ABH Internship report

Data Analytics: Month 2

Shapefiles of OSM data for Arizona is imported in postgis using QGIS. Only shape files of Arizona POIs are used.

**ST\_GeomFromText**: creates geometry out of WKT represented object

**ST\_Buffer**: creates a geometry/geography that represents all points whose distance from this Geometry/geography is less than or equal to distance (if SRID is 4326, unit is degrees)

**ST\_Transform**: Returns a new geometry with its coordinates transformed to a different spatial reference system

**ST\_MakePoint**: makes point out two variables, for example lat and long

**ST\_Intersection:** returns intersection (point, polygon…) from two geometries

**ST\_IsValid:** Test if an ST\_Geometry value is well formed. For geometries that are invalid, the PostgreSQL NOTICE will provide details of why it is not valid

**Data cleaning:** In OSM data, benches (3700), public toilets (811), waste-baskets (2100), fountains(364), atm (136), shelter(1109) are deleted

-Outlier: poi in Las Vegas (but state is AZ)

Intersection of pois\_points and yelp buffered (11 m) points and osm name is not null: 3217 records

Intersection of pois\_points and yelp buffered (111 m) points and osm name is not null: 57272 records

Intersection of pois\_points and yelp buffered (11 m) points and osm name is not null and osm name= yelp name: 588 records

Intersection of pois\_points and buffered (111 m) yelp points and lower(name) are equal: 1460 records

Intersection of pois\_points and buffered (55 m) yelp points and fuzzy string match used: 2560 records

Intersection of pois\_points and buffered (111 m) yelp points, osm\_name is null and yelp categories is not null: 8581 records

Intersection of osm without osm name: 8594 points and 14542 polygons, matched 136 polygons and 135 points, threshold for fuzzy was 70, radius 111 m

Intersected with fuzzy names: 3278 points and 3278 polygons

-problem with categories in business Arizona table: 32458 distinct categories, cannot be manually standardized

Normalization of names

-convert all to lower case

-strip blank spaces

-remove diacritics

-replace special characters

Convex Hull of yelp data

Intersection between buffered convex hull of Yelp data and osm data is done:

There are 9484 intersected osm points with yelp

There are 16736 intersected osm polygons with yelp

In total: 26220 osm polygons and points located in CH of yelp data

Categories standardization

30 most frequent Yelp categories:

|  |  |
| --- | --- |
| **Yelp categories** | **OSM fclass** |
| "Mexican, Restaurants" |  |
| "Restaurants, Mexican" |  |
| "Nail Salons, Beauty & Spas" |  |
| "Beauty & Spas, Nail Salons" |  |
| "Hair Salons, Beauty & Spas" |  |
| "Beauty & Spas, Hair Salons" |  |
| "Automotive, Auto Repair" |  |
| "Financial Services, Banks & Credit Unions" |  |
| "Banks & Credit Unions, Financial Services" |  |
| "Auto Repair, Automotive" |  |
| "Heating & Air Conditioning/HVAC, Home Services" |  |
| "Coffee & Tea, Food" |  |
| "Real Estate, Home Services, Apartments" |  |
| "Apartments, Real Estate, Home Services" |  |
| "Food, Coffee & Tea" |  |
| "Beauty & Spas, Barbers" |  |
| "Real Estate, Apartments, Home Services" |  |
| "Restaurants, Pizza" |  |
| "Apartments, Home Services, Real Estate" |  |
| "Home Services, Heating & Air Conditioning/HVAC" |  |
| "Barbers, Beauty & Spas" |  |
| "Pizza, Restaurants" |  |
| "Home Services, Real Estate, Apartments" |  |
| "Home Services, Apartments, Real Estate" |  |
| "Chinese, Restaurants" |  |
| "Pest Control, Local Services" |  |
| "Health & Medical, Chiropractors" |  |
| "Restaurants, Chinese" |  |
| "Chiropractors, Health & Medical" |  |

In 30 most frequent categories, there are "5378" records.