρουσιαστούν συνοπτικά στις επόμενες ενότητες.

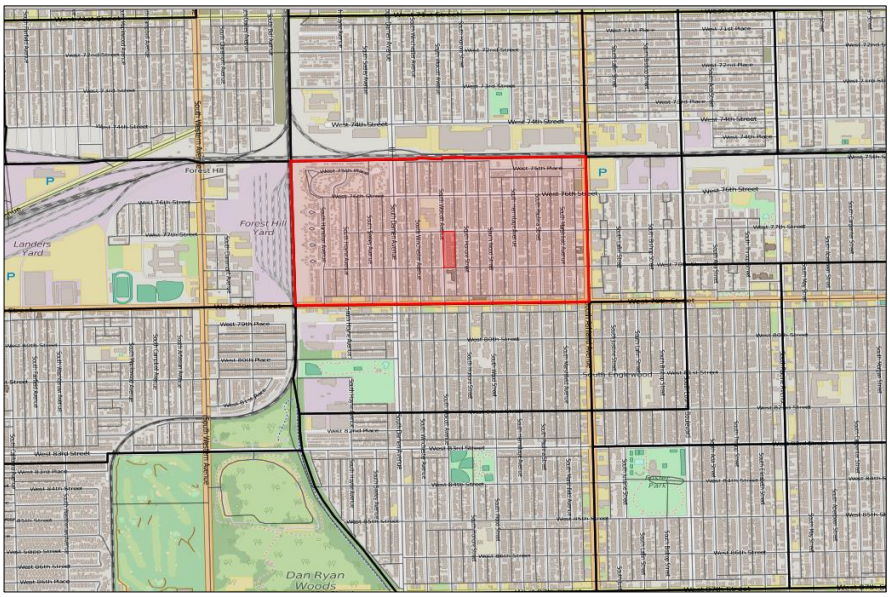
2.1 Χωροθέτηση των δεδομένων

Πολλά από τα ανοικτά δεδομένα που θα χρησιμοποιήσουμε έχουν χωρικές συντεταγμένες. Προκειμένου τα συγκεκριμένα δεδομένα να οπτικοποιηθούν σε ένα χάρτη θα πρέπει να συναθροιστούν σε χωρικό επίπεδο, δηλαδή να ομαδοποιηθούν σε ευρύτερες περιοχές. Αντί να ορίσουμε κάποιες αυθαίρετες περιοχές, είναι προτιμότερο να χρησιμοποιήσουμε υπάρχουσες διοικητικές υποδιαιρέσεις που ήδη χρησιμοποιούνται από τις κρατικές υπηρεσίες της πόλης του Chicago. Οιπροτεινόμενες υποδιαιρέσειςείναι: (i) CensusBlocks, (ii) CensusTractsκαι (iii) CommunityAreas. Ο λόγος που επιλέξαμε αυτές τις συγκεκριμένες διοικητικές υποδιαιρέσεις είναι ότι για τις δύο πρώτες υπάρχουν τα δεδομένα της τελευταίας απογραφής που πραγματοποιήθηκε στο Chicago το 2010, ενώ για τα CommunityAreas υπάρχουν σημαντικό πλήθος κοινωνικο-οικονομικών δεδομένων, καθώς και δεδομένα που αφορούν δείκτες υγείας. Στοιχεία για τις τρεις αυτές διοικητικές υποδιαιρέσεις παρέχονται παρακάτω, σύμφωνα με τη Βιβλιοθήκη του Chicago[20].

2.1.1 Census blocks

Η περιοχή του Chicago περιλαμβάνει συνολικά 46.311 CensusBlocks. Κάθε CensusBlock αντιστοιχεί πρακτικά σε ένα οικοδομικό τετράγωνο. Πολλά από τα δεδομένα απογραφής παρέχονται σε επίπεδο censusblock για λόγους ανωνυμοποίησης.

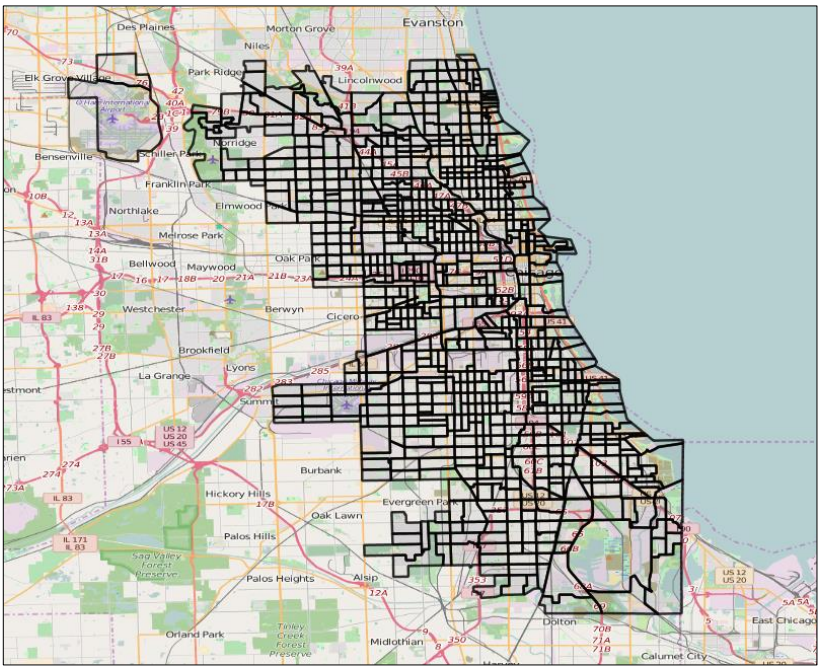
[18]https://data.cityofchicago.org/
[19]http://www.cityofchicago.org/city/en/narr/foia/CityData.html
[20]http://www.lib.uchicago.edu/e/collections/maps.moved/censusinfo.html



Εικόνα118. Census Blocks και census tracts

2.1.2 Census Tracts

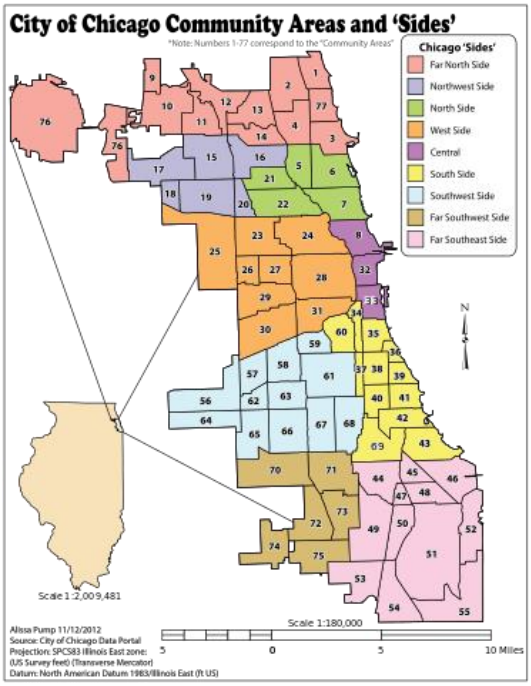
Η περιοχή του Chicago περιλαμβάνεισυνολικά 801 CensusTracts. Κάθε CensusTract περιγράφει μια μικρή περιοχή που θεωρείται πως είναι σχετικά ομοιογενήςκαι αντιστοιχεί ιδανικά σε περίπου 1200 νοικοκυριά (2000-4000 άτομα). Στο Σικάγο, ο πληθυσμός για κάθε CensusTract κυμαίνεται από 0 έως 10.000 και τα αντίστοιχα όρια έχουν παραμένει σχεδόν σταθερά από το 1920 (με εξαίρεση κυρίως στα προάστια), με αλλαγές όμως στο αντίστοιχο σύστημα αρίθμησης. ΚάθεCensus Tractμπορείναπεριλαμβάνειαπό 1-250 CensusBlocks.



Εικόνα129. Census tracts of Chicago

2.1.3 Community areas

Υπάρχουν 77 συνολικά CommunityAreas στο Chicago. Κάθε CommunityArea ιδανικάαντιστοιχεί σε μια γειτονιά (π.χ. Hyde Park και Uptown) που αναγνωρίζεται από τους κατοίκους της.Παρόλο που τα όρια των CommunityAreas δεν συμπίπτουν με τα όρια των censustracts, κάθε censustract θεωρείται πως ανήκει σε ένα μόνο CommunityArea.



Εικόνα 20. Chicago community areas. By Alissapump - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=22655083

2.2 Δεδομένα απογραφής

Στην συγκεκριμένη κατηγορία δεδομένων ανήκουν τα δεδομένα που προέρχονται κυρίως από την απογραφή του 2010. Προφανώς για κάθε CensusBlock είναι γνωστός ο αριθμός των αντίστοιχων κατοίκων. Υπάρχουν όμως και δεδομένα που αφορούν τα CommunityAreas και αφορούν κοινωνικό-οικονομικούς δείκτες και δείκτες υγείας.

2.2.1 Κοινωνικο-οικονομικοί δείκτες

Σε επίπεδο CommunityAreas υπάρχουν διαθέσιμοι συνολικά επτά κοινωνικό-οικονομικοί δείκτες[21] για τα έτη 2008 – 2012. Αυτοί οι δείκτες είναι:

- Ποσοστό κατοικιών που είναι γεμάτο
- Ποσοστό νοικοκυριών σε συνθήκες φτώχειας
- Ποσοστό ανέργων σε ηλικίες 16 και πάνω
- Ποσοστό ανθρώπων ηλικίας άνω των 25 χωρίς απολυτήριο λυκείου
- Ποσοστό ανθρώπων ηλικίας κάτω των 18 ή άνω των 64
- Κατά κεφαλήν εισόδημα
- Δείκτης κακουχίας (Hardship Index)

[21] https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

Οι πέντε πρώτοι δείκτες είναι σε επίπεδο ποσοστού.

	A	B	C	D	E	F	G	H	I
1	Community Area Number	COMMUNITY AREA NAME	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER CAPITA INCOME	HARDSHIP INDEX
2	1	Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39
3	2	West Ridge	7.8	17.2	8.8	20.8	38.5	23040	46
4	3	Uptown	3.8	24	8.9	11.8	22.2	35787	20
5	4	Lincoln Square	3.4	10.9	8.2	13.4	25.5	37524	17
6	5	North Center	0.3	7.5	5.2	4.5	26.2	57123	6
7	6	Lake View	1.1	11.4	4.7	2.6	17	60058	5
8	7	Lincoln Park	0.8	12.3	5.1	3.6	21.5	71551	2
9	8	Near North Side	1.9	12.9	7	2.5	22.6	88669	1
10	9	Edison Park	1.1	3.3	6.5	7.4	35.3	40959	8
11	10	Norwood Park	2	5.4	9	11.5	39.5	32875	21
12	11	Jefferson Park	2.7	8.6	12.4	13.4	35.5	27751	25
13	12	Forest Glen	1.1	7.5	6.8	4.9	40.5	44164	11
14	13	North Park	3.9	13.2	9.9	14.4	39	26576	33
15	14	Albany Park	11.3	19.2	10	32.9	32	21323	53
16	15	Portage Park	4.1	11.6	12.6	19.3	34	24336	35
17	16	Irving Park	6.3	13.1	10	22.4	31.6	27249	34
18	17	Dunning	5.2	10.6	10	16.2	33.6	26282	28
19	18	Montclaire	8.1	15.3	13.8	23.5	38.6	22014	50
20	19	Belmont Cragin	10.8	18.7	14.6	37.3	37.3	15461	70
21	20	Hermosa	6.9	20.5	13.1	41.6	36.4	15089	71
22	21	Avondale	6	15.3	9.2	24.7	31	20039	42
23	22	Logan Square	3.2	16.8	8.2	14.8	26.2	31908	23
24	23	Humboldt park	14.8	33.9	17.3	35.4	38	13781	85
25	24	West Town	2.3	14.7	6.6	12.9	21.7	43198	10
26	25	Austin	6.3	28.6	22.6	24.4	37.9	15957	73
27	26	West Garfield Park	9.4	41.7	25.8	24.5	43.6	10934	92
28	27	East Garfield Park	8.2	42.4	19.6	21.3	43.2	12961	83
29	28	Near West Side	3.8	20.6	10.7	9.6	22.2	44689	15
30	29	North Lawndale	7.4	43.1	21.2	27.6	42.7	12034	87
31	30	South Lawndale	15.2	30.7	15.8	54.8	33.8	10402	96
32	31	Lower West Side	9.6	25.8	15.8	40.7	32.6	16444	76
33	32	Loop	1.5	14.7	5.7	3.1	13.5	65526	3
34	33	Near South Side	1.3	13.8	4.9	7.4	21.8	59077	7
35	34	Armour Square	5.7	40.1	16.7	34.5	38.3	16148	82
36	35	Douglas	1.8	29.6	18.2	14.3	30.7	23791	47
37	36	Oakland	1.3	39.7	28.7	18.4	40.4	19252	78

Εικόνα213. Sample socioeconomic indicators for Chicago (CSV)

Στη συνέχεια παραθέτουμε ένα πίνακα με τις παρατηρούμενες τιμές για τα συγκεκριμένα μεγέθη (MIN, AVG, MAX).

Μέγεθος	MIN	AVG	MAX
Ποσοστό κατοικιών που είναι γεμάτο	0.3	4.9	15.8
Ποσοστό νοικοκυριών σε συνθήκες φτώχειας	3.3	21.8	56.5
Ποσοστό ανέργων σε ηλικίες 16 και πάνω	4.7	15.4	35.9
Ποσοστό ανθρώπων ηλικίας άνω των 25 χωρίς απολυτήριο λυκείου	2.5	20.3	54.8
Ποσοστό ανθρώπων ηλικίας κάτω των 18 ή άνω των 64	13.5	35.7	51.5
Κατά κεφαλή εισόδημα	8201.0	25563.2	88669.0
Δείκτης κακουχίας (Hardship Index)	1.0	49.5	98.0

2.2.2 Δείκτες Υγείας

Σε επίπεδο CommunityAreas υπάρχουν διαθέσιμοι συνολικά είκοσι επτά δείκτες υγείας[22]με τελευταία ενημέρωση στις 4 Ιουνίου 2013.Λεπτομέρειες για το συγκεκριμένο datasetείναι διαθέσιμεςκαι στη διαδικτυακή διεύθυνση https://data.cityofchicago.org/api/assets/2107948F-357D-4ED7-ACC2-2E9266BBFFA2. Αυτοί οι δείκτες είναι:

Γεννήσεις

- Ποσοστό γεννήσεων
- Γενικό ποσοστό γονιμότητας
- Χαμηλό βάρος νεογνών
- Προγεννητική φροντίδα αρχής γενομένης από το πρώτο τρίμηνο
- Πρόωρες γεννήσεις
- Ποσοστό γεννήσεων από έφηβες

Θάνατοι:

- Επίθεση / Ανθρωποκτονία
- Καρκίνος του μαστού (γυναίκες)
- Καρκίνος
- Καρκίνος του παχέος εντέρου
- Που σχετίζονται με Διαβήτη,
- Από πυροβόλο όπλο,
- ΒρεφικήΘνησιμότητα
- Καρκίνος του πνεύμονα
- Καρκίνος του προστάτη (άνδρες),
- Εγκεφαλικό επεισόδιο - Εγκεφαλική αγγειακή νόσος,
- Μετρήσεις στο επίπεδο μολύβδου στο αίμα, σε παιδική ηλικία
- Δηλητηρίαση από μόλυβδο σε παιδική ηλικία
- Γονόρροια (γυναίκες)
- Γονόρροια (άνδρες)
- Φυματίωση

Οικονομικοί δείκτες

- Ποσοστό που ζει κάτω από το επίπεδο της φτώχειας
- Συνωστισμός κατοικίας
- Ποσοστό προστατευόμενων ατόμων
- Ποσοστό ανθρώπων χωρίς απολυτήριο λυκείου
- Κατά κεφαλήν εισόδημα
- Ποσοστό ανεργίας

[22] https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Community Area	Community Area Name	Birth Rate	General Fertility Rate	Low Birth Weight	Prenatal Care Beginning in First Trimester	Preterm Births	Teen Birth Rate	Assault (Homicide)	Breast cancer in females	Cancer (All Sites)	Colorectal Cancer	Diabetes-related	Firearm-related	Infant Mortality Rate	Lung Cancer	Prostate Cancer in Males
2	1	Rogers Park	16.4	62	11	73	11.2	40.8	7.7	23.3	176.9	25.3	77.1	5.2	6.4	36.7	21.7
3	2	West Ridge	17.3	83.3	8.1	71.1	8.3	29.9	5.8	20.2	155.9	17.3	60.5	3.7	5.1	36	14.2
4	3	Uptown	13.1	50.5	8.3	77.7	10.3	35.1	5.4	21.3	183.3	20.5	80	4.6	6.5	50.5	25.2
5	4	Lincoln Square	17.1	61	8.1	80.5	9.7	38.4	5	21.7	153.2	8.6	55.4	6.1	3.8	43.1	27.6
6	5	North Center	22.4	76.2	9.1	80.4	9.8	8.4	1	16.6	152.1	26.1	49.8	1	2.7	42.4	15.1
7	6	Lake View	13.5	38.7	6.3	79.1	8.1	15.8	1.4	20.1	126.9	13	38.5	1.8	2.2	32.5	17
8	7	Lincoln Park	13.2	38.7	6.6	75.7	7.8	2.1	0.7	23.7	152.9	16.7	50.1	2.3	2.4	40	27.3
9	8	Near North Side	10.7	35.9	8.6	69.7	9.6	34	3.7	24	142.7	15.1	27	3.2	6.5	33.6	15.1
10	9	Edison Park	11.3	59.5	7.9	86.6	12.6	3.9	0	13.8	189.7	15.1	53	7.1	4.6	45.2	28
11	10	Norwood Park	10.4	59.6	4.9	89.4	8.3	3.4	4.7	20.7	180.8	18.9	47.3	8.7	4.4	44.5	26.4
12	11	Jefferson Park	13.8	67.8	6.6	82.9	7.7	28.6	4.8	18.4	208.2	23.2	49.2	4.2	8.3	55.7	32.1
13	12	Forest Glen	10	60.6	7.6	79.3	10.3	6.3	3.3	25	138.7	14.3	37.2	6.2	3.8	27	20.3
14	13	North Park	10.9	54.2	9.7	79.1	10.2	10.5	3	20.4	143.7	21.9	58.9	3.1	5.4	34.7	14.6
15	14	Albany Park	18.3	76.5	8.5	73.3	8.3	44.5	4.7	22.9	158.1	16.8	72.1	5.3	4.9	36.9	13.1
16	15	Portage Park	14.2	66.1	6.9	79.8	8.7	41.7	3.3	23.3	168.7	15.9	48.2	4.7	4.7	44.9	14.8
17	16	Irving Park	15.8	67.1	7.7	79.9	10.2	37	4.1	29.9	169.4	19.2	60.2	5.7	5.3	41.3	17.9
18	17	Dunning	12.5	64.7	6.8	82.7	9.9	19.9	3.7	23.7	191.5	25.9	42.5	5.2	4.9	53.9	24.4
19	18	Montclaire	17.1	83.5	8.3	77.6	8.8	61.5	8.6	29.9	151	15.4	89.6	6.5	4.6	33.4	21.9
20	19	Belmont Cragin	20	88.6	6.9	74.1	7.6	68.2	7	14.4	152.6	17.7	58.6	5.5	5.6	37.8	27.3
21	20	Hermosa	20.3	86.7	6.7	77.5	8.8	69.7	12.7	18.4	135.2	15.6	63.6	11.8	9.3	27.7	25.6
22	21	Avondale	18.5	77.7	7.3	74.4	7.5	63.4	4.7	16.6	133.9	13.4	52.7	4.6	5.7	32.5	37.7
23	22	Logan Square	18.2	63.5	7.2	78.2	9.4	66.1	8.6	9.2	148.7	13.7	75.7	10.2	4.3	37.7	17.5
24	23	Humboldt Park	19.2	80.7	12.3	70.9	11.7	77.9	29	26.3	211.1	26.6	94.1	22.7	9.8	48	52.5
25	24	West Town	18.8	60.4	9.1	75.5	10.8	49.2	8.5	14.5	139.6	12.4	107	6.6	5.1	27.4	16.6
26	25	Austin	18	80.1	15.4	72.9	14.3	81.8	34.4	33.7	261.9	29.8	113.9	28.5	13.3	74.6	69.8
27	26	West Garfield Park	20.1	88.4	17	71.4	17.5	114.9	40	54.7	291.5	31.4	118.2	36	19	65.3	75
28	27	East Garfield Park	19.4	80.8	17.5	73.2	16.3	93.2	38.4	21.7	236.8	24.8	97.3	37.1	11	56.3	78.1
29	28	Near West Side	18.2	55.6	9	77.6	10.5	36.7	12.7	33.4	202	19.2	62.3	9.3	9.1	66.2	33.6
30	29	North Lawndale	20.6	86.3	15.3	75.8	15.2	108.9	46.7	45.8	261.5	34.8	99.2	37.6	14.1	61.7	54
31	30	South Lawndale	19.5	94.9	7.6	85.1	9.6	77.5	11.1	13.2	127.4	9.2	65	8.6	5.9	15.9	32.7
32	31	Lower West Side	16.5	68.2	4.5	80.8	8.8	49	11.7	27.2	141.3	11.9	61.9	11.7	5.4	27.8	14.3
33	32	Loop	9.4	27.7	5.3	78.2	6.9	1.3	0.7	20.2	120.1	10.8	26.8	4	5.7	29.2	17.2
34	33	Near South Side	21.4	72.9	8.8	78.1	10.9	50.9	4.8	31.9	169	19.2	61.5	6.6	4.8	46.2	51.4
35	34	Armour Square	11.5	57.1	12.4	79.1	11.8	16.2	1.8	10.7	162.9	23.1	42.5	1.8	1.5	54.3	17.2
36	35	Douglas	10.3	42.2	11.7	76	10.2	34.2	13.6	34.3	269.9	33.2	119.1	9.1	13.4	74.5	85.5

Εικόνα 22. Sample health indicators for Chicago (CSV)

Στη συνέχεια παραθέτουμε ένα πίνακα με τις παρατηρούμενες τιμές για τα συγκεκριμένα μεγέθη (MIN, AVG, MAX), καθώς και τις απαραίτητες επεξηγήσεις για κάθε μέγεθος, όπως ακριβώς παρέχονται από το portal και το αντίστοιχο διευκρινιστικό έγγραφο.

Μέγεθος	**MIN**	**AVG**	**MAX**	**Σχόλιο**
Ποσοστό γεννήσεων	9.4	15.7	22.4	Percent of persons in labor force aged 16 years and older
Γενικό ποσοστό γονιμότητας	27.7	68.4	94.9	Per 1,000 females aged 15-44
Χαμηλό βάρος νεογνών	3.5	10.1	19.7	Percent of live births
Προγεννητική φροντίδα αρχής γενομένης από το πρώτο τρίμηνο	63.6	77.0	94.5	Percent of females delivering a live birth
Πρόωρες γεννήσεις	5.0	11.3	17.5	Percent of live births
Ποσοστό γεννήσεων από έφηβες	1.3	50.1	116.9	Per 1,000 females aged 15-19
Επίθεση / Ανθρωποκτονία	0.0	18.1	70.3	Per 100,000 persons (age adjusted)
Καρκίνος του μαστού (γυναίκες)	7.6	26.0	54.7	Per 100,000 persons (age adjusted)
Καρκίνος	120.1	194.3	291.5	Per 100,000 persons (age adjusted)
Καρκίνος του παχέος εντέρου	8.6	21.6	39.4	Per 100,000 persons (age adjusted)
Που σχετίζονται με Διαβήτη,	26.8	71.9	119.1	Per 100,000 persons (age adjusted)
Από πυροβόλο όπλο,	1.0	16.7	70.3	Per 100,000 persons (age adjusted)
Βρεφική Θνησιμότητα	1.5	8.6	22.6	Per 1,000 live births

Μέγεθος	MIN	AVG	MAX	Σχόλιο
Καρκίνος του πνεύμονα	15.9	51.5	89.6	Per 100,000 persons (age adjusted)
Καρκίνος του προστάτη (άνδρες),	0.0	36.8	92.9	Per 100,000 males (age adjusted)
Εγκεφαλικό επεισόδιο - Εγκεφαλική αγγειακή νόσος,	22.0	46.5	99.1	Per 100,000 persons (age adjusted)
Μετρήσεις στο επίπεδο μολύβδου στο αίμα, σε παιδική ηλικία	133.6	385.3	605.9	Per 1,000 children aged 0-6 years
Δηλητηρίαση από μόλυβδο σε παιδική ηλικία	0.0	0.9	3.7	Per 100
Γονόρροια (γυναίκες)	50.3	894.3	3193.3	Per 100,000 females aged 15 to 44 years
Γονόρροια (άνδρες)	52.7	840.7	2545.7	Per 100,000 males aged 15 to 44 years
Φυματίωση	0.0	6.8	22.7	Per 100,000 persons (age adjusted)
Ποσοστό νοικοκυριών που ζει κάτω από το επίπεδο της φτώχειας	3.1	20.3	61.4	
Συνωστισμός κατοικίας	0.2	4.9	17.6	Percent of occupied housing units
Ποσοστό προστατευόμενων ατόμων	15.5	35.8	50.2	Percent of persons aged less than 16 or more than 64 years
Ποσοστό ανθρώπων χωρίς απολυτήριο λυκείου	2.9	21.6	58.7	Percent of persons aged 25 years and older
Κατά κεφαλήν εισόδημα	8535.0	25106.7	87163.0	2011 inflation-adjusted dollars
Ποσοστό ανεργίας	4.2	13.3	40.0	Percent of persons in labor force aged 16 years and older

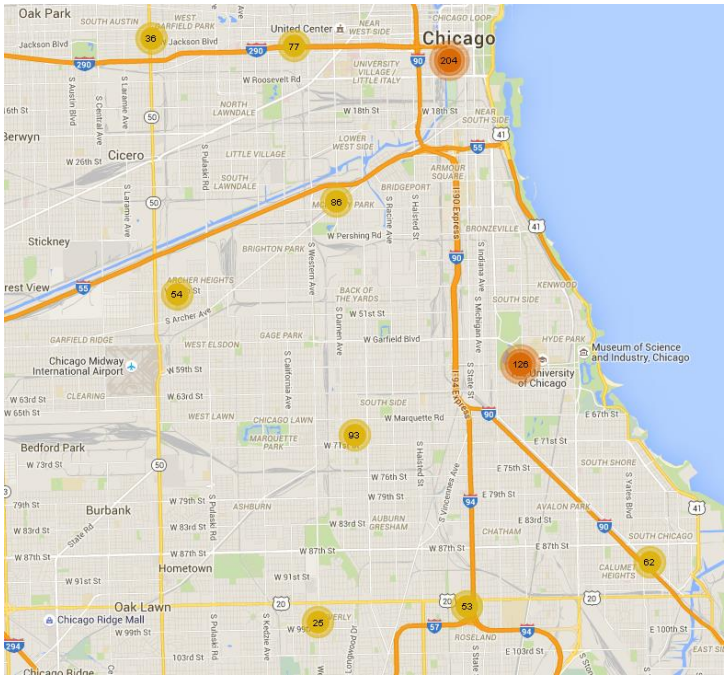
2.2.3 Δεδομένα κίνησης

Για το Chicagoκαι επιλεγμένες τοποθεσίες, υπάρχουν δεδομένα για την μέση ημερήσια κυκλοφορία, δηλαδή τον πραγματικό αριθμό των οχημάτων που διέρχονται από μια δεδομένη οδό κατά μέσο όρο μια ημέρα της εβδομάδας. Τα δεδομένα αυτά δεν ανανεώνονται πολύ τακτικά (κάθε δεκαετία), αλλά μπορούν να χρησιμοποιηθούν για να έχουμε μια πρόχειρη εκτίμηση της κυκλοφορικής κίνησης στους κυριότερους δρόμους της πόλης. Συνολικά, υπάρχουν διαθέσιμες μετρήσεις για 1279 τοποθεσίες. Τα συγκεκριμένα δεδομένα έχουν τα παρακάτω στοιχεία, όπως εμφανίζονται στην επόμενη εικόνα.

	ID	Traffic Volume Count Location Address	Street	Date of Count	Total Passing Vehicle Volume	Vehicle Volume By Each Direction of Traffic	Latitude	Longitude
2	1	2523 West	71st Street	02/28/2006	14600	East Bound: 7800 / West Bound: 6800	41.764641	-87.686772
3	2	1708 West	71st Street	03/09/2006	14600	East Bound: 6900 / West Bound: 7700	41.764877	-87.666635
4	3	1275 West	71st Street	02/28/2006	16500	East Bound: 7800 / West Bound: 8700	41.765008	-87.657067
5	5	920 West	71st Street	02/28/2006	18200	East Bound: 8800 / West Bound: 9400	41.765153	-87.647751
6	6	758 West	71st Street	02/28/2006	21600	East Bound: 10400 / West Bound: 11200	41.765204	-87.644371
7	7	240 East	71st St	02/28/2006	18300	East Bound: 9000 / West Bound: 9300	41.765644	-87.618476
8	9	2050 East	71st St	03/09/2006	8600	East Bound: 3600 / West Bound: 5000	41.76626	-87.574226
9	10	8539 South	Commercial Ave	03/07/2006	10000	North Bound: 5000 / South Bound: 5000	41.739836	-87.551476
10	11	8933 South	Commercial Ave	03/07/2006	10500	North Bound: 4900 / South Bound: 5600	41.732725	-87.55132
11	12	9379 South	Commercial Ave	03/07/2006	12700	North Bound: 6000 / South Bound: 6700	41.724444	-87.551124
12	13	9730 South	Commercial Ave	03/07/2006	9000	North Bound: 4000 / South Bound: 5000	41.718183	-87.551016
13	14	4107 South	Cottage Grove Ave	03/30/2006	10800	North Bound: 5600 / South Bound: 5200	41.820171	-87.606798
14	15	4750 South	Cottage Grove Ave	03/01/2006	14200	North Bound: 7000 / South Bound: 7200	41.807987	-87.606532
15	16	5325 South	Cottage Grove Ave	02/01/2006	13700	North Bound: 7200 / South Bound: 6500	41.797881	-87.606302
16	17	6144 South	Cottage Grove Ave	03/01/2006	20900	North Bound: 10300 / South Bound: 10600	41.78269	-87.605979
17	18	6533 South	Cottage Grove Ave	03/01/2006	19100	North Bound: 9800 / South Bound: 9300	41.775779	-87.605826
18	19	6820 South	Cottage Grove Ave	02/01/2006	21700	North Bound: 10300 / South Bound: 11400	41.770685	-87.605711
19	20	7346 South	Cottage Grove Ave	03/01/2006	17800	North Bound: 8700 / South Bound: 9100	41.760801	-87.605452
20	21	8800 South	Cottage Grove Ave	03/01/2006	22400	North Bound: 11100 / South Bound: 11300	41.734791	-87.604779
21	23	10101 South	Cottage Grove Ave	03/02/2006	12500	North Bound: 5800 / South Bound: 6700	41.710922	-87.60573
22	24	1603 South	Damen Ave	03/02/2006	18800	North Bound: 9600 / South Bound: 9200	41.859378	-87.676043
23	25	1959 South	Damen Ave	03/02/2006	15300	North Bound: 7700 / South Bound: 7600	41.854926	-87.67592
24	26	2259 South	Damen Ave	03/02/2006	14800	North Bound: 8000 / South Bound: 6800	41.850395	-87.67579
25	27	2566 South	Damen Ave	05/09/2006	28000	North Bound: 13800 / South Bound: 14200	41.844875	-87.675652
26	28	3460 South	Damen Ave	03/02/2006	8100	North Bound: 3800 / South Bound: 4300	41.830389	-87.675279
27	29	3630 South	Damen Ave	03/01/2006	7000	North Bound: 3200 / South Bound: 3800	41.827625	-87.675208
28	31	4936 South	Damen Ave	03/09/2006	10900	North Bound: 5900 / South Bound: 5000	41.803741	-87.674569
29	32	5929 South	Damen Ave	03/09/2006	14900	North Bound: 7600 / South Bound: 7300	41.785746	-87.674075
30	33	6755 South	Damen Ave	03/02/2006	16600	North Bound: 7900 / South Bound: 8700	41.770409	-87.673671
31	34	7509 South	Damen Ave	03/02/2006	12300	North Bound: 5600 / South Bound: 6700	41.757363	-87.673336
32	36	3030 South	Dr Martin Luther King Jr Dr	03/14/2006	19900	North Bound: 10000 / South Bound: 9900	41.838419	-87.617465
33	37	3748 South	Dr Martin Luther King Jr Dr	03/02/2006	15600	North Bound: 7300 / South Bound: 8300	41.825982	-87.617039
34	38	4358 South	Dr Martin Luther King Jr Dr	03/02/2006	14500	North Bound: 6600 / South Bound: 7900	41.814852	-87.616741
35	39	5450 South	Dr Martin Luther King Jr Dr	02/28/2006	11800	North Bound: 5600 / South Bound: 6200	41.795094	-87.615989
36	41	7718 South	Dr Martin Luther King Jr Dr	02/28/2006	16700	North Bound: 7900 / South Bound: 8800	41.754187	-87.615027
37	42	9243 South	Dr Martin Luther King Jr Dr	02/28/2006	15500	North Bound: 7800 / South Bound: 7700	41.726239	-87.614254
38	43	10326 South	Dr Martin Luther King Jr Dr	02/28/2006	11100	North Bound: 5300 / South Bound: 5800	41.706494	-87.613745

Εικόνα23. Chicago Traffic Data (CSV)

Ένα μικρό δείγμα αυτών των δεδομένων οπτικοποιημένα σε ένα χάρτη, φαίνεται στην παρακάτω εικόνα.



Εικόνα144. Chicago traffic data (Χάρτης)

2.3 Δεδομένα εγκλημάτων

Για την περιοχή του Chicago, υπάρχουν διαθέσιμα δεδομένα[23] σε μορφή CSV για τα εγκλήματα που πραγματοποιήθηκαν από το 2001 και μετά. Τα κυριότερα στοιχεία για τα παραπάνω εγκλήματα είναι:

- ID
- Αριθμός υπόθεσης
- Ημερομηνία / Ώρα
- Οικοδομικό τεράγωνο
- Έγκλημα
- Περιγραφή εγκλήματος
- Περιγραφήτοποθεσίας (π.χ., APARTMENT, STREET, ABANDONED BUILDING, RESIDENCE)
- Υπήρχε σύλληψη (Boolean)
- Οικιακόέγκλημα (Boolean)
- District
- Ward
- CommunityArea
- ΚωδικόςFBI
- Τελευταία ενημέρωση
- Γεωγραφικό πλάτος
- Γεωγραφικό μήκος

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	case number	crime date	block	iucr	primary type	description	location description	arrest	domestic	beat	district	ward	community area
2	3244819	HK265627	11/01/2004 17:00	061XX N KENMORE AVE	840	THEFT	FINANCIAL ID THEFT: OVER $300	APARTMENT	f	f	2433	24	49	77
3	3129898	HK119520	11/01/2004 17:00	122XX S WENTWORTH AVE	810	THEFT	OVER $500	RESIDENCE	f	f	523	5	9	53
4	3135639	HK125026	11/01/2004 17:00	031XX S THROOP ST	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET	f	f	924	9	11	60
5	3143403	HK133660	11/01/2004 17:00	058XX W MADISON ST	560	ASSAULT	SIMPLE	STREET	f	t	1513	15	29	25
6	3153482	HK144990	11/01/2004 17:00	027XX W 62ND ST	820	THEFT	$500 AND UNDER	DRIVEWAY - RESIDENTIAL	f	f	825	8	15	66
7	3130974	HK121561	11/01/2004 17:00	038XX W 63RD ST	1305	CRIMINAL DAMAGE	CRIMINAL DEFACEMENT	RESIDENCE	f	f	823	8	13	65
8	3129048	HK119567	11/01/2004 17:00	059XX S RICHMOND ST	2825	OTHER OFFENSE	HARASSMENT BY TELEPHONE	RESIDENCE	f	t	824	8	16	66
9	3129408	HK120235	11/01/2004 17:00	073XX S ROCKWELL ST	5002	OTHER OFFENSE	OTHER VEHICLE OFFENSE	STREET	f	f	835	8	18	66
10	3133097	HK122636	11/01/2004 17:00	026XX S WABASH AVE	910	MOTOR VEHICLE THEFT	AUTOMOBILE	STREET	f	f	2113	1	2	35
11	3133459	HK124050	11/01/2004 17:00	024XX N LAMON AVE	930	MOTOR VEHICLE THEFT	THEFT/RECOVERY: AUTOMOBILE	STREET	f	f	2521	25	31	19
12	3130590	HK120883	11/01/2004 17:00	047XX N WINTHROP AVE	1150	DECEPTIVE PRACTICE	CREDIT CARD FRAUD	RESIDENCE	f	f	2312	19	46	3
13	3142944	HK119467	11/01/2004 16:59	078XX S GREENWOOD AVE	430	BATTERY	AGGRAVATED: OTHER DANG WEAPON	STREET	t	f	624	6	8	69
14	3129241	HK119468	11/01/2004 16:55	037XX N LINCOLN AVE	1120	DECEPTIVE PRACTICE	FORGERY	SMALL RETAIL STORE	t	f	1923	19	47	5
15	3129617	HK119463	11/01/2004 16:55	034XX S DR MARTIN LUTHER KING JR DR	1330	CRIMINAL TRESPASS	TO LAND	DRUG STORE	t	f	2122	2	4	35
16	3129519	HK119461	11/01/2004 16:54	047XX N HARDING AVE	4387	OTHER OFFENSE	VIOLATE ORDER OF PROTECTION	STREET	t	t	1723	17	39	14
17	3137919	HK127218	11/01/2004 16:48	022XX N MILWAUKEE AVE	820	THEFT	$500 AND UNDER	PARKING LOT/GARAGE(NON.RESID.)	f	f	1431	14	1	22
18	3174711	HK119475	11/01/2004 16:45	049XX W SUPERIOR ST	2024	NARCOTICS	POSS: HEROIN(WHITE)	OTHER	t	f	1531	15	37	25
19	3129808	HK119464	11/01/2004 16:30	093XX S DR MARTIN LUTHER KING JR DR	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE	f	t	633	6	9	49
20	3132585	HK120452	11/01/2004 16:30	032XX W ROOSEVELT RD	820	THEFT	$500 AND UNDER	OTHER	f	f	1134	11	24	29
21	3172898	HK119437	11/01/2004 16:30	009XX N HUDSON AVE	1350	CRIMINAL TRESPASS	TO STATE SUP LAND	CHA PARKING LOT/GROUNDS	t	f	1823	18	27	8
22	3131748	HK120430	11/01/2004 16:30	032XX S ARCHER AVE	610	BURGLARY	FORCIBLE ENTRY	OTHER	f	f	922	9	11	59
23	3129594	HK120263	11/01/2004 16:30	045XX S RICHMOND ST	610	BURGLARY	FORCIBLE ENTRY	RESIDENCE-GARAGE	f	f	912	9	14	58
24	3129310	HK119449	11/01/2004 16:30	114XX S HALSTED ST	560	ASSAULT	SIMPLE	RESTAURANT	f	f	2233	22	34	49
25	3130711	HK120414	11/01/2004 16:30	018XX N TALMAN AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET	f	f	1421	14	1	22
26	3130540	HK119584	11/01/2004 16:30	022XX S WHIPPLE ST	560	ASSAULT	SIMPLE	SIDEWALK	f	f	1033	10	12	30
27	3130892	HK119420	11/01/2004 16:30	057XX S WOOD ST	460	BATTERY	SIMPLE	STREET	f	f	715	7	15	67
28	3211562	HK221975	11/01/2004 16:25	100XX W OHARE ST	2890	PUBLIC PEACE VIOLATION	OTHER VIOLATION	AIRPORT/AIRCRAFT	t	f	1651	16	41	76
29	3130213	HK119428	11/01/2004 16:25	051XX W WELLINGTON AVE	560	ASSAULT	SIMPLE	SIDEWALK	f	f	2521	25	31	19
30	3136108	HK119378	11/01/2004 16:22	003XX W ADAMS ST	460	BATTERY	SIMPLE	COMMERCIAL / BUSINESS OFFICE	t	f	112	1	2	32
31	3137540	HK119429	11/01/2004 16:20	0000X N STATE ST	860	THEFT	RETAIL THEFT	DEPARTMENT STORE	t	f	122	1	42	32
32	3129563	HK119394	11/01/2004 16:15	001XX N PARKSIDE AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	STREET	f	t	1512	15	29	25
33	3131763	HK119435	11/01/2004 16:15	0000X E WALTON ST	870	THEFT	POCKET-PICKING	DRUG STORE	t	f	1833	18	42	8
34	3132075	HK119423	11/01/2004 16:15	011XX N LARRABEE ST	1350	CRIMINAL TRESPASS	TO STATE SUP LAND	CHA PARKING LOT/GROUNDS	t	f	1823	18	27	8
35	3130224	HK119404	11/01/2004 16:10	040XX W GRAND AVE	1330	CRIMINAL TRESPASS	TO LAND	RESTAURANT	f	f	2534	25	30	23
36	3129179	HK119453	11/01/2004 16:10	002XX N CALIFORNIA AVE	460	BATTERY	SIMPLE	STREET	f	f	1331	12	2	27
37	3130062	HK119387	11/01/2004 16:06	005XX W MADISON ST	860	THEFT	RETAIL THEFT	GROCERY FOOD STORE	t	f	111	1	42	28
38	3129776	HK119397	11/01/2004 16:06	065XX S FRANCISCO AVE	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT	f	f	831	8	15	66
39	3167851	HK119413	11/01/2004 16:05	063XX W ADDISON ST	1811	NARCOTICS	POSS: CANNABIS 30GMS OR LESS	RESIDENCE-GARAGE	t	f	1633	16	36	17

Εικόνα155. Sample crime data for Chicago (CSV)

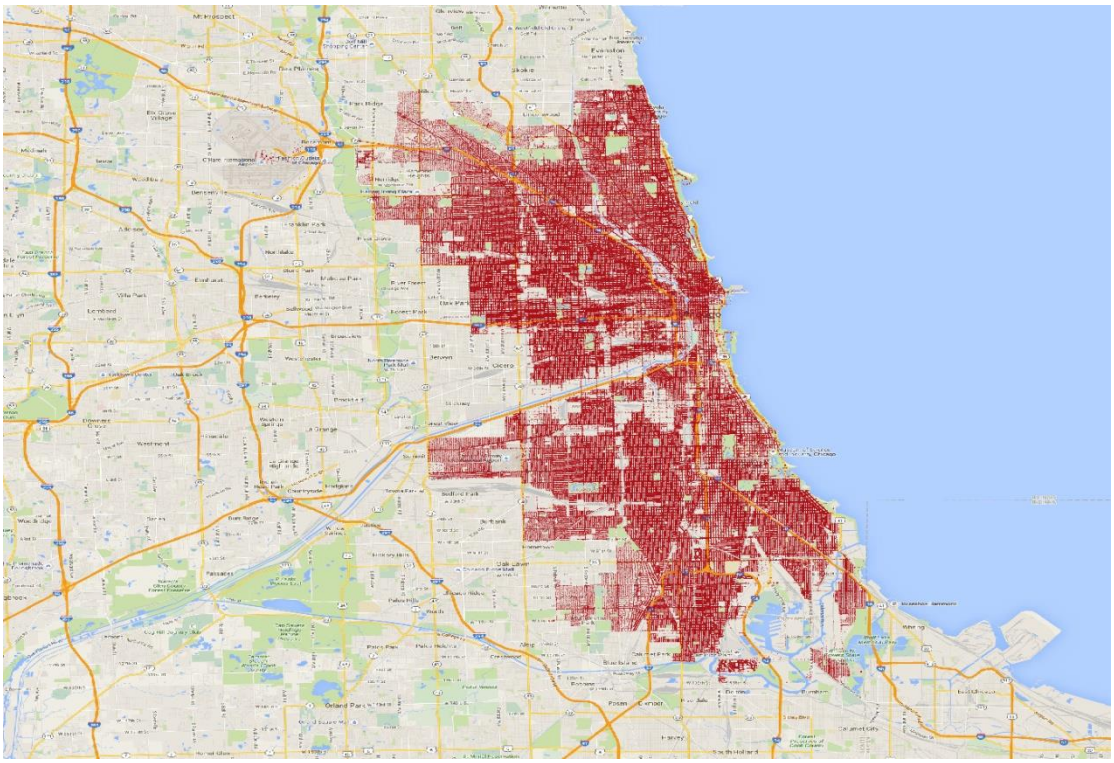
Το συγκεκριμένο dataset μπορεί να θεωρηθεί ιδιαίτερα πλούσιο, αφού περιέχει περισσότερες από 5.900.000 εγγραφές.

[23]https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

2.3.1 Οπτικοποίηση των δεδομένων εγκλημάτων

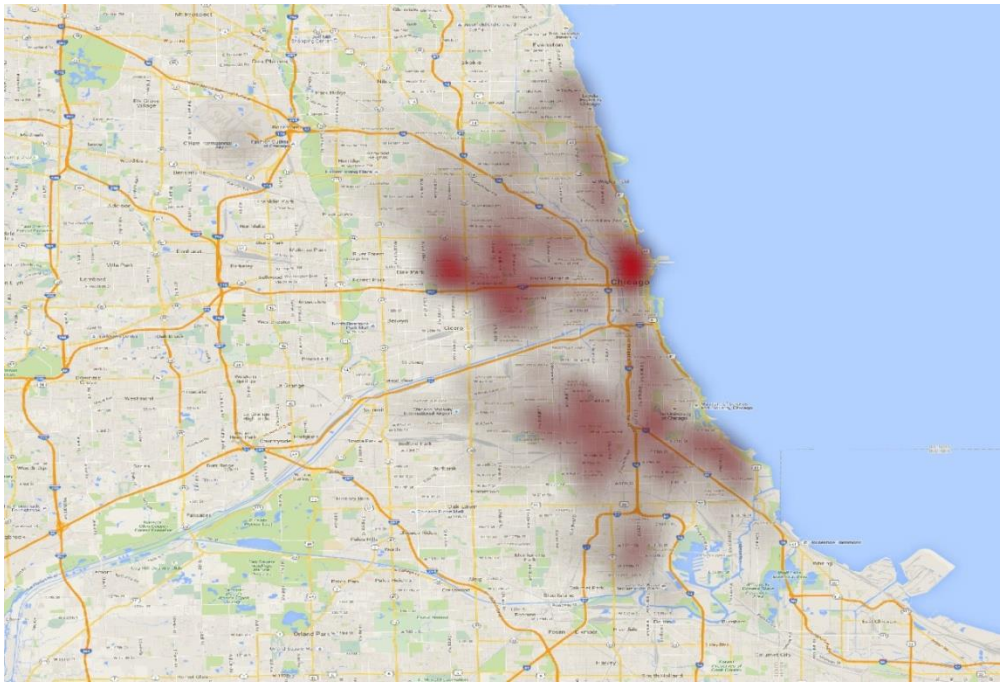
Στην παρούσα ενότητα θα περιγράψουμε κάποιους από τους πιθανούς τρόπους οπτικοποίησης των διαθέσιμων δεδομένων εγκλημάτων, ώστε να επιδείξουμετις διαφορετικές εναλλακτικές μεθόδους που έχουμε στη διάθεσή μας, καθώς και πως η συνάθροιση των δεδομένων σε ευρύτερες περιοχές επηρεάζει το τελικό αποτέλεσμα

Στην πρώτη εικόνα απεικονίζονται το σύνολο των εγκληματών στην περιοχή του Chicago.



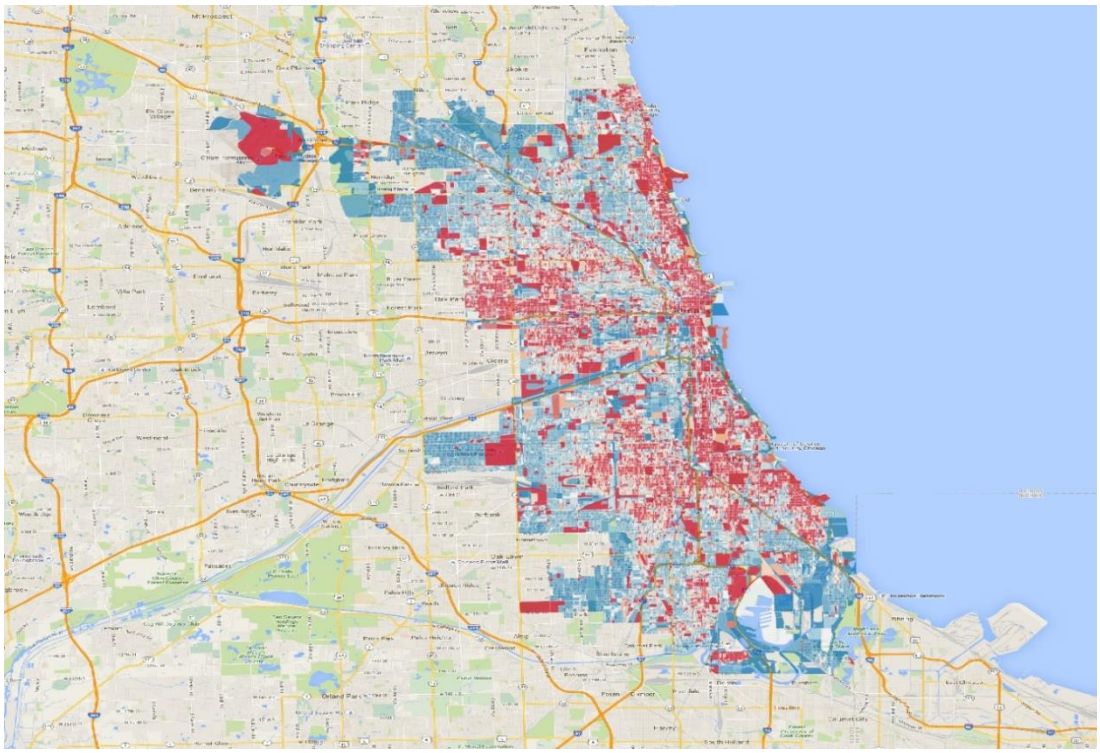
Εικόνα 166. Εγκλήματα στο Chicago

Καθώς ο αριθμός των εγκλημάτων είναι πολύ μεγάλος, η σημειακή απεικόνιση τους σε ευρεία κλίμακα δεν βοηθάει οπτικά στην εξαγωγή χρήσιμης πληροφορίας. Προκειμένου να υπερκεράσουμε αυτόν τον περιορισμό, στην επόμενη εικόνα παρουσιάζεται ένα Heatmap των εγκλημάτων, όπου όσο μεγαλύτερη είναι η πυκνότητα εγκλημάτωνμιας περιοχήςτόσο πιο έντονο είναι το κόκκινο νέφος πάνω από αυτή την περιοχή.Η συγκεκριμένη απεικόνιση βελτιώνει σημαντικά τον εντοπισμό των περιοχών με ασυνήθιστα μεγάλη εγκληματικότητα.



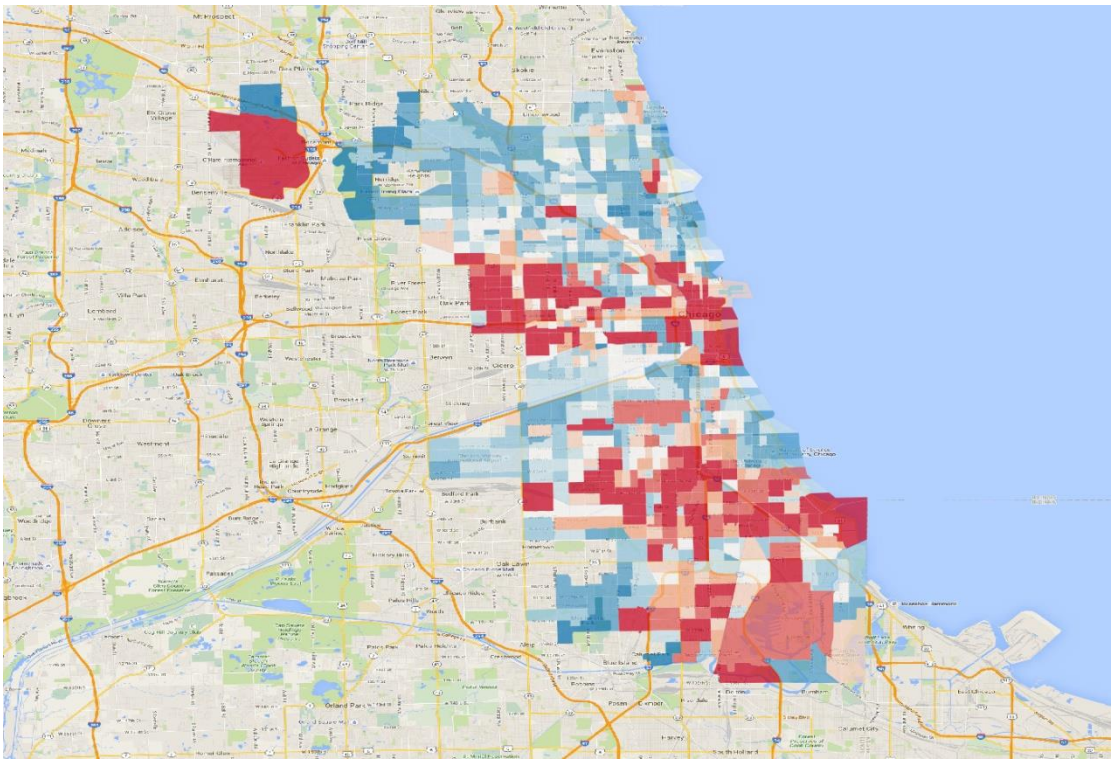
Εικόνα 177. Heatmap εγκλημάτων στο Chicago

Στην επόμενη εικόνα απεικονίζεται η επικινδυνότητα κάθεcensusblock του Chicago, σε σχέση με το μέσο όρο που προκύπτει από όλα τα censusblocks. Τα κόκκινα censusblocksεμφανίζουν εγκληματικότητα που είναι πάνω από το συνολικό μέσο όρο. Το αντίθετο ισχύει για τα αντίστοιχα μπλε censusblocks. Βλέπουμε πως η συγκεκριμένη απεικόνιση είναι ιδιαίτερα αποτελεσματική στο να εντοπίζει μικρές σχετικά περιοχές με ιδιαίτερα πλούσια εγκληματικότητα.



Εικόνα 188. Εγκληματική δραστηριότητα ανά census block του Chicago

Στην επόμενη εικόνα απεικονίζεται η επικινδυνότητα κάθεcensustract του Chicago, σε σχέση με το μέσο όρο που προκύπτει από όλα τα censustracts. Η συγκεκριμένη εικόνα δεν παρουσιάζει σαφή εποπτεία της επικινδυνότητας,επειδή κάθε censustract αντιστοιχεί σε σχετικά μεγάλη γεωγραφική περιοχή και συνεπώς χάθηκε σημαντική πληροφορία σε σύγκριση με την προηγούμενη εικόνα. Συνεπώς για τα συγκεκριμένα δεδομένα εγκλημάτων, η συνάθροιση σε επίπεδο censusblock αποτελεί τη βέλτιστη λύση.



Εικόνα 199. Εγκληματική δραστηριότητα κατά censustract του Chicago

2.4 Δεδομένα καιρού

Τα συγκεκριμένα δεδομένα καιρού[24]αφορούν ολόκληρη την αστική περιοχή του Chicago και είναι διαθέσιμα για πολλά έτη (ακόμα και πριν το 2000). Αναλυτικά, μπορεί να κατεβάσει κανείς για κάθε έτος και κάθε ημέραδεδομένα για τα παρακάτω μεγέθη (σε CSVμορφή):

- Ημερομηνία
- Θερμοκρασία (Μέγιστη, Ελάχιστη, Μέση)
- Σημείο Δρόσου (Μέγιστη, Ελάχιστη, Μέση)
- Υγρασία (Μέγιστη, Ελάχιστη, Μέση)
- Πίεση σε επίπεδο θάλασσας (Μέγιστη, Ελάχιστη, Μέση)
- Ορατότητα (Μέγιστη, Ελάχιστη, Μέση)
- Νεφοκάλυψη
- Διεύθυνση Ανέμου
- Άνεμος (km/hr) (Μέγιστη, Μέση)
- Ύψος βροχόπτωσης (mm) (Συνολική)
- Γεγονότα (π.χ. καταιγίδες)

[24]https://www.wunderground.com

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CST	Max Temperature C	Mean Temperature C	Min Temperature C	Dew Point C	Mean DewPointC	Min DewpointC	Max Humidity	Mean Humidity	Min Humidity	Max Sea Level Pressure hPa	Mean Sea Level Pressure hPa	Min Sea Level Pressure hPa	Max Visibility Km	Mean Visibility Km	Min Visibility kM	Max Wind Speed Km/h	Mean Wind Speed Km/h
2	01/01/2010	-8	-11	-14	-14	-16	-18	76	69	57	1030	1028	1025	16	15	8	23	16
3	02/01/2010	-11	-13	-16	-17	-20	-21	73	63	49	1033	1032	1030	16	16	11	26	19
4	03/01/2010	-8	-12	-17	-14	-17	-21	73	67	59	1033	1031	1029	16	15	10	26	19
5	04/01/2010	-7	-10	-13	-12	-14	-17	74	70	65	1029	1026	1023	16	16	8	24	18
6	05/01/2010	-5	-7	-9	-10	-12	-14	78	71	65	1025	1024	1022	16	15	3	24	16
7	06/01/2010	-5	-9	-13	-9	-12	-16	84	77	68	1025	1024	1021	16	15	10	13	10
8	07/01/2010	-6	-8	-9	-8	-10	-12	93	85	77	1022	1015	1011	10	3	1	19	8
9	08/01/2010	-2	-6	-9	-5	-9	-14	93	77	65	1031	1022	1014	16	10	1	35	18
10	09/01/2010	-5	-9	-13	-11	-14	-17	76	68	59	1035	1033	1030	16	16	16	19	13
11	10/01/2010	-7	-13	-18	-12	-17	-21	83	72	61	1035	1030	1021	16	16	13	29	14
12	11/01/2010	-3	-6	-8	11	-9	-13	74	69	66	1029	1023	1017	16	14	10	29	19
13	12/01/2010	-1	-3	-7	-6	-7	-10	81	75	64	1033	1031	1028	16	14	10	24	16
14	13/01/2010	1	-3	-8	-4	-7	-11	81	74	61	1027	1022	1019	16	15	11	24	18
15	14/01/2010	4	1	-1	0	-2	-5	92	76	69	1022	1018	1016	16	15	6	24	16
16	15/01/2010	1	-1	-4	-1	-3	-6	92	86	81	1027	1025	1022	8	5	3	21	8
17	16/01/2010	-2	-3	-4	-4	-5	-6	92	87	81	1026	1022	1017	8	5	2	19	13
18	17/01/2010	2	-2	-6	-3	-5	-7	96	90	64	1016	1013	1010	16	5	2	21	6
19	18/01/2010	-1	-2	-2	-2	-3	-4	96	91	85	1017	1015	1014	8	5	2	14	10
20	19/01/2010	2	-1	-4	-4	-5	-7	86	73	59	1018	1017	1015	16	15	10	14	6
21	20/01/2010	1	0	-1	-3	-4	-6	80	75	66	1017	1015	1013	16	15	11	34	21
22	21/01/2010	2	0	-1	-2	-4	-6	82	73	66	1014	1011	1009	16	15	6	32	26
23	22/01/2010	2	1	1	-1	-2	-3	89	83	72	1017	1015	1010	16	9	2	23	16
24	23/01/2010	7	4	1	6	2	-1	100	92	85	1016	1010	998	10	5	2	26	16
25	24/01/2010	8	3	-1	7	4	-4	97	89	76	997	992	989	16	11	2	42	23
26	25/01/2010	0	-3	-7	-4	-6	-11	86	76	66	1002	995	991	16	11	2	37	24
27	26/01/2010	-5	-8	-11	-9	-12	-16	78	72	67	1020	1012	1002	16	15	10	35	26
28	27/01/2010	-4	-8	-12	-7	-12	-17	86	74	60	1024	1022	1019	16	10	2	35	21
29	28/01/2010	-9	-11	-13	-14	-19	-22	67	56	47	1036	1032	1023	16	16	16	32	21
30	29/01/2010	-7	-11	-15	-10	-16	-22	80	61	48	1036	1033	1027	16	15	4	19	10
31	30/01/2010	-5	-7	-8	-11	-12	-14	74	64	55	1028	1025	1022	16	16	14	19	13
32	31/01/2010	-2	-6	-9	-10	-12	-15	81	63	36	1028	1025	1022	16	16	13	26	11
33	01/02/2010	-1	-4	-8	-6	-10	-13	72	62	54	1028	1026	1022	16	14	10	13	3
34	02/02/2010	0	-2	-4	-3	-4	-7	93	85	69	1022	1018	1015	10	5	1	21	10
35	03/02/2010	-1	-3	-5	-6	-7	-8	84	75	64	1031	1026	1020	16	9	5	19	11
36	04/02/2010	1	-2	-4	-4	-6	-7	81	74	66	1031	1028	1024	10	8	5	21	10

Εικόνα 30. Sample weather data for Chicago (CSV)

2.5 Αποθήκευση δεδομένων

Τα παραπάνω ανοικτά δεδομένα όλων των κατηγοριών ήταν διαθέσιμα σε μορφή CSV. Προκειμένου να τα επεξεργαστούμε, αποθηκεύτηκαν όλα στην ίδια βάση PostgreSQL που χρησιμοποιείται και από τον areaprofiler. Για κάθε CSV αρχείο δημιουργήσαμε και ένα διαφορετικό πίνακα στη βάση και για κάθε πεδίο διαθέσιμο στο αρχείο, έχουμε αντιστοιχίσει και μια διαφορετική στήλη σε κάθε πίνακα στη βάση, ώστε αρχικά να έχουμε 1-1 αντιστοιχία μεταξύ των original δεδομένων και των πινάκων στη βάση. Η τελική μορφή των δεδομένων, όπως αυτά θα χρησιμοποιηθούν από την τελική εφαρμογή θα οριστικοποιηθεί στη συνέχεια του έργου.

2.6 Σύνοψη

Στην παρούσα ενότητα περιγράψαμεκάποια από τα διαθέσιμα ανοικτά δεδομένα που θα χρησιμοποιηθούν στο έργο Citysense. Καθώς κύριος σκοπός του έργου είναι η αποτύπωση του «ίχνους» μιας αστικής περιοχής, δόθηκε έμφαση στα δεδομένα που σχετίζονται με την ευημερία και καλοζωία των κατοίκων. Τέτοια δεδομένα είναι η εγκληματικότητα, στοιχεία σχετικά με την οικονομική κατάσταση των κατοίκων μιας περιοχής, στοιχεία σχετικά με την υγεία των κατοίκων (που έμμεσα συνδέεται με τις συνθήκες μόλυνσης που επικρατούν στην περιοχή), στοιχεία κυκλοφορικής κίνησης που επηρεάζει σημαντικά την ποιότητα ζωής, καθώς και δεδομένα καιρού για όλη την πόλη του Chicago, αφού ο καιρός αποτελεί την πρώτη πληροφορία που χρειάζεται κάποιος επισκέπτης να ξέρει πριν επισκεφτεί μια καινούρια περιοχή. Τα δεδομένα που συλλέξαμε μπορούν να συνοψιστούν στον παρακάτω πίνακα.

Είδος Δεδομένων	**Πηγή**	**Αριθμός πεδίων**	**Αριθμός Εγγραφών**	**Format**
POIs	Google Places	11	184.392	API (JSON)
POIs	Foursquare	12	93.893	API (JSON)
Κοινωνικο-οικονομικοί δείκτες	Chicago Portal	8	77	CSV
Δείκτες Υγείας	Chicago Portal	28	77	CSV
Δεδομένα Κυκλοφορικής Κίνησης	Chicago Portal	8	1279	CSV
Εγκλήματα	Chicago Portal	22	5.915.820	CSV
Καιρός	Weather Underground	23	365 x Έτη	CSV

Στην επόμενη ενότητα, θα αναφερθούμε στο οδικό δίκτυο του Chicagoκαι την απαραίτητη προεπεξεργασία που πρέπει να υποστεί, ώστε να είναι δυνατή η χρήση του στα πλαίσια έργου Citysense.

3 Επεξεργασία Οδικού δικτύου

Μια από τις κύριες πληροφορίες που χρειάζεται κάποιος χρήστης για κάποια περιοχή ενδιαφέροντος είναι το οδικό της δίκτυο. Στη συνέχεια, μπορεί εύκολα να υπολογίσει τις πραγματικές αποστάσεις μεταξύ σημείων ενδιαφέροντος χρησιμοποιώντας το οδικό δίκτυο, αντί για την ευκλείδια απόσταση που θα εισήγαγε σημαντικό λάθος στους υπολογισμούς. Για το σκοπό αυτό θα χρησιμοποιήσουμε δεδομένα από το OpenStreetMaps[25](OSM) που είναι η σημαντικότερη διαδικτυακή υπηρεσία πληθοπορισμού για χάρτες και η οποία καλύπτει τις περισσότερες αστικές περιοχές του δυτικού κόσμου. Προκειμένου να μετατρέψουμε τα δεδομένα του OSMσε ένα routablegraphnetwork, θα χρησιμοποιήσουμε διάφορα εργαλεία , όπως το Osmosiskαι το Open Source Routing Machine (OSRM), τα οποία θα περιγράψουμε στη συνέχεια.

3.1 OpenStreetMap Δεδομένα

Για να κατεβάσει κανείς δεδομένα από το OpenStreetMap μπορεί να χρησιμοποιήσει το δικτυακό τόπο GEOFABRIK [26] , το οποίο διαθέτει χάρτες για όλης τηνυφήλιο, ομαδοποιημένους ανά μεγάλες περιοχές. Στην κατηγορία downloads του GEOFABRIK είναι διαθέσιμα τόσο ολόκληρη η Βόρεια Αμερική, όσο και η πολιτείατου Illinois, στην οποία ανήκει το Chicago. Για τους σκοπούς λοιπόν του Citysense, κατεβάσαμε τα δεδομένα του Illinois. Τα δεδομένα του OpenStreetMap είναι διαθέσιμα σε μορφή XML (αρχεία OSM) και binary/compressed (αρχεία PBF). Τα εργαλεία που χρησιμοποιούμε δουλεύουν και με τα δύο formats, αλλά εμείς κατεβάσαμε το αντίστοιχο XMLαρχείο για την πολιτεία του Illinois, κυρίως για να είναι humanreadableγια καλύτερη επισκόπηση των διαθέσιμων δεδομένων

3.2 Εξαγωγή περιοχής ενδιαφέροντος από ευρύτερο χάρτη

Έχοντας διαθέσιμο το OSM / XMLαρχείο για την πολιτεία του Illinois,, το εργαλείο που απαιτείται για την απομόνωση και εξαγωγή της περιοχής του Chicagoείναι το Osmosis[27]. ToOsmosis είναιμια command line Java εφαρμογήγια την επεξεργασία OSM δεδομένων.Προκειμένου να εξάγουμε μόνο τα δεδομένα που αφορούν την πόλη του Chicago, ήταν αναγκαίο να προσδιορίσουμε το MinimumBoundingRectangle(MBR) που περικλείει την πόλη. Για το Chicago τα απαραίτητα σημεία είναι αντίστοιχα (εκφρασμένα σε γεωγραφικό πλάτος και γεωγραφικό μήκος) τα (42.066516, -87.977981) και (41.606604, -87.462997). Έχοντας ως είσοδο το αρχείο του Ιλινόις και τα δύο διαγώνια σημεία του MinimumBoundingRectangle, το Osmosis εξάγει ένα μικρότερο οδικό δίκτυο που καλύπτει το οδικό δίκτυο του Chicago και μπορεί στη συνέχεια να χρησιμοποιηθεί από το Open Source Routing Machine (OSRM).

[25]https://www.openstreetmap.org
[26]http://www.geofabrik.de/
[27]http://wiki.openstreetmap.org/wiki/Osmosis

3.3 Open Source Routing Machine

Το Open Source Routing Machine (OSRM)[28] είναι ένα εργαλείο δρομολόγησης σε οδικά δίκτυα. Χρησιμοποιεί δεδομένα οδικού δικτύου από το OpenStreetMap και πραγματοποιεί δρομολόγηση βασισμένη σε μία υλοποίηση του αλγόριθμου Contraction Hierarchies [1]. Για τις ανάγκες του CitySense, το OSRM τροποποιήθηκε, ώστε να εξάγει δεδομένα τα οποία μπορούν να εισαχθούν σε PostgreSQL και δεδομένα που να μπορούν να χρησιμοποιηθούν για την εξαγωγή των Hub Labels [2] του οδικού δικτύου. Τα HubLabelsεκτός από το ότι αποτελούν την πιο γρήγορη μέθοδο δρομολόγησης σε οδικά δίκτυα, έχουν το πλεονέκτημα ότι μπορούν να χρησιμοποιηθούν και ενσωματωμένα σε μια σχεσιακή βάση δεδομένων [3],[4],[5], πράγμα που ταιριάζει απόλυτα με τους σκοπούς του Citysense.

3.4 Τεχνικές λεπτομέρειες

Το OSRM δέχεται ως είσοδο το οδικό δίκτυο (είτε OSM/XML, είτε PBF) και χρησιμοποιεί εσωτερικές δομές και παραγόμενα δυαδικά αρχεία για την υλοποίηση των αλγορίθμων επεξεργασίας του οδικού δικτύου και δρομολόγησης.

Το οδικό δίκτυο που παρέχεται από το OpenStreetMap είναι σε μορφή node-based γράφου με κόμβους και ακμές.Κάθε κόμβος αναπαριστά μια τοποθεσία (lon, lat) και κάθε ακμή αναπαριστά τη σύνδεση μεταξύ δύο σημείων, δηλαδή ένα κομμάτι δρόμου. Το οδικό δίκτυο περιέχει, επίσης, περιορισμούς στροφών. Το OSRM, μετά την αρχική εισαγωγή του δικτύου, μετατρέπει το node-basedOSMγράφο σε edge-based γράφο. Στη νέα αναπαράσταση, οι κόμβοι αναπαριστούν κομμάτια δρόμων και οι ακμές αναπαριστούν τη σύνδεση μεταξύ των κομματιών. Αν υπάρχει σύνδεση ανάμεσα σε κομμάτια δρόμων, τότε είναι δυνατή η μετάβαση από το ένα κομμάτι δρόμου στο άλλο. Διαφορετικά θεωρείται ότι δεν υπάρχει δυνατότητα μετάβασης από το ένα κομμάτι δρόμου στο άλλο. Η edge-based αναπαράσταση του οδικού δικτύου καθιστά δυνατήτη δρομολόγηση που λαμβάνει υπόψη της την ύπαρξη απαγορευμένων στροφών, κατά τον υπολογισμό των συντομότερων διαδρομών.

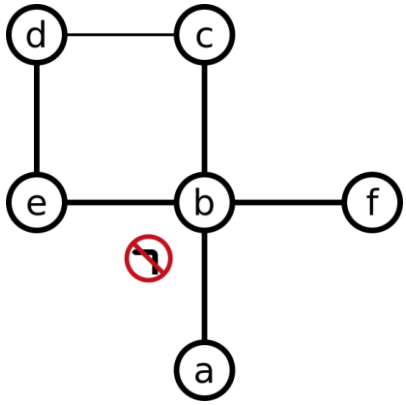
3.5 Τροποποίηση του OSRM

Για τις ανάγκες του CitySense λαμβάνονται δεδομένα οδικού δικτύου από το OpenStreetMap. Καθώς χρησιμοποιείται το σύστημα διαχείρισης βάσεων δεδομένωνPostgreSQL για το CitySense, πρέπει τα δεδομένα OpenStreetMap και τα αποτελέσματα της επεξεργασίας του OSRM να ληφθούν σε μία μορφή που να είναι συμβατήμε την PostgreSQL. Για αυτό το λόγο τροποποιήθηκε το OSRM, ώστε να παράγει tab-delimitedCSV αρχεία, τα οποία μπορούν να φορτωθούν σε αντίστοιχους σχεσιακούς πίνακες της PostgreSQL με την εντολή COPY. Επίσης τροποποιήθηκε το OSRM, ώστε να παράγει τα αποτελέσματα του contraction σε μορφή GR[29] και να εξάγει τη σειρά των κόμβων σε ένα αρχείο ORDER. Τα αρχεία GR/ORDERθα μπορούν να δίνονται στη συνέχεια ως είσοδος σε εφαρμογή για τον υπολογισμό των HubLabels του δικτύου.

[28]http://project-osrm.org/
[29]http://www.dis.uniroma1.it/challenge9/format.shtml#graph

3.6 Παράδειγμα οδικού δικτύου

Προκειμένου να επιδείξουμε τα ενδιάμεσα στάδια της επεξεργασίας των δεδομένων του OSMαπό το εργαλείο OSRM, θα χρησιμοποιήσουμε ένα ενδεικτικό παράδειγμα. Έστω το οδικό δίκτυο που παρουσιάζεται στο παρακάτω σχήμα. Στο συγκεκριμένο παράδειγμα, κάθε δρόμος είναι 100m και διπλής κατεύθυνσης. Όλοι οι δρόμοι είναι κύριοι (Primary) εκτός από τον "cd" που είναι τοπικός (residential). Υπάρχει μία απαγόρευση αριστερής στροφής από τον "ab" στον "eb". Οι υπόλοιπες στροφές επιτρέπονται κανονικά.



Εικόνα 31. Παράδειγμα οδικού δικτύου

3.7 Αρχείο OSM

Το αντίστοιχο αρχείο OSM που περιγράφει το παραπάνω οδικό δίκτυο θα έχειτην εξής μορφή:

```
<?xmlversion="1.0"encoding="UTF-8"?>
<osmgenerator="xnakos"version="0.6">
<nodeid="10000"version="1"uid="1"user="xnakos"timestamp="2008-09-21T21:37:45Z"lon="-
87.619726"lat="41.838527">
<tagk="name"v="a"/>
</node>
<nodeid="10001"version="1"uid="1"user="xnakos"timestamp="2008-09-21T21:37:45Z"lon="-
87.619724"lat="41.839427">
<tagk="name"v="b"/>
</node>
<nodeid="10002"version="1"uid="1"user="xnakos"timestamp="2008-09-21T21:37:45Z"lon="-
87.619721"lat="41.840326">
<tagk="name"v="c"/>
</node>
<nodeid="10003"version="1"uid="1"user="xnakos"timestamp="2008-09-21T21:37:45Z"lon="-
87.620928"lat="41.840325">
<tagk="name"v="d"/>
</node>
<nodeid="10004"version="1"uid="1"user="xnakos"timestamp="2008-09-21T21:37:45Z"lon="-
87.620933"lat="41.839426">
<tagk="name"v="e"/>
</node>
<nodeid="10005"version="1"uid="1"user="xnakos"timestamp="2008-09-21T21:37:45Z"lon="-
87.618518"lat="41.839427">
<tagk="name"v="f"/>
</node>
<wayid="20000"version="1"uid="1"user="xnakos"timestamp="2008-09-21T21:37:45Z">
<ndref="10000"/>
<ndref="10001"/>
<tagk="highway"v="primary"/>
<tagk="oneway"v=""/>
<tagk="name"v="ab"/>
</way>
<wayid="20001"version="1"uid="1"user="xnakos"timestamp="2008-09-21T21:37:45Z">
<ndref="10001"/>
<ndref="10002"/>
<tagk="highway"v="primary"/>
```

```
<tag k="oneway" v="" />
<tag k="name" v="bc" />
</way>
<way id="20002" version="1" uid="1" user="xnakos" timestamp="2008-09-21T21:37:45Z">
<nd ref="10002" />
<nd ref="10003" />
<tag k="highway" v="residential" />
<tag k="oneway" v="" />
<tag k="name" v="cd" />
</way>
<way id="20003" version="1" uid="1" user="xnakos" timestamp="2008-09-21T21:37:45Z">
<nd ref="10003" />
<nd ref="10004" />
<tag k="highway" v="primary" />
<tag k="oneway" v="" />
<tag k="name" v="de" />
</way>
<way id="20004" version="1" uid="1" user="xnakos" timestamp="2008-09-21T21:37:45Z">
<nd ref="10004" />
<nd ref="10001" />
<tag k="highway" v="primary" />
<tag k="oneway" v="" />
<tag k="name" v="eb" />
</way>
<way id="20005" version="1" uid="1" user="xnakos" timestamp="2008-09-21T21:37:45Z">
<nd ref="10001" />
<nd ref="10005" />
<tag k="highway" v="primary" />
<tag k="oneway" v="" />
<tag k="name" v="bf" />
</way>
<relation id="30000" version="1" uid="1" user="xnakos" timestamp="2008-09-21T21:37:45Z">
<member type="way" ref="20000" role="from" />
<member type="way" ref="20004" role="to" />
<member type="node" ref="10001" role="via" />
<tag k="restriction" v="no_left_turn" />
<tag k="type" v="restriction" />
</relation>
</osm>
```

Η πρώτη βασική οντότητα που εμφανίζεται στα OSMαρχεία είναι το node. Το node αναπαριστά μια γεωγραφική τοποθεσία (lon, lat). Βασικά του γνωρίσματα είναι το αναγνωριστικό του, με το οποίο αναφέρονται σε αυτό άλλες οντότητες του αρχείου, το γεωγραφικό μήκος και το γεωγραφικό πλάτος.

Η δεύτερη βασική οντότητα στα OSM αρχεία είναι το way. Το way αναπαριστά μία διαδρομή. Αποτελείται από μία orderedλίστα από αναφορές σε nodes (βάσει των αναγνωριστικών τους), τα οποία αναπαριστούν τα σημεία από τα οποία περνάει η διαδρομή, καθώς και γνωρίσματα που δείχνουν το είδος του δρόμου (αν ο δρόμος αποτελεί κύρια οδική αρτηρία, δευτερεύουσα οδική αρτηρία, κ.λπ.), το όνομα του δρόμου και το αν είναι μονής ή διπλής κατεύθυνσης. Αν σε μια τοποθεσία τέμνονται δύο δρόμοι, τότε τα ways τα οποία τους αναπαριστούν στο σημείο τομής θα έχουν ένα κοινό node. Έναway μπορεί να είναι κλειστό, αν το πρώτο και το τελευταίο node της λίστας συμπίπτουν.

Η τρίτη βασική οντότητα που εμφανίζεται στα OSMαρχεία είναι το relation. Αποτελείται από μία λίστα από nodes, ways, άλλα relations, καθώς και λοιπές οντότητες (tags), ώστε το relation να αναπαριστά μία σχέση ανάμεσα στις οντότητες από τις οποίες αποτελείται. Στο παραπάνω παράδειγμα υπάρχει ένα relation ανάμεσα σε δύο ways και ένα node, το οποίο αναπαριστά την απαγόρευση της αριστερής στροφής (από τα tags προκύπτει το "noleftturnrestriction") από το way με ρόλο "from" στο way με ρόλο "to" μέσω του node με ρόλο "via" (σημείο τομής των ways).

Το παραπάνω οδικό δίκτυο, όταν δοθεί ως είσοδος στο τροποποιημένο OSRM, με τις εντολές "osrm-extract" και "osrm-prepare", παράγει τα αρχεία που περιγράφονται παρακάτω.

3.8 Node-based graph

Αρχικά το OSRMμετατρέπει το OSMαρχείο που δέχεται ως είσοδο σε node-basedγράφο. Για λόγους πληρότητας, onode-basedγράφος εξάγεται σε 4CSVαρχείαπου περιγράφονται στη συνέχεια.

3.8.1 Node-based nodes CSV

Περιλαμβάνει τους κόμβους (σημεία) του δικτύου. Συγκεκριμένα για κάθε κόμβο περιλαμβάνει το αναγνωριστικό του (με βάση το οποίο αναφέρονται σε αυτόν παραγόμενες δομές από το OSRM), το γεωγραφικό πλάτος, το γεωγραφικό μήκος, και το OSM αναγνωριστικό του (το οποίο χρησιμεύει για την ταυτοποίηση του κόμβου με το αρχικό OSMοδικό δίκτυο).

id	lat	lon	osm_id
0	41838527	-87619726	10000
1	41839427	-87619724	10001
2	41840326	-87619721	10002
3	41840325	-87620928	10003
4	41839426	-87620933	10004
5	41839427	-87618518	10005

3.8.2 Node-based edges CSV

Περιλαμβάνει τις ακμές (δρόμους) του δικτύου. Συγκεκριμένα για κάθε ακμή περιλαμβάνει το αναγνωριστικό του κόμβου αρχής, το αναγνωριστικό του κόμβου τέλους, το αναγνωριστικό του ονόματος του δρόμου, το αναγνωριστικό του τύπου του δρόμου, το βάρος της ακμής, αν ο δρόμος έχει την κατεύθυνση από την αρχή προς το τέλος, και αν ο δρόμος έχει την κατεύθυνση από το τέλος προς την αρχή.

source	target	name_id	highway_id	weight	forward	backward
0	1	1	1	57	1	1
1	2	2	1	57	1	1
1	4	6	1	57	1	1
1	5	4	1	57	1	1
2	3	5	2	116	1	1
3	4	3	1	57	1	1

3.8.3 Names CSV

Περιλαμβάνει τα ονόματα των δρόμων του δικτύου. Συγκεκριμένα για κάθε όνομα δρόμου περιλαμβάνει το αναγνωριστικό του και το όνομα του δρόμου.

name_id	name
0	
1	ab
2	bc
3	de
4	bf
5	cd
6	eb

3.8.4 Highways CSV

Περιλαμβάνει τους τύπους των δρόμων του δικτύου. Συγκεκριμένα για κάθε τύπο δρόμου περιλαμβάνει το αναγνωριστικό του και τον τύπο του δρόμου.

highway_id	highway
0	
1	primary
2	residential

3.9 Edge-based graph

Μια από τις πιο συνηθισμένες τεχνικές προκειμένου να εισάγουμε απαγορευμένες στροφές στους αλγόριθμους δρομολόγησης, είναι η μετατροπή του κλασσικής node-basedαναπαράστασης του οδικού δικτύου (όπου οι κόμβοι παριστάνουν διασταυρώσεις και οι ακμές παριστάνουν τους δρόμους μεταξύ δύο διασταυρώσεων) στην edge-basedαναπαράσταση, όπου οι κόμβοι παριστάνουν δρόμους και οι ακμές παριστάνουν τις διασταυρώσεις. Αυτήν την τεχνική χρησιμοποιεί και το OSRM. Στη συνέχεια θα περιγράψουμε τα CSVαρχεία που παράγονται κατά τη διαδικασία αυτή.

3.9.1 Edge-based nodes CSV

Το αρχείο περιλαμβάνει τους κόμβους του edge-based graph. Κάθε κόμβος αναπαριστά ένα δρόμο του δικτύου. Κάθε γραμμή του αρχείου περιλαμβάνει το αναγνωριστικό του κόμβου που αναπαριστά την προς-τα-εμπρός κίνηση, δηλαδή την κίνηση από το u στο v, για δύο σημεία u και v του δικτύου (κόμβοι του node-based graph), το αναγνωριστικό του κόμβου που αναπαριστά την προς-τα-πίσω κίνηση, δηλαδή από το v στο u, για τα ίδια σημεία, το αναγνωριστικό του σημείου u στο node-based graph, το αναγνωριστικό του σημείου v στο node-based graph, το αναγνωριστικό του ονόματος δρόμου, το βάρος του δρόμου για την κατεύθυνση από το u στο v, και το βάρος του δρόμου για την κατεύθυνση το v στο u.

forward_edge based_node_id	reverse_edge based_node_id	u	v	name_id	Forward weight	Reverse weight
0	1	0	1	1	57	57
2	5	1	2	2	57	57
3	9	1	4	6	57	57
4	11	1	5	4	57	57
6	7	2	3	5	116	116
8	10	3	4	3	57	57

3.9.2 Edge-based edges CSV

Περιλαμβάνει τις ακμές του edge-based graph. Κάθε ακμή αναπαριστά τη δυνατότητα μετάβασης από ένα κόμβο-δρόμο σε άλλον. Συγκεκριμένα για κάθε ακμή περιλαμβάνει το αναγνωριστικό κόμβου αρχής, το αναγνωριστικό κόμβου τέλους, το αναγνωριστικό της ακμής, το βάρος της ακμής, το αν η ακμή συνδέει τον κόμβο αρχής με τον κόμβο τέλους με αυτήν την κατεύθυνση, και το αν η ακμή συνδέει τον κόμβο τέλους με τον κόμβο αρχής με αυτή την κατεύθυνση. Αν υπάρχουν περιορισμοί στροφών στο οδικό δίκτυο, τότε αφαιρούνται οι αντίστοιχες ακμές από τον edge-based graph, με αποτέλεσμα οι απαγορευμένες μεταβάσεις να μην είναι διαθέσιμες στη συνέχεια.

source	target	edge_id	weight	forward	backward
0	2	0	57	1	0
0	4	1	65	1	0
1	0	2	90	1	0
2	6	3	69	1	0
3	10	4	65	1	0
4	11	5	90	1	0
5	1	6	57	1	0
5	3	7	65	1	0
5	4	8	69	1	0
6	8	9	127	1	0
7	5	10	124	1	0
8	9	11	69	1	0
9	1	12	65	1	0
9	2	13	68	1	0
9	4	14	57	1	0
10	7	15	65	1	0
11	1	16	68	1	0
11	2	17	65	1	0
11	3	18	57	1	0

3.10 Contracted edge-based graph

ΑφούτοOSRMμετέτρεψετον αρχικόnode-basedγράφο που προέκυψε από την μετατροπή του αρχικού OSMαρχείου στην αντίστοιχη edge-based αναπαράσταση, μετά εκτελεί την επεξεργασία του ContractionHierarchies [1] αλγόριθμου στην edge-based αναπαράσταση. Έξοδος της παραπάνω διαδικασίας είναι η ιεράρχηση των κόμβων σε επίπεδα (υψηλότερο επίπεδο έχουν οι κόμβοι που συμμετέχουν σε πολλά μονοπάτια συντομότερης διαδρομής) και ένας contractedγράφος που περιέχει shortcuts που επιταχύνουν την εύρεση της συντομότερης διαδρομής μεταξύ δύο κόμβων. Κατόπιν, κατά την εκτέλεση ερωτημάτων συντομότερης διαδρομής, αρκεί να τρέξουμε μια τροποποιημένη έκδοση του bidirectionalDijkstraστον contractedγράφο, κατά τέτοιον τρόπο ώστε κάθεμια από τις δύο επιμέρους αναζητήσεις να κατευθύνεται σε κόμβους ίδιου ή μεγαλυτέρου επιπέδου. Για περισσότερες πληροφορίες σχετικά με τον αλγόριθμο ContractionHierarchies κανείς μπορεί να συμβουλευτεί το [1].

3.10.1 Edge-based contracted edges CSV

Περιλαμβάνει τις ακμές που προκύπτουν ως αποτέλεσμα του contraction. Κάθε γραμμή περιλαμβάνει το αναγνωριστικό του κόμβου αρχής, το αναγνωριστικό του κόμβου τέλους, την απόσταση, το αν η ακμή αποτελεί συντόμευση ή όχι (αν αποτελεί συντόμευση, τότε δεν υπάρχει στον originaldge-based graph αλλά δημιουργήθηκε κατά το contraction), αν η ακμή έχει κατεύθυνση προς-τα-εμπρός και αν η ακμή έχει κατεύθυνση προς-τα-πίσω.

source	target	distance	is_shortcut	forward	backward
0	1	90	f	f	t
0	2	57	f	t	f
0	4	65	f	t	f
1	2	147	t	t	f
1	4	155	t	t	f
1	9	65	f	f	t
1	11	68	f	f	t
3	5	254	t	t	f
3	5	65	f	f	t
3	11	57	f	f	t
4	9	57	f	f	t
4	11	90	f	t	f
4	11	223	t	f	t
5	1	57	f	t	f
5	4	69	f	t	f
5	11	311	t	f	t
6	2	69	f	f	t
6	9	196	t	t	f
7	5	124	f	t	f
7	10	65	f	f	t
8	6	127	f	f	t
8	9	69	f	t	f

source	target	distance	is_shortcut	forward	backward
9	2	68	f	t	f
9	2	265	t	f	t
10	3	65	f	f	t
10	5	189	t	t	f
11	2	65	f	t	f
11	9	147	t	f	t

3.10.2 Edge-based node levels CSV

Περιλαμβάνει τα επίπεδα των κόμβων στο Contraction Hierarchy. Κάθε γραμμή του αρχείου περιλαμβάνει το αναγνωριστικό του κόμβου και το αντίστοιχο επίπεδό του.

node_id	level
0	2
1	4
2	8
3	2
4	5
5	3
6	1
7	0
8	0
9	7
10	1
11	6

3.10.3 Edge-based contracted edges, αρχείο GR

Περιλαμβάνειτον contracted edge-based graph. Η πρώτη γραμμή δηλώνει το πλήθος των κόμβων του γράφου και το πλήθος των ακμών του γράφου (που ταυτίζεται με το πλήθος των υπόλοιπων γραμμών). Κάθε γραμμή στη συνέχεια αποτελεί μία ακμή και περιλαμβάνει τον κόμβο αρχής, τον κόμβο τέλους, και το βάρος της ακμής. Η διαφορά από τις προηγούμενες αναπαραστάσεις είναι ότι η αρίθμηση/αναγνώριση των κόμβων στην GR αναπαράσταση ξεκινάει από το "1" (και όχι από το 0).

```
p sp 12 28
a 1 3 57
a 1 5 65
a 2 1 90
a 2 3 147
a 2 5 155
a 3 7 69
a 3 10 265
a 4 6 254
a 4 11 65
a 5 12 90
a 6 2 57
a 6 4 65
a 6 5 69
a 7 9 127
a 7 10 196
a 8 6 124
a 9 10 69
a 10 2 65
a 10 3 68
a 10 5 57
a 10 12 147
a 11 6 189
a 11 8 65
a 12 2 68
a 12 3 65
a 12 4 57
a 12 5 223
a 12 6 311
```

3.10.4 Edge-based node order, αρχείο ORDER

Περιλαμβάνει σε φθίνουσα σειρά επιπέδων Contraction Hierarchy τα αναγνωριστικά των κόμβων του edge-based graph. Η πρώτη σειρά περιλαμβάνει το πλήθος των κόμβων του γράφου και οι επόμενες σειρές περιλαμβάνουν τα αναγνωριστικά των κόμβων από το υψηλότερο επίπεδο προς το χαμηλότερο επίπεδο. Η αρίθμηση/αναγνώριση των κόμβων ξεκινάει από το "0".

```
12
2
9
11
4
1
5
0
3
6
10
7
8
```

3.11 Δημιουργία Hub Labels

Ο αλγόριθμος hub labels αποτελεί την πιο γρήγορη μέθοδο δρομολόγησης σε οδικά δίκτυα [2]. Επιπλέον, έχει το πλεονέκτημα ότι είναι η μόνη μέθοδος δρομολόγησης που μπορεί να χρησιμοποιηθεί και σε μια σχεσιακή βάση δεδομένων (βλέπε [3],[4],[5]). Περισσότερα για το συγκεκριμένο αλγόριθμο δρομολόγησης μπορεί κανείς να βρει στο [2].

Προκειμένου να φτιάξουμε τα hub labels θα χρησιμοποιήσουμε τα αρχεία ORDER και GR (της προηγούμενης ενότητας) ως είσοδο. Στη συνέχεια, χρησιμοποιούμε την εφαρμογή "akiba" από το "Hub Labeling Algorithms" repository[30]. Περισσότερα για τον συγκεκριμένο αλγόριθμο δημιουργίας Hub labels μπορεί κανείς να βρει στο [6]. Τα παραγόμενα hub labels εξάγονται σε δύο αρχεία forward και reverse που φαίνονται στις παρακάτω ενότητες.

3.11.1 Forward Hub Labels

source	target	weight
0	0	57
0	2	155
0	3	65
0	6	0
1	0	147
1	2	245
1	3	155
1	4	0
2	0	0
3	0	458
3	2	413
3	7	0
4	0	155
4	2	90
4	3	0
5	0	204
5	2	159
5	3	69
5	4	57
5	5	0
6	0	264
6	8	0
7	0	328
7	2	283
7	3	193
7	4	181
7	5	124

[30] https://github.com/savrus/hl

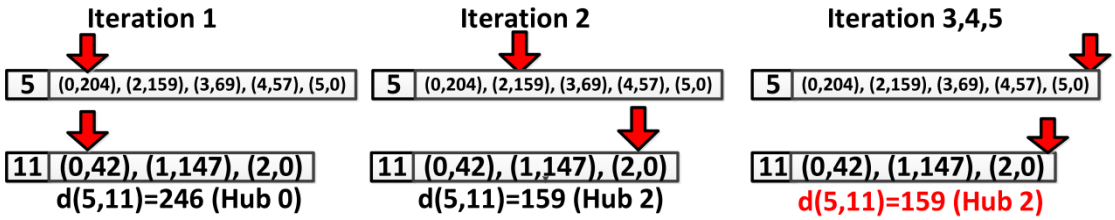
source	target	weight
7	10	0
8	0	137
8	1	69
8	11	0
9	0	68
9	1	0
10	0	393
10	2	348
10	4	246
10	5	189
10	9	0
11	0	65
11	2	0

3.11.2 Reverse Hub Labels

source	target	weight
0	0	420
0	1	155
0	2	158
0	4	90
0	6	0
1	0	330
1	1	65
1	2	68
1	4	0
2	0	0
3	0	469
3	1	204
3	2	57
3	5	65
3	7	0
4	0	322
4	1	57
4	2	223
4	3	0
5	0	723
5	1	458
5	2	311
5	5	0
6	0	69

source	target	weight
6	8	0
7	0	599
7	1	334
7	2	187
7	5	195
7	7	130
7	9	65
7	10	0
8	0	196
8	8	127
8	11	0
9	0	265
9	1	0
10	0	534
10	1	269
10	2	122
10	5	130
10	7	65
10	9	0
11	0	412
11	1	147
11	2	0

Αφού έχουν υπολογιστεί τα forwardκαι reversehublabels για κάθε κόμβο, προκειμένου να υπολογίσουμε το κόστος της συντομότερης διαδρομής μεταξύ δύο κόμβων υκαι v, αρκεί να χρησιμοποιήσουμε τα forwardlabelstου u(αρχή) και τα reverselabelstου v(προορισμός) και να βρούμε το hubwπου ελαχιστοποιεί το άθροισμα d(u,w)+d(w,v) (Δες την παρακάτω εικόνα).



Εικόνα 3220. A HL query μεταξύ των κόμβων 5 και 11

3.12 Στατιστικά οδικού δικτύου περιοχής ενδιαφέροντος (Chicago)

Στην παρούσα ενότητα θα αναφερθούμε στα ποσοτικά χαρακτηριστικά του οδικού δικτύου για την περιοχή ενδιαφέροντος του Chicago.

3.12.1 Chicago OSM

Το αρχείο OSM για το Chicago, το οποίο προέκυψε από την εφαρμογή του Osmosis πάνω στο αρχείο OSM για το Illinois βάσει του MinimumBoundingRectangle που ορίζεται από τις συντεταγμένες (41.606604, -87.462997) και (42.066516, -87.977981) περιέχει 304.778 nodes, 69.740 ways, και 2.267 relations. Από τα συνολικά relations, τα 518 αναπαριστούν απαγορευμένες στροφές (turning restrictions) και τα οποία θα χρησιμοποιηθούν στη συνέχεια από το OSRM.

3.12.2 Node-based graph γιατο Chicago

Ο node-based graph για το Chicago, όπως προκύπτει από το OSRM, περιέχει 297.113 nodeskαι 378.830 edges, που αναπαριστούν συνδέσεις μεταξύ δύο γεωγραφικών σημείων, είτε προς τη μία κατεύθυνση (unidirectional), είτε και προς τις δύο κατευθύνσεις (bidirectional).

3.12.3 Edge-based graph γιατο Chicago

Ο edge-based graph για το Chicago, όπως προκύπτει από το OSRM, περιέχει 419.398 nodes, που αναπαριστούν δρόμους μεταξύ δύο γεωγραφικών σημείων κατά μία κατεύθυνση, δηλαδή έγκυρες διαδρομές μίας κατεύθυνσης και 929.477 edges, που αναπαριστούν έγκυρες συνδέσεις μεταξύ των διαδρομών αυτών. Αυτός είναι ο τελικός edge-based graph, (στον οποίον δεν συμπεριλαμβάνονται τα edges τα οποία ακυρώνονται από τις 518 απαγορευμένες στροφές) και συνεπώς είναι κατάλληλος για δρομολόγηση.

3.12.4 Edge-based contracted graph γιατο Chicago

Ο edge-based contracted graph γιατο Chicago, όπωςπροκύπτειαπότο OSRM, περιέχει419.398nodes και3.273.044edges. Όπωςαναμενόταν, οedge-basedcontractedgraphγιατοChicagoέχειτονίδιοαριθμόnodesμετονedge-basedgraphγιατοChicago, αλλάπολλάπερισσότεραedges (3.5χπερισσότερα)λόγωτηςπροσθήκηςτωναπαραίτητωνshortcutsπουαπαιτούνταιγιατοναλγόριθμοContractionHierarchies.

3.12.5 Hub Labels γιατο Chicago

Όπωςπροκύπτειαπότηνεφαρμογήτηςεφαρμογή "akiba" απότο "Hub Labeling Algorithms" repositoryπάνωστον edge-based contracted graph γιατο Chicago, τα forward hub labels γιατο Chicago είναισυνολικά91.052.682καιτα reverse hub labels γιατο Chicago είναισυνολικά93.410.912. Συνεπώς, τομέσο label size είναι219,915 labels ανά node καιτομέγιστο label size είναι735 labels.

3.13 Σύνοψη

Στην παρούσα ενότητα περιγράψαμε την πολύπλοκη διαδικασία, προκειμένου να πάρουμε το οδικό δίκτυο μιας περιοχής ενδιαφέροντος από το OpenStreetMapsκαι να την μετατρέψουμε σε τέτοια μορφή (HubLabels)ώστε να μπορούμε εύκολα να απαντάμε ερωτήματα συντομότερης διαδρομής. Επίσης, περιγράψαμε τα αποτελέσματα της συγκεκριμένης διαδικασίας για την πόλη του Chicago. Βέβαια, ανάλογα με τις τελικές απαιτήσεις της εφαρμογής που θα οριστικοποιηθούν στους επόμενους μήνες, θα δούμε με ποιο ακριβώς τρόπο η παρούσα διαδικασία θα ενταχθεί στην τελική εφαρμογή – demotou Citysense.

4 Αποθήκευση Δεδομένων του CitySense

Για την αποθήκευση των δεδομένων, που θα χρησιμοποιηθούν στην εφαρμογή του CitySense, χρησιμοποιείται μία σχεσιακή βάση δεδομένωνμε σκοπό την ορθή οργάνωση δεδομένων που συλλέχθηκαν από τον AreaProfiler και αφορούν τα σημεία ενδιαφέροντος και τα γεωχωρικά tweets. Με την χρήση της βάσης είναι εφικτή η άμεση ανάκτηση της πληροφορίας με διάφορα κριτήρια σύμφωνα με τις απαιτήσεις της εφαρμογής.

4.1 Ανάλυση Απαιτήσεων Βάσης Δεδομένων

Οι απαιτήσεις που πρέπει να ικανοποιεί η βάση δεδομένων αφορούν την οργάνωση πληροφοριών για τα σημεία ενδιαφέροντος, τα ανοικτά δεδομένα, αλλάκαι τα γεωχωρικά tweets, που έχουν συλλεχθεί. Οι απαιτήσεις που θέτουμε σε αυτήν την ενότητα είναι αρκετές τόσο σε σχέση με τις ανάγκες της εφαρμογής, όσο και σε σχέση με αλλαγές που μπορεί να προκύψουν στις απαιτήσεις της εφαρμογής. Οι απαιτήσεις της βάσης δεδομένων περιγράφονται παρακάτω ανά κατηγορία δεδομένων.

4.1.1 Σημεία Ενδιαφέροντος

Για τα σημεία ενδιαφέροντος απαιτείται η αποθήκευση των χαρακτηριστικών που τα συνοδεύουν, όπως το όνομα, η διεύθυνση, ώρες λειτουργίας, τα στοιχεία επικοινωνίας, το tract στο οποίο βρίσκονται. Παράλληλα για τα σημεία ενδιαφέροντος είναι απαραίτητη η αποθήκευση της κατηγορίας που ανήκουν (εστίαση, νοσοκομεία, κλπ), αλλά και οι βαθμολογίες των χρηστών.Επιπλέον, η βάση πρέπει να περιέχει στοιχεία και παραμέτρους που αφορούν τους crawlers που χρησιμοποιήθηκαν για την συλλογή των σημείων ενδιαφέροντος. Τέλος, για τα σημεία ενδιαφέροντος δεν χρειάζεται να υπάρχει ξεχωριστή χρονική πληροφορία, καθώς θεωρούνται στατικά ως προς το χρόνο.

4.1.2 Ανοικτά Δεδομένα

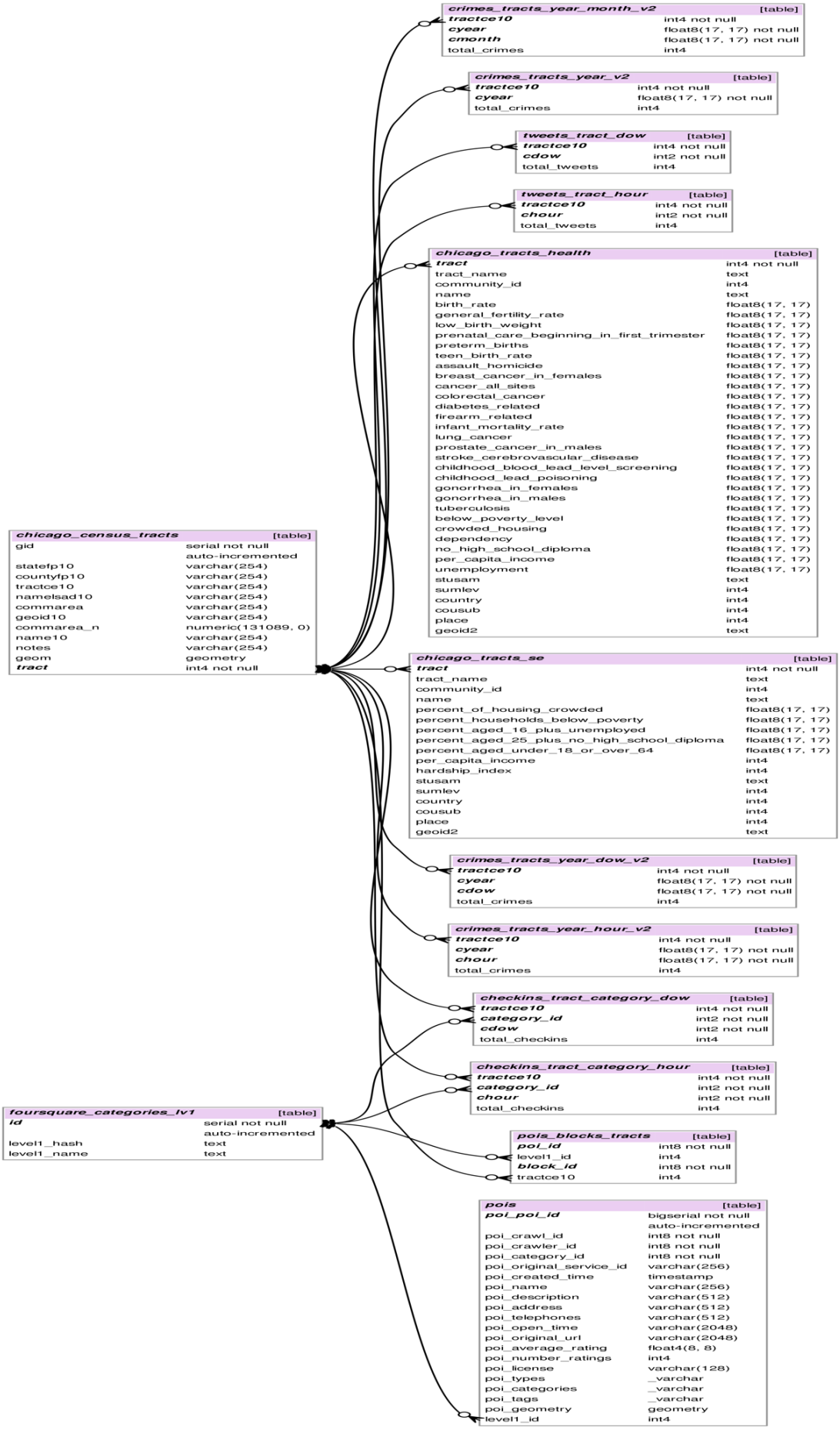
Τα ανοικτά δεδομένα απαρτίζονται τόσο από στατικά όσο και από χρονικά δεδομένα. Για παράδειγμα τα δεδομένα απογραφής, οι κοινωνικο-οικονομικοί δείκτες, οι δείκτες υγείας παρουσιάζονται σαν στατικά δεδομένα καιεπομένως για αυτά τα δεδομένα δεν είναι απαραίτητη η αποθήκευση χρονικής πληροφορίας. Αντίθετα για τα δεδομένα των εγκλημάτων χρειάζεται η αποθήκευση της πληροφορίας για την κατανομή τους στο χρόνο, ανά μήνα, ημέρα, κλπ.

4.1.3 StreamingΔεδομένααπόSocialMedia

ΤαStreamingΔεδομένααπόSocialMedia αποτελούν χρονικά δεδομένα, επομένως απαιτείται η αποθήκευση της πληροφορίας για την κατανομή τους στο χρόνο, ανά ημέρα και ώρα. Επιπλέον απαιτείται πέρα από τη χωρική πληροφορία, να αποθηκεύονται επίσης ξεχωριστά τα tweetsπου αφορούν check-insσε κάποιο σημείο ενδιαφέροντος μαζί με την κατηγορία του.

4.2 Σχεδιασμός Βάσης Δεδομένων

Η βάση δεδομένων που παρουσιάζεται εδώ αποτελεί την βάση της εφαρμογής του CitySense. Η βάση αυτή είναι σχεδιασμένη ώστε να υποστηρίζει το σύνολο των απαιτήσεων που τέθηκαν και επιπλέον να είναι εύκολα προσαρμόσιμη σε μελλοντικές απαιτήσεις και εισαγωγή νέων δεδομένων που ενδεχομένως να προκύψουν. Το αντίστοιχο ER διάγραμμα της βάσης δεδομένων απεικονίζεται στο παρακάτω σχήμα.


crimes_tracts_year_month_v2 [table]
tractce10 int4 not null
cyear float8(17, 17) not null
cmonth float8(17, 17) not null
total_crimes int4
crimes_tracts_year_v2 [table]
tractce10 int4 not null
cyear float8(17, 17) not null
total_crimes int4
tweets_tract_dow [table]
tractce10 int4 not null
cdow int2 not null
total_tweets int4
tweets_tract_hour [table]
tractce10 int4 not null
chour int2 not null
total_tweets int4
chicago_tracts_health [table]
tract int4 not null
tract_name text
community_id int4
name text
birth_rate float8(17, 17)
general_fertility_rate float8(17, 17)
low_birth_weight float8(17, 17)
prenatal_care_beginning_in_first_trimester float8(17, 17)
preterm_births float8(17, 17)
teen_birth_rate float8(17, 17)
assault_homicide float8(17, 17)
breast_cancer_in_females float8(17, 17)
cancer_all_sites float8(17, 17)
colorectal_cancer float8(17, 17)
diabetes_related float8(17, 17)
firearm_related float8(17, 17)
infant_mortality_rate float8(17, 17)
lung_cancer float8(17, 17)
prostate_cancer_in_males float8(17, 17)
stroke_cerebrovascular_disease float8(17, 17)
childhood_blood_lead_level_screening float8(17, 17)
childhood_lead_poisoning float8(17, 17)
gonorrhea_in_females float8(17, 17)
gonorrhea_in_males float8(17, 17)
tuberculosis float8(17, 17)
below_poverty_level float8(17, 17)
crowded_housing float8(17, 17)
dependency float8(17, 17)
no_high_school_diploma float8(17, 17)
per_capita_income float8(17, 17)
unemployment float8(17, 17)
stusam text
sumlev int4
country int4
cousub int4
place int4
geoid2 text
chicago_census_tracts [table]
gid serial not null auto-incremented
statefp10 varchar(254)
countyfp10 varchar(254)
tractce10 varchar(254)
namelsad10 varchar(254)
commarea varchar(254)
geoid10 varchar(254)
commarea_n numeric(131089, 0)
name10 varchar(254)
notes varchar(254)
geom geometry
tract int4 not null
chicago_tracts_se [table]
tract int4 not null
tract_name text
community_id int4
name text
percent_of_housing_crowded float8(17, 17)
percent_households_below_poverty float8(17, 17)
percent_aged_16_plus_unemployed float8(17, 17)
percent_aged_25_plus_no_high_school_diploma float8(17, 17)
percent_aged_under_18_or_over_64 float8(17, 17)
per_capita_income int4
hardship_index int4
stusam text
sumlev int4
country int4
cousub int4
place int4
geoid2 text
crimes_tracts_year_dow_v2 [table]
tractce10 int4 not null
cyear float8(17, 17) not null
cdow float8(17, 17) not null
total_crimes int4
crimes_tracts_year_hour_v2 [table]
tractce10 int4 not null
cyear float8(17, 17) not null
chour float8(17, 17) not null
total_crimes int4
checkins_tract_category_dow [table]
tractce10 int4 not null
category_id int2 not null
cdow int2 not null
total_checkins int4
checkins_tract_category_hour [table]
tractce10 int4 not null
category_id int2 not null
chour int2 not null
total_checkins int4
foursquare_categories_lv1 [table]
id serial not null auto-incremented
level1_hash text
level1_name text
pois_blocks_tracts [table]
poi_id int8 not null
level1_id int4
block_id int8 not null
tractce10 int4
pois [table]
poi_poi_id bigserial not null auto-incremented
poi_crawl_id int8 not null
poi_crawler_id int8 not null
poi_category_id int8 not null
poi_original_service_id varchar(256)
poi_created_time timestamp
poi_name varchar(256)
poi_description varchar(512)
poi_address varchar(512)
poi_telephones varchar(512)
poi_open_time varchar(2048)
poi_original_url varchar(2048)
poi_average_rating float4(8, 8)
poi_number_ratings int4
poi_license varchar(128)
poi_types _varchar
poi_categories _varchar
poi_tags _varchar
poi_geometry geometry
level1_id int4

Στη συνέχεια ακολουθεί μια αναλυτική περιγραφή της χρήσης και των δεδομένων για όλους τους πίνακες της βάσης δεδομένων.

chicago_census_tracts: Ο πίνακας αυτός περιέχει όλα τα tracts του Chicago που υποστηρίζει η εφαρμογή. Αναλυτικότερα, περιέχει το μοναδικό αναγνωριστικό, το όνομα, συνοπτικές πληροφορίες, τη γεωμετρία του πολυγώνου που περικλείει το tract.

crimes_tracts_year_v2: Ο πίνακας περιέχει τα δεδομένα εγκλημάτων που χρησιμοποιεί η εφαρμογή. Στον πίνακα αποθηκεύεται το πλήθος των εγκλημάτων, ανά tract σε χρονική ανάλυση έτους.

crimes_tracts_year_month_v2: Ο πίνακας περιέχει τα δεδομένα εγκλημάτων που χρησιμοποιεί η εφαρμογή. Στον πίνακα αποθηκεύεται το πλήθος των εγκλημάτων, ανά tract σε χρονική ανάλυση μήνα.

crimes_tracts_year_dow_v2: Ο πίνακας περιέχει τα δεδομένα εγκλημάτων που χρησιμοποιεί η εφαρμογή. Στον πίνακα αποθηκεύεται το πλήθος των εγκλημάτων, ανά tract σε χρονική ανάλυση ημέρας της εβδομάδας.

crimes_tracts_year_hour_v2: Ο πίνακας περιέχει τα δεδομένα εγκλημάτων που χρησιμοποιεί η εφαρμογή. Στον πίνακα αποθηκεύεται το πλήθος των εγκλημάτων, ανά tract σε χρονική ανάλυση ώρας της ημέρας.

chicago_tracts_health: Ο συγκεκριμένος πίνακας χρησιμοποιείται για να αποθηκεύσει τα δεδομένα των δεικτών υγείας που χρησιμοποιεί η εφαρμογή. Αναλυτικότερα, περιέχει τους δείκτες βρεφικής θνησιμότητας, πρόωρων γεννήσεων, γονιμότητας κλπ., ανά tract.

chicago_tracts_se: Ο συγκεκριμένος πίνακας χρησιμοποιείται για να αποθηκεύσει τα δεδομένα των κοινωνικό-οικονομικών δεικτών που χρησιμοποιεί η εφαρμογή. Αναλυτικότερα, περιέχει το ύψος του κατα κεφαλήν εισοδήματος, ποσοστά φτώχειας, το μέγεθος της ανεργίας κλπ., ανά tract.

tweets_tract_dow: Ο πίνακας περιέχει τα δεδομένα των γεωχωρικών tweets που χρησιμοποιεί η εφαρμογή. Στον πίνακα αποθηκεύεται το πλήθος των tweets, ανά tract σε χρονική ανάλυση ημέρας της εβδομάδας.

tweets_tract_hour: Ο πίνακας περιέχει τα δεδομένα των γεωχωρικών tweets που χρησιμοποιεί η εφαρμογή. Στον πίνακα αποθηκεύεται το πλήθος των tweets, ανά tract σε χρονική ανάλυση ώρας της ημέρας.

checkins_tract_category_dow: Ο πίνακας περιέχει τα δεδομένα των γεωχωρικών tweets που αφορούν checkin σε κάποιο σημείο ενδιαφέροντος. Στον πίνακα αποθηκεύεται το πλήθος των checkins, η κατηγορία των σημείων ενδιαφέροντος, ανά tract σε χρονική ανάλυση ημέρας της εβδομάδας.

checkins_tract_category_hour: Ο πίνακας περιέχει τα δεδομένα των γεωχωρικών tweets που αφορούν checkin σε κάποιο σημείο ενδιαφέροντος. Στον πίνακα αποθηκεύεται το πλήθος των checkins, η κατηγορία των σημείων ενδιαφέροντος, ανά tract σε χρονική ανάλυση ώρας της ημέρας.

pois_blocks_tracts: Ο συγκεκριμένος πίνακας χρησιμοποιείται για να αποθηκεύσει τα σημεία ενδιαφέροντος που χρησιμοποιεί η εφαρμογή. Οι πληροφορίες που αποθηκεύονται για κάθε σημείο ενδιαφέροντος είναι το μοναδικό αναγνωριστικό, η κατηγορία του, το block και το tract, στο οποίο ανήκει.

pois: Ο συγκεκριμένος πίνακας χρησιμοποιείται για να αποθηκεύσει τα σημεία ενδιαφέροντος που χρησιμοποιεί η εφαρμογή. Οι πληροφορίες που αποθηκεύονται είναι το μοναδικό αναγνωριστικό του σημείου ενδιαφέροντος, το γεωμετρικό σημείο στο οποίο εντοπίζεται, η κατηγορία του, η διεύθυνση, οι ώρες λειτουργίας, τα στοιχεία επικοινωνίας, το tract στο οποίο βρίσκεται, οι βαθμολογίες των χρηστών κλπ. Παράλληλα σε αυτόν τον πίνακα αποθηκεύεται και το μοναδικό αναγνωριστικό των crawlers που χρησιμοποιήθηκαν για τη συλλογή των σημείων ενδιαφέροντος.

foursquare_categories_lv1: Ο συγκεκριμένος πίνακας χρησιμοποιείται για να προσδιορίζει τις κατηγορίες των σημείων ενδιαφέροντος που χρησιμοποιεί η εφαρμογή. Οι πληροφορίες που αποθηκεύονται είναι το μοναδικό αναγνωριστικό της κατηγορίας και το όνομά της.

5 Επίλογος και μελλοντική δουλειά

Βασικός στόχοστου έργου Citysense είναι η εκμετάλλευση όσο το δυνατόν περισσότερων δεδομένων που αφορούν μια συγκεκριμένη περιοχή, προκειμένου να έχουμε μια ρεαλιστική απεικόνιση του "ίχνους" της σε πολλαπλά επίπεδα. Στο παρών παραδοτέο περιγράψαμε λοιπόν τρία κυρίως αντικείμενα: α) Την εφαρμογή που αναπτύχθηκε (τον "Area-Profiler") που αντλεί δεδομένα σημείων ενδιαφέροντος (POIs) για μια γεωγραφική περιοχή από τα GooglePlacesAPI καιFoursquareAPI) και τη συλλογή των γεωχωρικών tweetsαπό το TwitterStreamingAPI,β) τα ανοικτά δεδομένα που υπάρχουν για τη συγκεκριμένη περιοχή και τα οποία μπορούμε να κατεβάσουμε στην ολότητά τους για τοπική επεξεργασία, είτε αυτά προέρχονται από επίσημες πηγές (την κρατική υπηρεσία της πόλης του Chicago), είτε από διαδικτυακούς τόπους (δεδομένα καιρού από το WeatherUnderground), καθώςκαι την επεξεργασίατων δεδομένων χαρτών που προέρχονται από τη γνωστή υπηρεσία πληθοπορισμού χαρτών OpenStreetMaps και γ) την αποθήκευση των δεδομένων του CitySense, σε μία βάση δεδομένων σχεδιασμένη, ώστε να καλύπτει τις απαιτήσεις που τέθηκαν για την ανάπτυξη της εφαρμογής.

Αναφορές

[1] Robert Geisberger, Peter Sanders, Dominik Schultes, Daniel Delling: Contraction Hierarchies: Faster and Simpler Hierarchical Routing in Road Networks. WEA 2008: 319-333

[2] I. Abraham, D. Delling, A. V. Goldberg, and R. F. Werneck. A hub-based labeling algorithm for shortest paths in road networks. In P. Pardalos and S. Rebennack, editors,Experimental Algorithms, volume 6630 of Lecture Notes in Computer Science, pages 230–241. Springer Berlin Heidelberg, 2011.

[3] I. Abraham, D. Delling, A. Fiat, A. V. Goldberg, and R. F. Werneck. Hldb: Location based services in databases. In SIGSPATIAL GIS. ACM, November 2012.

[4] A. Efentakis, C. Efstathiades, and D. Pfoser. Cold. revisiting hub labels on the database for large-scale graphs. In C. Claramunt, M. Schneider, R. C.-W. Wong, L. Xiong, W.-K.Loh, C. Shahabi, and K.-J. Li, editors, Advances in Spatial and Temporal Databases, volume 9239 of Lecture Notes in Computer Science, pages 22–39. Springer InternationalPublishing, 2015.

[5] A. Efentakis. Scalable public transportation queries on the database. In Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016,Bordeaux, France, March 15-16, 2016, Bordeaux, France, March 15-16, 2016., pages 527–538, 2016.

[6] T. Akiba, Y. Iwata, and Y. Yoshida. Fast exact shortest-path distance queries on large networks by pruned landmark labeling. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, USA, pages 349–360, 2013.