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

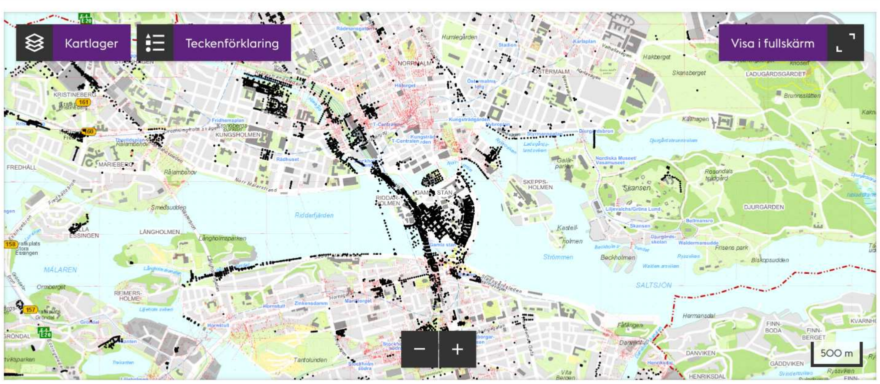

FIGURE 2: Geoarkivet Coverage for Stockholm, Sweden

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.

3 Methodology

We will begin first by parsing publicly available geographic data for Boston (via ARCGIS) and Stockholm (via Geoarkivet) to grab a simple list of unique neighborhood names. For this purpose I have used the "squares" of Boston and the "districts" of Stockholm. Neither squares nor districts are political boundaries: they simply reflect how the residents of the respective cities have organized their sense of local communities.

	City	Neighborhood	Latitude	Longitude
0	Stockholm, Sweden	Abrahamsberg	59.336468	17.953763
1	Stockholm, Sweden	Alvik	59.333401	17.982279
2	Stockholm, Sweden	Älvsjö	59.275849	18.001890
3	Stockholm, Sweden	Aspudden	59.306466	18.001373
4	Stockholm, Sweden	Axelsberg	59.304325	17.974414
...	...	...	...	...
63	Stockholm, Sweden	Svedmyra	59.277491	18.067128
64	Stockholm, Sweden	Tallkrogen	59.271529	18.086059
65	Stockholm, Sweden	Vårberg	59.275887	17.889953
66	Stockholm, Sweden	Vasastan	59.341350	18.048947
67	Stockholm, Sweden	Västertorp	59.291315	17.966692

67 rows × 4 columns

	City	Neighborhood	Latitude	Longitude
1	Boston, MA	Boston City Hall Plaza	42.360401	-71.057682
2	Boston, MA	Bowdoin Square	42.361394	-71.062120
3	Boston, MA	Brigham Circle	42.334608	-71.103692
4	Boston, MA	Central Square	42.375020	-71.039378
6	Boston, MA	Copley Square	42.349990	-71.076440
7	Boston, MA	Day Square	42.379203	-71.027788
8	Boston, MA	Dewey Square	42.352984	-71.055554
9	Boston, MA	Dock Square	42.360933	-71.054914
10	Boston, MA	Edward Everett Square	42.320174	-71.061451
11	Boston, MA	Franklin Square	42.338815	-71.072554
12	Boston, MA	Haymarket Square	42.362950	-71.057845
13	Boston, MA	Kenmore Square	42.348952	-71.096254

FIGURE 5: Neighborhoods in Stockholm and Boston augmented with Geocoordinate Data

We then use the Folium library to map the different neighborhoods:

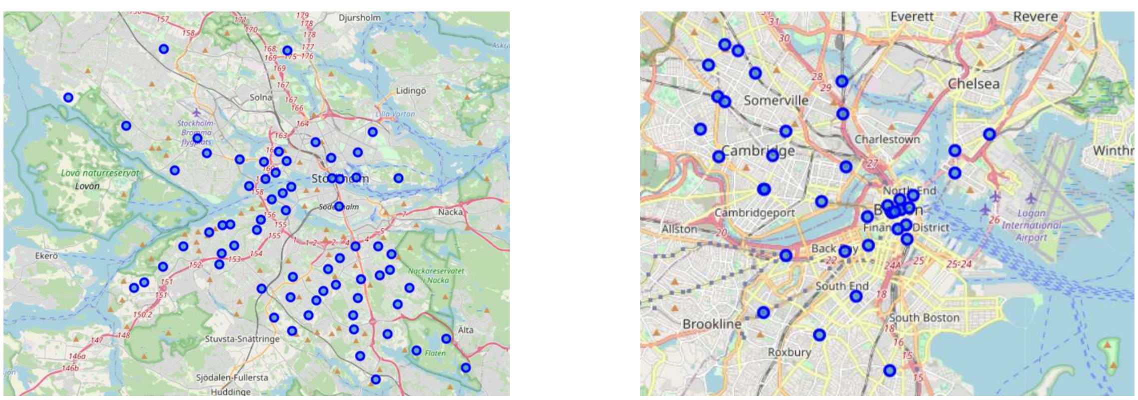

FIGURE 6: Neighborhoods in Stockholm and Boston augmented with Geocoordinate Data

Already we can see that the 'neighborhoods' in Boston are much more densely packed, and indeed it may make sense to drop a few of the closest packed squares. Naively at this point I'm expecting any similarity analysis or clustering algorithm to see more homogeneity in Boston than in Stockholm, but we'll confirm this in the final steps of the analysis.

We'll use the techniques we learned earlier in the unit to utilize the Foursquare API to grab interesting venue data in each neighborhood in each city!

	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Stockholm, Sweden	Abrahamsberg	59.336468	17.953763	Abrahamsberg Sushi	59.336311	17.952846	Japanese Restaurant
1	Stockholm, Sweden	Abrahamsberg	59.336468	17.953763	Abrahamsbergsvägen / Drottningholmsvägen	59.337151	17.952862	Intersection
2	Stockholm, Sweden	Abrahamsberg	59.336468	17.953763	ICA Abrahamsberg	59.336172	17.952363	Grocery Store
3	Stockholm, Sweden	Abrahamsberg	59.336468	17.953763	Bonne Femme	59.336121	17.952388	Bakery
4	Stockholm, Sweden	Abrahamsberg	59.336468	17.953763	Abrahamsberg T-Bana	59.336617	17.953475	Metro Station

	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Boston, MA	Boston City Hall Plaza	42.360401	-71.057682	Saus Restaurant	42.361076	-71.057054	Belgian Restaurant
1	Boston, MA	Boston City Hall Plaza	42.360401	-71.057682	Wachusett Boston Brew Yard at City Hall	42.359788	-71.057296	Beer Garden
2	Boston, MA	Boston City Hall Plaza	42.360401	-71.057682	The New England Holocaust Memorial	42.361059	-71.057179	Historic Site
3	Boston, MA	Boston City Hall Plaza	42.360401	-71.057682	Food Trucks at City Hall	42.359678	-71.057599	Street Food Gathering
4	Boston, MA	Bowdoin Square	42.361394	-71.062120	Target	42.361032	-71.062670	Convenience Store

FIGURE 7: Examples of venues in neighborhoods in Stockholm and Boston

We transform the data using one-hot encoding in order to build a matrix of the frequency of each category of venue in each neighborhood. We can peak at this data to get a sense of the most popular kinds of place in each neighborhood, as shown:

	City	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Boston, MA	Boston City Hall Plaza	Historic Site	Belgian Restaurant	Street Food Gathering	Beer Garden	Yoga Studio
1	Boston, MA	Bowdoin Square	Japanese Restaurant	Donut Shop	Bar	Convenience Store	Dessert Shop
2	Boston, MA	Brigham Circle	Convenience Store	Pub	Sandwich Place	Sushi Restaurant	Food Truck
3	Boston, MA	Central Square	Pharmacy	Liquor Store	Fast Food Restaurant	Sandwich Place	Wings Joint
4	Boston, MA	Copley Square	Plaza	Bagel Shop	Farmers Market	Lounge	Gym
...	...	...	...	...	...	...	...
73	Stockholm, Sweden	Södermalm	Pizza Place	Indian Restaurant	Vietnamese Restaurant	Gym	Yoga Studio
74	Stockholm, Sweden	Tallkrogen	Metro Station	Indian Restaurant	Yoga Studio	Farmers Market	Food Truck
75	Stockholm, Sweden	Vasastan	Sushi Restaurant	Pizza Place	Pool Hall	Grocery Store	Gym / Fitness Center
76	Stockholm, Sweden	Västertorp	Metro Station	Yoga Studio	Farmers Market	Food Truck	Food
77	Stockholm, Sweden	Vårberg	Shopping Mall	Metro Station	Flea Market	Grocery Store	Yoga Studio

FIGURE 8: The most popular venue types in each neighborhood

We're going to use k-means analysis to cluster together similar neighborhoods across cities. In order to choose a value for k we're going to fit our data and plot the distortion, and look for the 'elbow.'

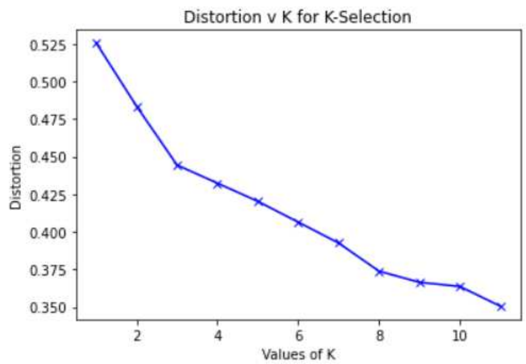

FIGURE 9: Selecting a value of k for k-means clustering

4 Results and Discussion

We use the more prominent kink in the graph at the value k=8 though it's not much of an elbow, and we'll explore other clustering algorithms in the future. We can see eight clusters, of which five are quite prominent. We map them side-by-side between Stockholm and Boston. As we hypothesized, Boston has much less diversity than Stockholm, likely owing to the lower mean distance between neighborhoods.

The matter of answering our initial question then becomes somewhat simplified. We want to know which neighborhoods in Stockholm are most similar to our desired neighborhood in the origin city, which in this case was Davis Square in Somerville in the Boston area. It's easy to query our dataframe for all neighborhoods in Stockholm that match the same cluster as Davis Square in Boston:

- Alvik, Stockholm
- Gamla stan, Stockholm
- Johanneshov, Stockholm
- Långholmen, Stockholm
- Liljeholmen, Stockholm
- Norrmalm, Stockholm
- Riddarholmen, Stockholm
- Skeppsholmen, Stockholm
- Södermalm, Stockholm

It's interesting that downtown Stockholm most closely resembles the neighborhoods of Boston, which is a much more dense city. This is what I would conventionally expect. One interesting thing is that the relative popularity of public transit in Stockholm cleaved off several neighborhoods into the red cluster. These are areas where the most popular attraction by far is the metro station, likely neighborhoods popular with commuters. By contrast, Boston proper had no neighborhoods where the Metro was the most popular venue. Another interesting cluster was teal, consisting of three neighborhoods (two in Boston and one in Stockholm) with were marked by parks, yoga studios, farmers' markets, and food trucks. I think colloquially we would call these "hipster" neighborhoods.

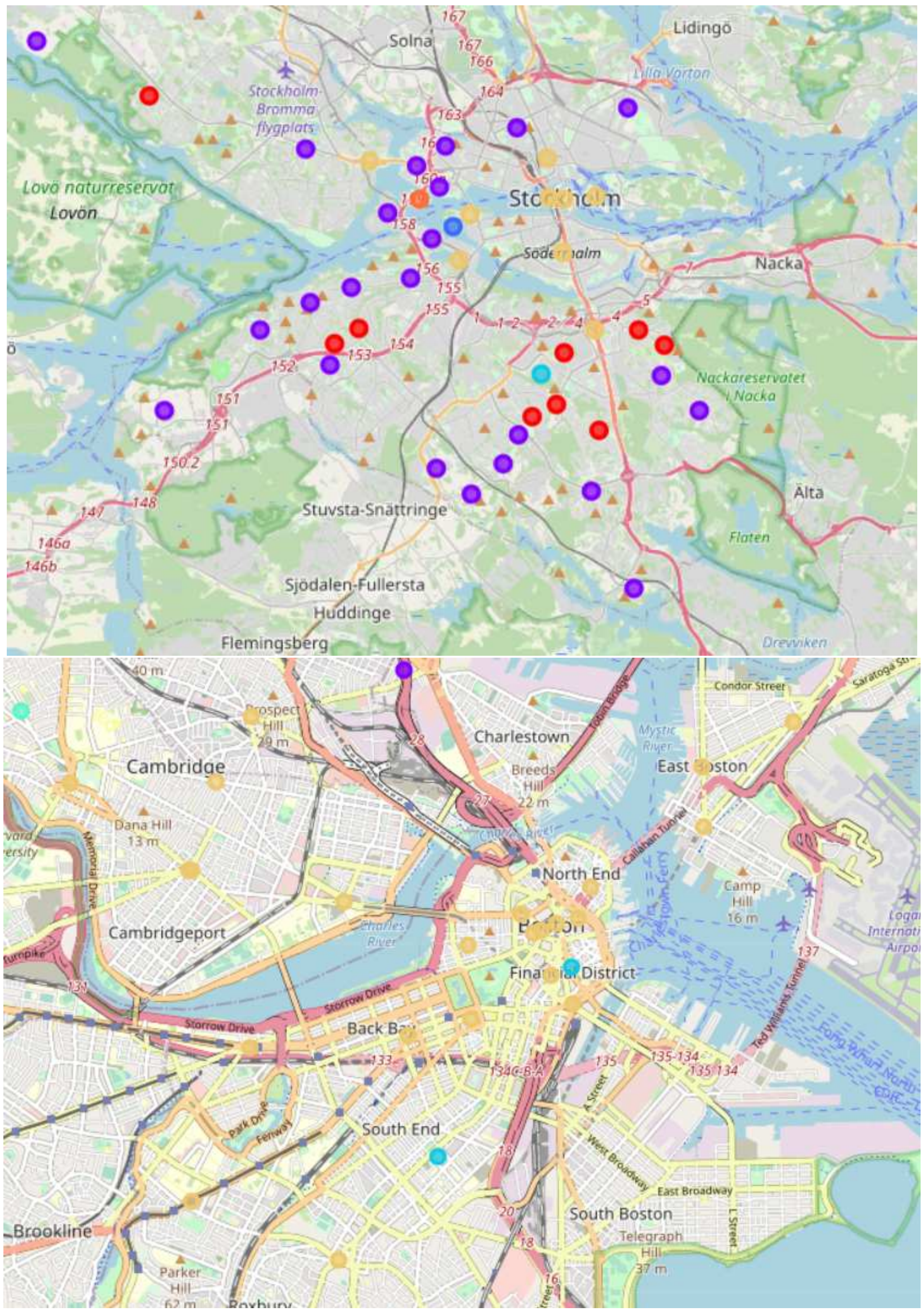

FIGURE 10: Neighborhoods in Stockholm and Boston Marked By Cluster Color

5 Conclusions

Ultimately we demonstrate that clustering is an interesting tool not just for exploring similarities between neighborhoods in the same city but also for giving us an interesting way of comparing neighborhoods in cities on different continents. Someone looking to relocate from one city to another might utilize such a tool to help them determine where to begin their residence search. More pragmatically it might have value for people or companies looking to expand business to an area with similarity to an already successful location or franchise. This analysis was meant largely as a proof of concept and a demonstration of learned techniques, and there's a lot of room to explore other machine learning techniques, particularly other clustering algorithms, and determine their relative effectiveness for exploration.