The Battle of the Neighborhoods

Timur Sahin

1 Introduction and Business Problem

I have chosen a specific hypothetical situation wherein my spouse is relocating for work reasons. In this scenario, they are moving from their favorite city in the United States, Boston, Massachusetts to a completely new city in Europe: Stockholm, Sweden. My goal is to use the Foursquare API as well as other publicly available data to determine which neighborhoods in which cities are the most similar to my spouse's favorite neighborhood in Boston: Davis Square.

I hope to both answer this specific question, but also to produce a more generic framework for the automated clustering of neighborhoods in two distinct cities that can recommend similar neighborhoods across geographical boundaries. The scope of the project is to build a framework that takes in a source city, a destination city, and some input or measure of priority of factors (such as similarity of nearby venues, climate data, public transportation data, housing data, etc), do some unsupervised learning on that data, and then produce mapping or categorization of similar neighborhoods across both cities. This project will use Boston, Massachusetts as a source city and Stockholm, Sweden as a destination city, but will be written to allow for other inputs (provided they are on the Foursquare API).

2 Data

Data from ARCGIS, Geoarkivet, and OpenStreetMaps will be used to determine the count and geographic centers of the neighborhoods in Boston and in various cities in Sweden. The Foursquare Places API will be used to collect information about surrounding venues of interest. I will explore whether or not there are meaningful climate differences with NOAA, though I don't expect this to be materially different within each city.

2.1 ARCGIS Neighborhood Boundary Data for Boston, MA

Link: ARCGIS Neighborhood Boundary Data for Boston, MA

Showing 1 to 10 of 26 — Hint: Filter columns using

OBJECTID	Name	Acres	Neighborhood_ID	SqMiles	Shape.STArea()	Shape.STLength()
27	Roslindale	1605.5682375	15	2.51	69938272.6743164	53563.912597056624
28	Jamaica Plain	2519.24539377	11	3.94	109737890.39697266	56349.93716141023
29	Mission Hill	350.8535636	13	0.55	15283120.097167969	17918.724113458415
30	Longwood	188.61194672	28	0.29	8215903.537109375	11908.757147546603
31	Bay Village	26.53983916	33	0.04	1156070.7705078125	4650.635493295902
32	Leather District	15.63990811	27	0.02	681271.671875	3237.140536982458
33	Chinatown	76.32440999	26	0.12	3324678	9736.590412617801
34	North End	126.91043901	14	0.2	5527505.988525391	16177.82681542302
35	Roxbury	2108.46907176	16	3.29	91844545.38745117	49488.80048473105
36	South End	471.53535561	32	0.74	20539997.932617188	17912.3335694887

FIGURE 1: Example of ARCGIS Open Data for Boston Neighborhoods

ARCGIS Data is free data that contains geographic information for most major places in the United States and will be used to get a list of Neighborhoods and their geographical coordinates. This will be mostly useful in the exploratory phase.

2.2 Stockholm Geoarkivet

Link: Stockholm Geoarkivet

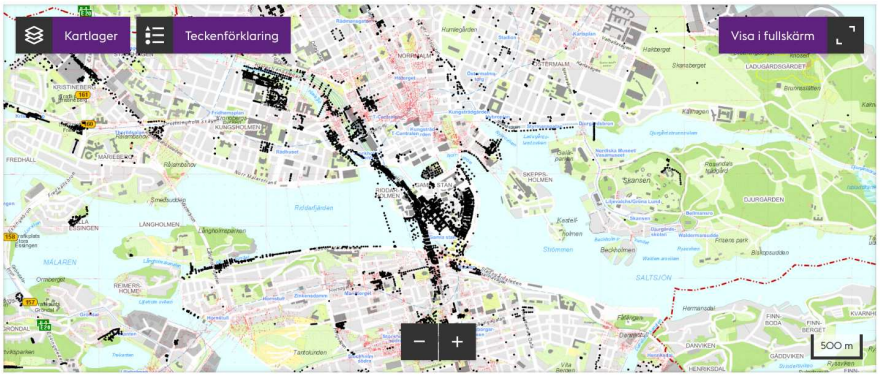

FIGURE 2: Geoarkivet Coverage for Stockholm, Sweden

Geoarkivet has a wealth of geographic data pertaining to Sweden and especially to Stockholm in particular, available in a machine-readable format. Like ARCGIS, I will use this information to extract neighborhood information, but also for exploratory analysis.

2.3 Nominatim OpenStreetMaps

Link: Nominatim OpenStreetMaps API


```
{
  "type": "FeatureCollection",
  "licence": "Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright",
  "features": [
    {
      "type": "Feature",
      "properties": {
        "place_id": "35811445",
        "osm_type": "node",
        "osm_id": "2846295644",
        "display_name": "17, Strada Pictor Alexandru Romano, Bukarest, Bucharest, Sector 2, Buc
        "place_rank": "30",
        "category": "place",
        "type": "house",
        "importance": 0.62025
      },
      "bbox": [
        26.1156689,
        44.4354754,
        26.1157689,
        44.4355754
      ],
      "geometry": {
        "type": "Point",
        "coordinates": [
          26.1157189,
          44.4355254
        ]
      }
    }
  ]
}
```


FIGURE 3: Sample GeoJSON response from Nominatim

Nominatim is an OpenStreetMap API that returns geographic data in a convenient JSON or GeoJSON format (pictured above), which includes a rich dataset of open-sourced geographical information in a very easy-to-query API. I'll use this as the chief source of graphic data after I'm done cross-referencing with the ARCGIS and Geoarkivet data because of its ease of use. It will principally help me get the latitude and longitude information consistently and accurately for all the neighborhoods discovered in the other datasets.

2.4 Foursquare Places API

Link: Foursquare Places API

Out[15]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Parkwoods	43.753259	-79.329656	Corrosion Service Company Limited	43.752432	-79.334661	Construction & Landscaping
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

FIGURE 4: Sample data enriched with Foursquare Venue Information

Foursquare Places API will be the main API used to pull venue information to determine similarity and compute distance metrics between the neighborhoods in our data sets, using either k-means or DBSCAN clustering methods.

2.5 NOAA Climate Data Web Services

Link: NOAA Climate Data Web Services API


```
{
    "date": "2010-05-01T00:00:00",
    "datatype": "PRCP",
    "station": "GHCND:USC00010008",
    "attributes": ",,,0",
    "value": 6.6
},
{
    "date": "2010-05-01T00:00:00",
    "datatype": "SNOW",
    "station": "GHCND:USC00010008",
    "attributes": ",,,0",
    "value": 0
}
```


FIGURE 5: Sample NOAA summary precipitation and snow JSON data

While I don't anticipate finding any meaningful distinctions in weather patterns between neighborhoods of a city, I want to include it in this dataset as a "canary" to confirm that my model can accurately recognize this data is not meaningful. I have written several functions to grab and compare historical weather data from NOAA and am expecting to find that it is not useful in clustering or classifying.