In [40]:

```
import os
import _pickle as pickle
import pandas as pd
import numpy as np
from collections import defaultdict
import folium
from folium import plugins
from pandas import DataFrame, Series
#matplotlib inline
from elasticsearch import Elasticsearch, helpers
```

In [41]:

```
import json
issues_pulled = [json.loads(line) for line in open('SPM587SP18issues.json')]#Loading the json file of the issues created in the git repo
```

In [42]:

```
issues_df = DataFrame(issues_pulled)#Putting the issues into a panda dataframe
```

```
In [43]: issues_df #Printing issues_df
```

Out[43]:

	Author	State	closed_at	created_at	issue_number	labe
0	HSP18SCM50W	closed	2018-04-22	2018-04-20	475	[Category:Inquiry, DetectionPhase:Field, Origi...
1	HSP18SCM50W	closed	2018-04-22	2018-04-19	474	[Address:59 W Grand A Chicago IL 60654, Cate
2	YSP18SCM40K	closed	2018-04-14	2018-04-14	472	[Category:Inquiry, DetectionPhase:Field, Origi...
3	RSP18SCM19N	closed	2018-04-14	2018-04-14	470	[Category:Enhancemen DetectionPhase:Testing
4	CSP18SCM32L	closed	2018-04-13	2018-04-13	466	[Category:Inquiry, DetectionPhase:Field, Origi...
5	RSP18SCM19N	closed	2018-04-13	2018-04-13	464	[Address: 2400 N Linco Ave Chicago IL 60614,..
6	MSP18SCM65B	closed	2018-04-13	2018-04-13	461	[Category:Bug, DetectionPhase:Testing Origina...
7	JSP18SCM63J	closed	2018-04-13	2018-04-13	459	[]
8	YSP18SCM35Z	closed	2018-04-13	2018-04-13	454	[Category:Enhancemen DetectionPhase:Testing
9	YSP18SCM40K	closed	2018-04-13	2018-04-13	452	[Address:225 S Canal S Chicago IL 60606, Cate
10	ZSP18SCM44L	closed	2018-04-13	2018-04-13	449	[Category:Inquiry, DetectionPhase:Field, Origi...
11	DSP18SCM14S	closed	2018-04-13	2018-04-13	447	[Category:Inquiry, DetectionPhase:Field, Origi...
12	RSP18SCM25A	closed	2018-04-13	2018-04-13	445	[Category:Enhancemen DetectionPhase:Testing
13	MSP18SCM01M	closed	2018-04-13	2018-04-13	444	[Category:Enhancemen DetectionPhase:Testing
14	FSP18SCM78A	closed	2018-04-13	2018-04-13	442	[Category:Inquiry, DetectionPhase:Field, Origi...

...	Author	State	closed_at	created_at	issue_number	labe
15	HSP18SCM81C	closed	2018-04-13	2018-04-13	436	[Category:Bug, DetectionPhase:Testing Origina...
16	YSP18SCM71Z	closed	2018-04-13	2018-04-13	435	[Category:Enhancemen DetectionPhase:Testing
17	ASP18SCM05S	closed	2018-04-13	2018-04-13	434	[Category:Enhancemen DetectionPhase:Testing
18	PSP18SCM99P	closed	2018-04-13	2018-04-13	432	[Category:Enhancemen DetectionPhase:Testing
19	SSP18SCM41S	open	None	2018-04-13	430	[Category:Enhancemen DetectionPhase:Testing
20	VSP18SCM42K	closed	2018-04-13	2018-04-13	424	[Category:Inquiry, DetectionPhase:Field, Origi...
21	ASP18SCM05A	closed	2018-04-13	2018-04-13	422	[Category:Enhancemen DetectionPhase:Testing
22	SSP18SCM19P	closed	2018-04-13	2018-04-13	420	[Category:Bug, DetectionPhase:Testing Origina...
23	NSP18SCM35K	closed	2018-04-13	2018-04-13	415	[Category:Enhancemen DetectionPhase:Testing
24	PSP18SCM99P	closed	2018-04-13	2018-04-13	411	[Category:Enhancemen DetectionPhase:Testing
25	SSP18SCM10S	closed	2018-04-13	2018-04-13	410	[Category:Inquiry, DetectionPhase:Field, Origi...
26	TSP18SCM03A	closed	2018-04-13	2018-04-13	409	[Category:Enhancemen DetectionPhase:Testing
27	KSP18SCM22B	closed	2018-04-13	2018-04-13	404	[Category:Inquiry, DetectionPhase:Field, Origi...
28	PSP18SCM73A	closed	2018-04-13	2018-04-13	403	[Category:Enhancemen DetectionPhase:Testing
29	ASP18SCM22S	closed	2018-04-13	2018-04-13	401	[Category:Inquiry, DetectionPhase:Field, Origi...
...	...	...	...	...	...	...

	Author	State	closed_at	created_at	issue_number	labe
225	HSP18SCM69D	open	None	2018-04-09	31	[Address:119 NORTH WABASH, Category:Bu Detec...
226	HSP18SCM69D	open	None	2018-04-09	30	[Address:111 W JACKSON, Category:Enhancement ...
227	HSP18SCM69D	open	None	2018-04-09	29	[Address:23 S CLARK, Category:Bug, DetectionPh...
228	HSP18SCM69D	open	None	2018-04-09	28	[Address:1951 N WESTERN AVE, Category:Bug, Det...
229	HSP18SCM69D	open	None	2018-04-09	27	[Address:645 N MCCLURG CT, Category:Inquiry, D...
230	HSP18SCM69D	open	None	2018-04-09	26	[Address:1951 N WESTERN AVE, Category:Enhancem...
231	HSP18SCM69D	open	None	2018-04-09	25	[Address:645 N MCCLURG CT, Category:Bug, Detec...
232	HSP18SCM69D	open	None	2018-04-09	24	[Address:600 E GRAND AVE, Category:Inquiry, De...
233	HSP18SCM69D	open	None	2018-04-09	23	[Address:119 NORTH WABASH, Category:Enhancemen.
234	HSP18SCM69D	open	None	2018-04-09	22	[Address:233 W JACKSON, Category:Bug, Detectio
235	HSP18SCM69D	open	None	2018-04-09	21	[Address:111 W JACKSON, Category:Inquiry, Dete..
236	HSP18SCM69D	open	None	2018-04-09	20	[Address:119 NORTH WABASH, Category:Enhancemen.
237	HSP18SCM69D	open	None	2018-04-09	19	[Address:119 NORTH WABASH, Category:Bu Detec...

	Author	State	closed_at	created_at	issue_number	labe
238	HSP18SCM69D	open	None	2018-04-09	18	[Address:111 W JACKSON, Category:Inquiry, Dete..
239	HSP18SCM69D	open	None	2018-04-09	17	[Address:23 S CLARK, Category:Enhancement Det...
240	HSP18SCM69D	open	None	2018-04-09	16	[Address:23 S CLARK, Category:Inquiry, Detecti...
241	HSP18SCM69D	open	None	2018-04-09	15	[Address:2525 S Martin Luther King Drive, Cate
242	HSP18SCM69D	open	None	2018-04-09	14	[Address:1951 N WESTERN AVE, Category:Bug, Det...
243	HSP18SCM69D	open	None	2018-04-09	13	[Address:645 N MCCLURG CT, Category:Enhancemen..
244	HSP18SCM69D	open	None	2018-04-09	12	[Address:645 N MCCLURG CT, Category:Inquiry, D...
245	HSP18SCM69D	open	None	2018-04-09	11	[Address:600 E GRAND AVE, Category:Bug, Detect...
246	HSP18SCM69D	open	None	2018-04-09	10	[Address:233 W JACKSON, Category:Enhancement ...
247	HSP18SCM69D	open	None	2018-04-09	9	[Address:119 NORTH WABASH, Category:Enhancemen..
248	HSP18SCM69D	open	None	2018-04-09	8	[Address:119 NORTH WABASH, Category:Bu Detec...
249	HSP18SCM69D	open	None	2018-04-09	7	[Address:111 W JACKSON, Category:Inquiry, Dete..
250	HSP18SCM69D	open	None	2018-04-09	6	[Address:111 W JACKSON, Category:Bug, Detectio

	Author	State	closed_at	created_at	issue_number	labe
251	HSP18SCM69D	open	None	2018-04-09	5	[Address:23 S CLARK, Category:Enhancement Det...
252	HSP18SCM69D	open	None	2018-04-09	4	[Address:23 S CLARK, Category:Enhancement Det...
253	SPM587SP18	closed	2018-04-09	2018-04-08	3	[Address:2525 S Martin Luther King Drive, Cate
254	SPM587SP18	closed	2018-04-06	2018-03-30	2	[Address:2525 S Martin Luther King Drive, Cate

255 rows × 6 columns

In [44]:

```
wrangled_issues_df = issues_df[['Author','State','closed_at','created_a
t','issue_number','labels']]# as per the code given in the tutorial,filt
ering and arranging the Dataframe
wrangled_issues_df.loc[0:len(wrangled_issues_df), 'OriginationPhase']= n
p.NaN
wrangled_issues_df.loc[0:len(wrangled_issues_df),'DetectionPhase']= np.N
aN
wrangled_issues_df.loc[0:len(wrangled_issues_df),'Category']= np.NaN
wrangled_issues_df.loc[0:len(wrangled_issues_df),'Priority']= np.NaN
wrangled_issues_df.loc[0:len(wrangled_issues_df),'Status']= np.NaN
```

In [45]:

```
newList = list() #creating a new list
for i in range(0, len(wrangled_issues_df)):#Since in the json file, the
 labels are not in form of key value pair, but an array of string, they
 cannot be accessed
#thus the label part of dataframe is split into a new dictioanry of key
 value pair and updated into the new list.
    tempDictionary = dict()
    if wrangled_issues_df.iloc[i]['labels']:
        for label in wrangled_issues_df.iloc[i]['labels']:
            label_name= (label.split(':'))[0]
            label_value= (label.split(':'))[1]
            tempDictionary.update({label_name : label_value})
    tempDictionary.update({'issue_number' : int(wrangled_issues_df.iloc[
i]['issue_number'])})#Since the panda dataframe uses numpy iteger, casti
ng it into
    # primitive integer, as elastic search only accepts primitve data ty
pes.
    newList.append(tempDictionary)
```

In [46]:

```
#newList
```

In [47]:

```
actions = list() #updating elastic search database as per given
es = Elasticsearch()
for data in newList:
    action = {
        '_index':'issues_database',
        '_type':'gitRepo',
        '_id':data['issue_number'],
        '_source':data
    }
    actions.append(action)
helpers.bulk(es,actions)
```

Out[47]:

```
(255, [])
```

In [48]:

```
first_query = { #first query where all the issues are pulled from the da
tabase
    'size' : 500,
    'query' : {
        'match_all' : {}
    }
}
queried_output_first = es.search(index = 'issues', body=first_query,scro
ll='1h') #issues are stored in json format
```

In [80]:

```
queried_output_first['hits']['hits'][0]
```

Out[80]:

```
{'_id': '470',
 '_index': 'issues',
 '_score': 1.0,
 '_source': {'Category': 'Enhancement',
  'DetectionPhase': 'Testing',
  'OriginationPhase': 'Design',
  'Priority': 'Major',
  'Status': 'Completed',
  'issue_number': 470},
 '_type': 'shreyas'}
```

In [50]:

```
sid = queried_output_first['_scroll_id'] #As per the tutorial code, sett
ing the scroll size.
scroll_size = queried_output_first['hits']['total']
```

In [51]:

```
count = 0
first_query_coord = []
while(scroll_size > count):
    for doc in queried_output_first['hits']['hits']:  #As per the code i
n the tutorial, accessing the values in the key(data['hits']['hits'])
# This the key value format generated by the elastic search when pulled
 from the database. The value of this key contains all the various label
s.
        location_ll = []
        results = doc['_source']
        count = count +1
        if 'Latitude' in results.keys():
            if 'Longitude' in results.keys():
                if(results['Latitude'] != None and results['Longitude']
!= None):
                    location_ll.append(float(results['Latitude']))
                    location_ll.append(float(results['Longitude']))
                    first_query_coord.append(location_ll)
```

In [52]:

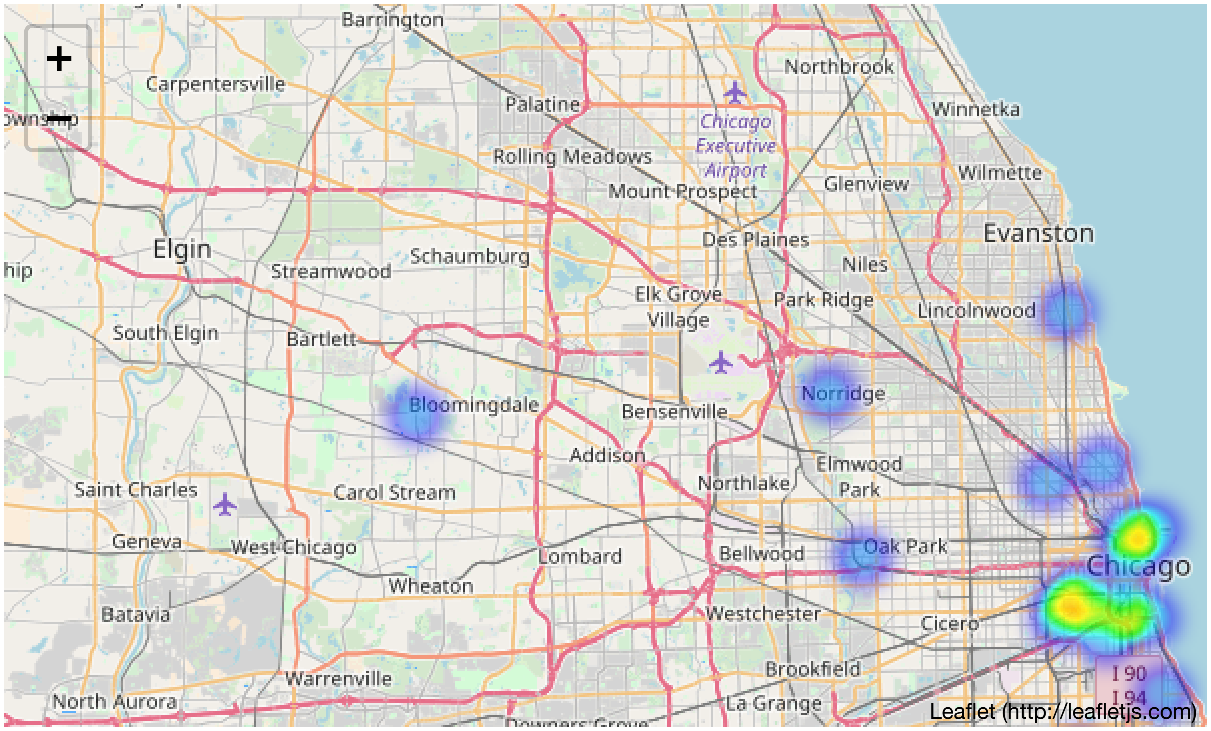
```
print(len(first_query_coord))
```

```
137
```

In [53]:

```
first_query_heat_map = folium.Map([41.891551, -87.607375],zoom_start = 1
6)
first_query_heat_map.add_child(plugins.HeatMap(first_query_coord,radius=
15))
```

Out[53]:



In [54]:

```
second_query = { #second query to match the labels given
    'size' : 500,
    'query' : {
        'bool':{
            'must' : [{'match':{'DetectionPhase':'Field'}},
                      {'match':{'Priority':'Critical'}}]
        }
    }
}
queried_output_second = es.search(index = 'issues', body=second_query,sc
roll='1h')#This is a dictionary variable. Pulls data and stores as key v
alue pair in the dictionary.
```

In [55]:

```
sid = queried_output_second['_scroll_id']
scroll_size = queried_output_second['hits']['total']
```

In [56]:

```
count = 0 #As per the tutorial, accessing the Latitude and Longitude val
ues from key, and storing them into an array. The if else condition chec
ks
#if the issues have those values or not, if not then that issue is skipp
ed.
second_query_coord = []
while(scroll_size > count):
    for doc in queried_output_second['hits']['hits']:
        location_ll = []
        results = doc['_source']
        count = count +1
        if 'Latitude' in results.keys():
            if 'Longitude' in results.keys():
                if(results['Latitude'] != None and results['Longitude']
!= None):
                    location_ll.append(float(results['Latitude']))
                    location_ll.append(float(results['Longitude']))
                    second_query_coord.append(location_ll)
```

In [57]:

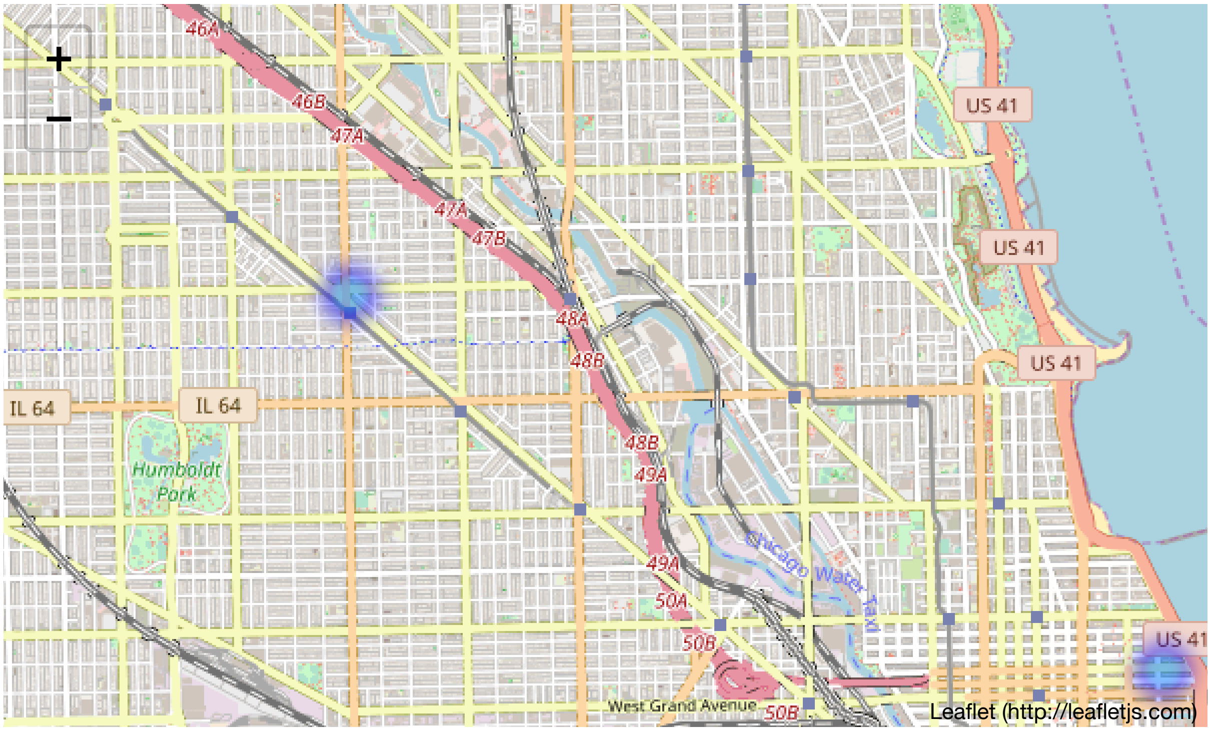
```
print(len(second_query_coord))
```

```
3
```

In [58]:

```
second_query_heat_map = folium.Map([41.891551, -87.607375],zoom_start =
16) #Syntax of folium map. Setting the default location tp view.
second_query_heat_map.add_child(plugins.HeatMap(second_query_coord,radiu
s=15)) #plot the coordinates onto the map.
```

Out[58]:



In [59]:

```
third_query = { #Third query to match the given labels
    'size' : 500,
    'query' : {
        'bool':{
            'must' : [{'match':{'DetectionPhase':'Field'}},
                      {'match':{'Status':'Completed'}}]
        }
    }
}
queried_output_third = es.search(index = 'issues', body=third_query,scro
ll='1h')#This is a dictionary variable. Pulls data and stores as key val
ue pair in the dictionary.
```

In [60]:

```
sid = queried_output_third['_scroll_id']
scroll_size = queried_output_third['hits']['total']
```

In [61]:

```
count = 0
third_query_coord = []
while(scroll_size > count):
    for doc in queried_output_third['hits']['hits']:
        location_ll = []
        results = doc['_source']
        count = count +1
        if 'Latitude' in results.keys():
            if 'Longitude' in results.keys():
                if(results['Latitude'] != None and results['Longitude']
!= None):
                    location_ll.append(float(results['Latitude']))
                    location_ll.append(float(results['Longitude']))
                    third_query_coord.append(location_ll)
```

In [62]:

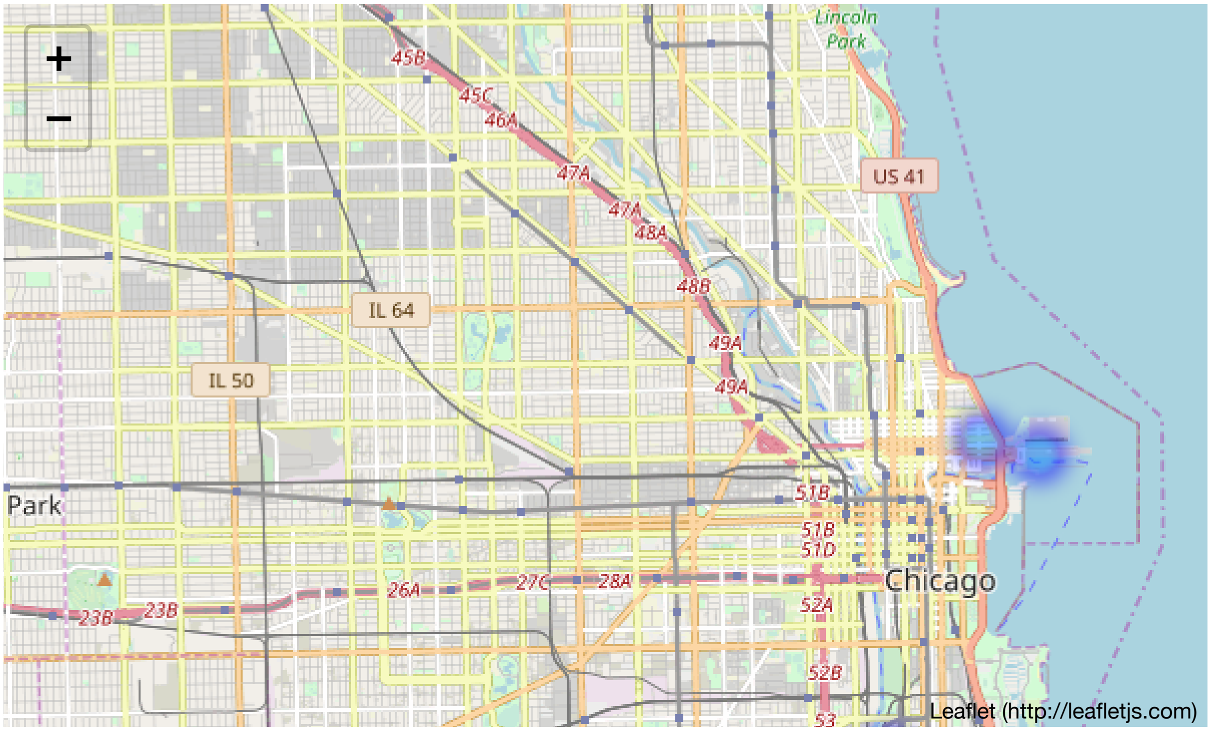
```
print(len(third_query_coord))
```

```
4
```

In [63]:

```
third_query_heat_map = folium.Map([41.891551, -87.607375],zoom_start = 1
6)
third_query_heat_map.add_child(plugins.HeatMap(third_query_coord,radius=
15))
```

Out[63]:



In [64]:

```
fourth_query = {#Fourth query to match the given labels
    'size' : 500,
    'query' : {
        'bool':{
            'must' : [{'match':{'DetectionPhase':'Field'}},
                      {'match':{'Priority':'Critical'}},
                      {'match':{'Status':'Approved'}}]
        }
    }
}
queried_output_fourth = es.search(index = 'issues', body=fourth_query,sc
roll='1h')#This is a dictionary variable. Pulls data and stores as key v
alue pair in the dictionary.
```

In [65]:

```
sid = queried_output_fourth['_scroll_id']
scroll_size = queried_output_fourth['hits']['total']
```

In [66]:

```
count = 0
fourth_query_coord = []
while(scroll_size > count):
    for doc in queried_output_fourth['hits']['hits']:
        location_ll = []
        results = doc['_source']
        count = count +1
        if 'Latitude' in results.keys():
            if 'Longitude' in results.keys():
                if(results['Latitude'] != None and results['Longitude']
!= None):
                    location_ll.append(float(results['Latitude']))
                    location_ll.append(float(results['Longitude']))
                    fourth_query_coord.append(location_ll)
```

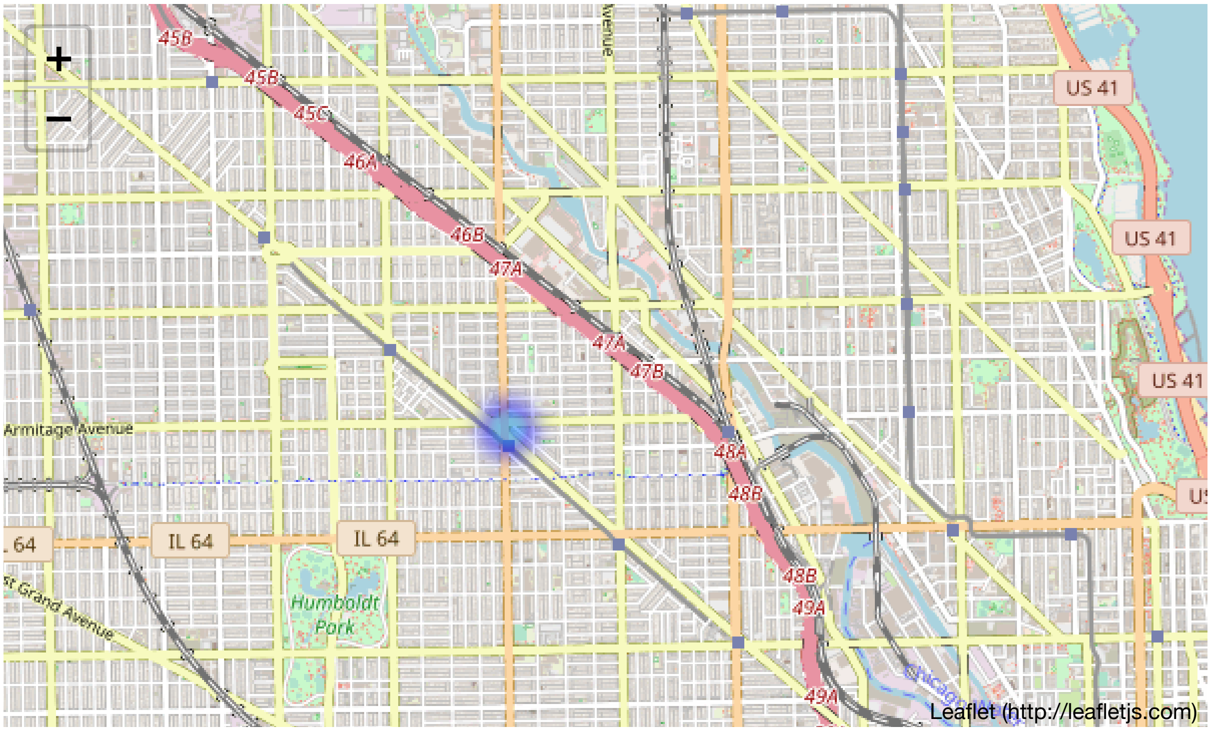
In [67]:

```
print(len(fourth_query_coord))
```

```
2
```

In [68]:
```
fourth_query_heat_map = folium.Map([41.891551, -87.607375],zoom_start = 16)
fourth_query_heat_map.add_child(plugins.HeatMap(fourth_query_coord,radius=15))
```

Out[68]:



In [69]:
```
fifth_query = {#Fifth query to match the given labels
    'size' : 500,
    'query' : {
        'bool':{
            'must' : [{'match':{'DetectionPhase':'Field'}},
                      {'match':{'Priority':'Critical OR High'}},
                      {'match':{'Status':'Approved OR inProgress'}}]
    }
}
}
queried_output_fifth = es.search(index = 'issues', body=fifth_query,scroll='1h')#This is a dictionary variable. Pulls data and stores as key value pair in the dictionary.
```

In [70]:
```
sid = queried_output_fifth['_scroll_id']
scroll_size = queried_output_fifth['hits']['total']
```

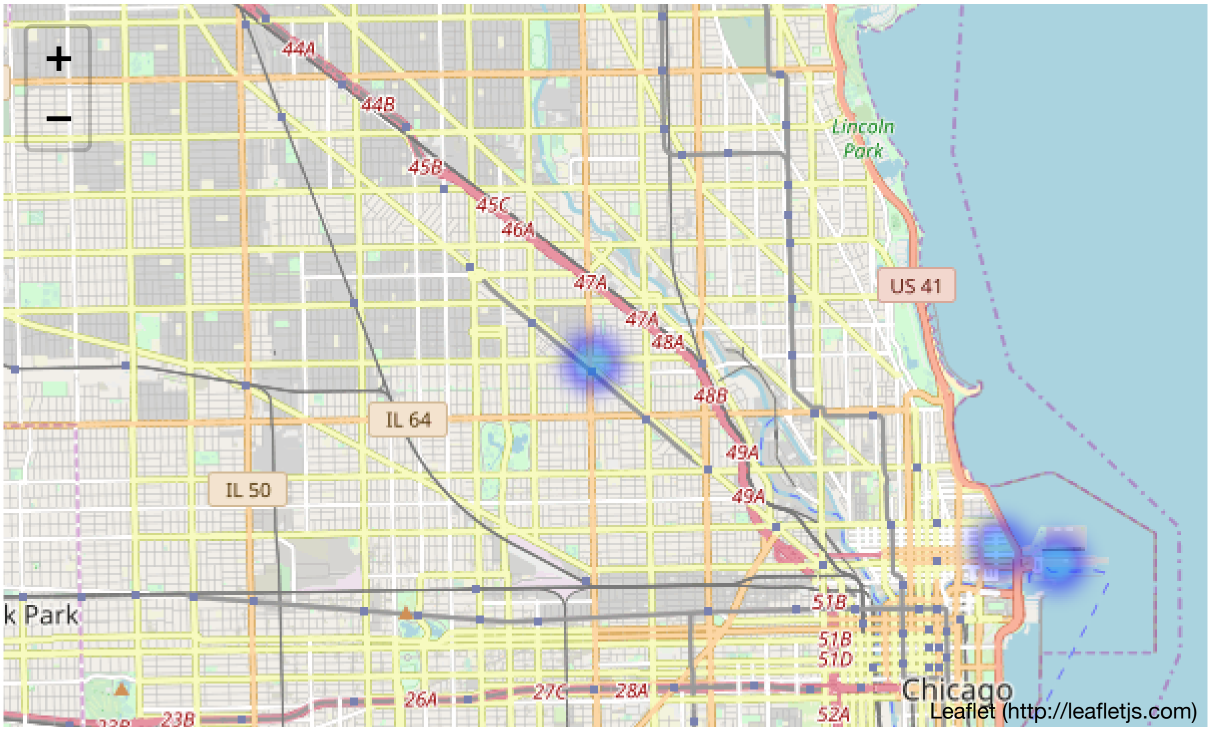
In [71]:
```
count = 0
fifth_query_coord = []
while(scroll_size > count):
    for doc in queried_output_fifth['hits']['hits']:
        location_ll = []
        results = doc['_source']
        count = count +1
        if 'Latitude' in results.keys():
            if 'Longitude' in results.keys():
                if(results['Latitude'] != None and results['Longitude']
!= None):
                    location_ll.append(float(results['Latitude']))
                    location_ll.append(float(results['Longitude']))
                    fifth_query_coord.append(location_ll)
```

In [72]:
```
print(len(fifth_query_coord))
```

```
4
```

In [73]:
```
fifth_query_heat_map = folium.Map([41.891551, -87.607375],zoom_start = 1
6)
fifth_query_heat_map.add_child(plugins.HeatMap(fifth_query_coord,radius=
15))
```

Out[73]:



In [74]:

```
sixth_query = { #Sixth query to match the given labels
    'size' : 500,
    'aggs' : {
        'data' : {
            'terms' : {
                'field' : 'Latitude.keyword',
                'field' : 'Longitude.keyword',
                'min_doc_count' : 5,
                'size' : 500
            },
            'aggs' : {
                'top_selection' : {
                    'top_hits' : {
                        'size' : 10
                    }
                }
            }
        }
    }
}
queried_output_sixth = es.search(index = 'issues', body=sixth_query,scro
ll='1h') #This is a dictionary variable. Pulls data and stores as key va
lue pair in the dictionary.
```

In [75]:

```
sid = queried_output_sixth['_scroll_id']
scroll_size = queried_output_sixth['hits']['total']
```

In [76]:

```
count = 0
sixth_query_coord = []
for i in queried_output_sixth['aggregations']['data']['buckets']:
        location_ll = []
        results = i['top_selection']['hits']['hits'][0]['_source']
        if 'Latitude' in results.keys():
               if 'Longitude' in results.keys():
                if(results['Latitude'] != None and results['Longitude']
!= None):
                    location_ll.append(float(results['Latitude']))
                    location_ll.append(float(results['Longitude']))
                    sixth_query_coord.append(location_ll)
```

In [77]:

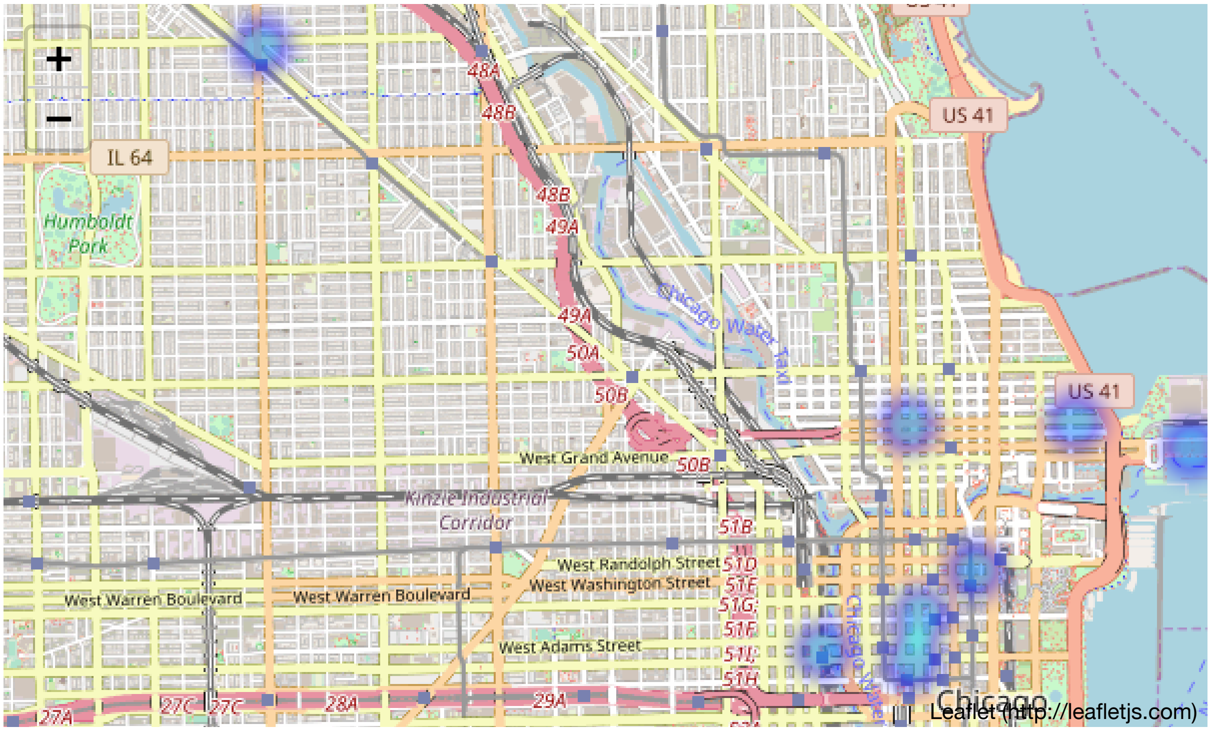
```
print(len(sixth_query_coord))
```

```
14
```

In [78]:

```
sixth_query_heat_map = folium.Map([41.891551, -87.607375],zoom_start = 16)
sixth_query_heat_map.add_child(plugins.HeatMap(sixth_query_coord,radius=15))
```

Out[78]:


+
−
IL 64
Humboldt Park
48A
48B
48B
49A
49A
50A
50B
50B
US 41
US 41
Chicago Water Taxi
West Grand Avenue
Kinzie Industrial Corridor
51B
West Randolph Street
West Washington Street
51D
51E
51G
51F
51L
51H
West Warren Boulevard
West Warren Boulevard
West Adams Street
27A
27C
27C
28A
29A
Chicago Water
Chicago
Leaflet (http://leafletjs.com)