COMPARING THE SIMILIRATY OF NEIGBORHOOD BETWEEN NEW YORK AND TORONTO

Muhammad Ayoub

Introduction

The decision to move out of your city can be pretty hard whether it is for a new job of for college or any other reason.

Since not only are we leaving our friends and family but also we will have to adapt to a new lifestyle than the one we are usually used to.

One way to ease the transition is to try to pick an area on neighborhood that is similar to our previous neighborhood to help us and adapt faster to this new lifestyle.

So in this project we are going to see if we can use machine learning methods to try to find the most similar neighborhood between New York and Toronto.

Finally, we see that this problem can be important to anybody thinking of moving out of his/her current neighborhood whether they are moving to a different city or to a different neighborhood in the same city, also we can use this if we needed a change of pace and decided to move to neighborhood that is entirely different to the one we are currently living in.

DATA

For this project, we are going to need two sets of data:

- First for new York Data we used a the same dataset we have used previously in the 3rd week lab in the course[1] which provided a json file that contains an already well organized data that didn't need much cleaning and looked eventually like this:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

- Then for the Toronto data we retrieved the boroughs and neighborhoods from a Wikipedia page[2] using web scraping techniques and then we used the geospatial data file that was provided in the 3rd week assignment in order to add the coordinate for each area and after some cleaning by removing all unassigned boroughs and grouping neighborhoods that belongs to the same postal code we came with the below table:

Postcode	Borough	Neighborhood	Latitude	Longitude
M1B	Scarborough	Rouge,Malvern	43.8067	-79.1944
M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.7845	-79.1605
M1E	Scarborough	Guildwood,Morningside,West Hill	43.7636	-79.1887
M1G	Scarborough	Woburn	43.771	-79.2169
M1H	Scarborough	Cedarbrae	43.7731	-79.2395

- And after that we did some formatting for the Toronto neighborhood data frame to match the same format of the New York Neighborhood data frame in order to concatenate them into one data frame that contains all Neighborhoods after adding a suffix to each Neighborhood name to indicate the country the neighborhood belongs to, this step can be very useful if we decided to expand the project in the future to include more cities
- Eventually we used foursquare API to find the venues surrounding each neighborhood as within 700 (m) and then we extracted the 7 most common venues around each neighbor hood

Note: all the values of the parameters that we have chosen so far and in later sections are chosen after some trial and error and can possibly be tuned to achieve a more accurate results

Methodology

- **Features:**

  For the purpose of this project, we are going to be considering the venues surrounding a neighborhood to be the only deciding factor to determine the similarity between neighborhoods, as we will extract the X most common venues to use as the features of our model

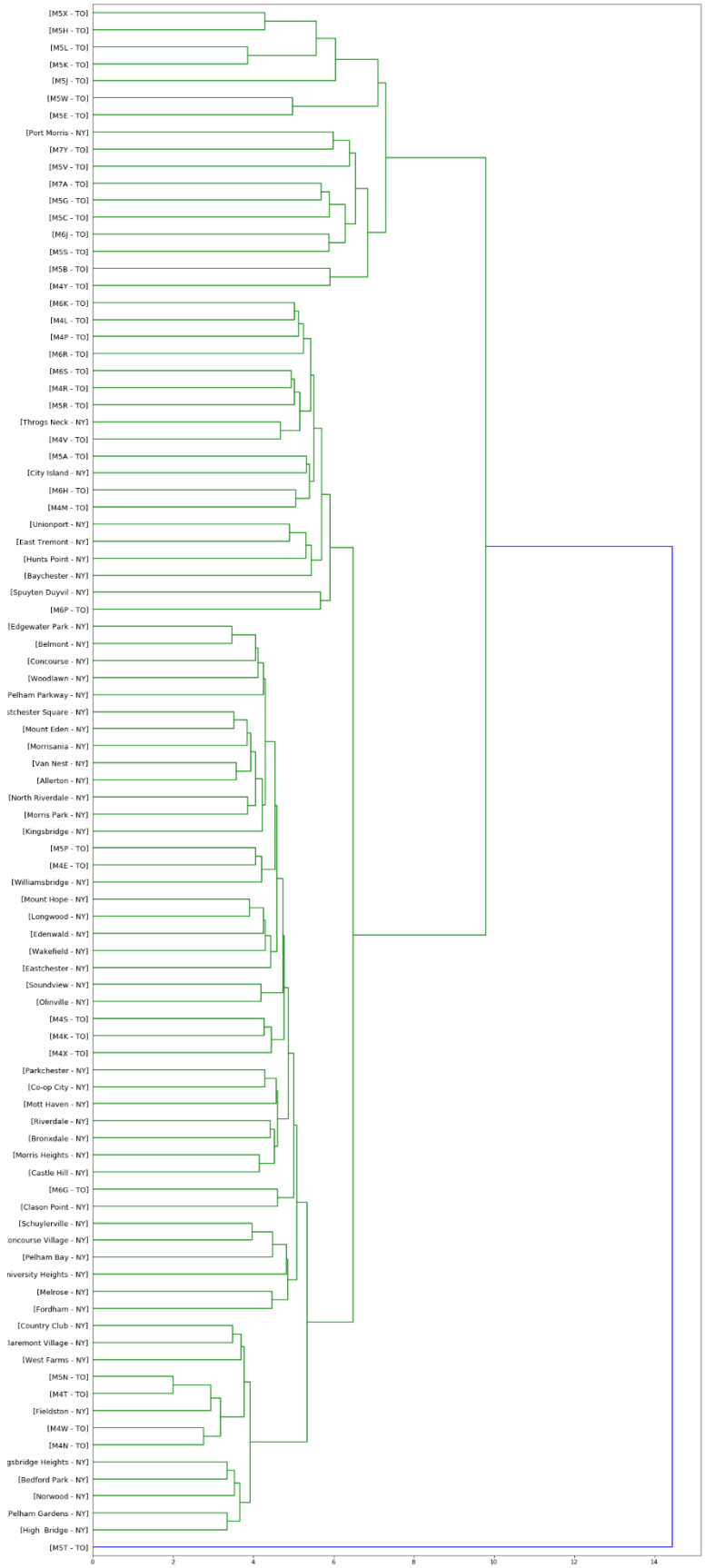
- **Model:**

  For our model we chose a hierarchal clustering model specifically Agglomerative Clustering since it is the most commonly used hierarchal clustering model, and the reason behind this choice is that for our goal the hierarchal model can give us different levels of similarities as opposed to classical partition based clustering like K-means which can only define the similarity to be between points that belong to same cluster without considering hoe similar are they, whereas in a hierarchal model we can easily find the neighborhood most similar and dissimilar for any neighborhood which can give more options for our goal

- **Libraries:**
  - for web scraping we used beautiful soup library
  - for our model we used scipy library
  - for retrieving venues we used Foursquare API

Results

For our first experiment we are going choose a subset of our datasets so we get a more clear dendrogram so we chose the Bronx borough from New York and all boroughs that contains the word Toronto from the Toronto Dataset and we chose the max depth of the tree to be 4 and after training our model we got the below dendrogram:


[M5X - TO]
[M5H - TO]
[M5L - TO]
[M5K - TO]
[M5J - TO]
[M5W - TO]
[M5E - TO]
[Port Morris - NY]
[M7Y - TO]
[M5V - TO]
[M7A - TO]
[M5G - TO]
[M5C - TO]
[M6J - TO]
[M5S - TO]
[M5B - TO]
[M4Y - TO]
[M6K - TO]
[M4L - TO]
[M4P - TO]
[M6R - TO]
[M6S - TO]
[M4R - TO]
[M5R - TO]
[Throgs Neck - NY]
[M4V - TO]
[M5A - TO]
[City Island - NY]
[M6H - TO]
[M4M - TO]
[Unionport - NY]
[East Tremont - NY]
[Hunts Point - NY]
[Baychester - NY]
[Spuyten Duyvil - NY]
[M6P - TO]
[Edgewater Park - NY]
[Belmont - NY]
[Concourse - NY]
[Woodlawn - NY]
Pelham Parkway - NY]
itchester Square - NY]
[Mount Eden - NY]
[Morrisania - NY]
[Van Nest - NY]
[Allerton - NY]
[North Riverdale - NY]
[Morris Park - NY]
[Kingsbridge - NY]
[M5P - TO]
[M4E - TO]
[Williamsbridge - NY]
[Mount Hope - NY]
[Longwood - NY]
[Edenwald - NY]
[Wakefield - NY]
[Eastchester - NY]
[Soundview - NY]
[Olinville - NY]
[M4S - TO]
[M4K - TO]
[M4X - TO]
[Parkchester - NY]
[Co-op City - NY]
[Mott Haven - NY]
[Riverdale - NY]
[Bronxdale - NY]
[Morris Heights - NY]
[Castle Hill - NY]
[M6G - TO]
[Clason Point - NY]
[Schuylerville - NY]
oncourse Village - NY]
[Pelham Bay - NY]
niversity Heights - NY]
[Melrose - NY]
[Fordham - NY]
[Country Club - NY]
laremont Village - NY]
[West Farms - NY]
[M5N - TO]
[M4T - TO]
[Fieldston - NY]
[M4W - TO]
[M4N - TO]
gsbridge Heights - NY]
[Bedford Park - NY]
[Norwood - NY]
Pelham Gardens - NY]
[High Bridge - NY]
[M5T - TO]
0
2
4
6
8
10
12
14

As we can see most neighborhoods that are very similar are within the same city but there a few exceptions like port Morris new York which we find that the most similar neighborhood to it is the ones inside the postal area M7Y which means that those two neighborhoods are closer to each other more than they are similar to other neighborhood within their same city.

Now to understand the idea of different levels of similarities we take the same previous example of Port Morris and M7Y and move up one level in the tree to find that both of these neighborhood are also similar with a lesser degree of similarities with M5V neighborhoods and we can repeat the same process for any other neighborhood to find the level of similarities depending on how far up the tree we need to move in order to find a common node.

Discussion

As we saw previously we were able to reach a model that can answer our question, However it is somewhat hard to measure the accuracy of mentioned model as it is the case with most clustering problems but that is okay in a way since the goal of this model is to give an approximate recommendation for someone looking to move to a new neighborhood and doesn't want to change his/her lifestyle dramatically but there is still one big limitation in our project and it is that we didn't use enough features to have the best decision possible since we considered the 7 most common venues to be the only deciding factor. So for future work we can add more features like land area, population density and average monthly temperature and other properties that can factor in one's decision to choose a new living area. We can also try different models that can help us achieve our goal such as fuzzy clustering model which can be very useful to us since it supports a data

point belonging to multiple clusters in different degrees on possibility.

Conclusion

In conclusion, we find that we can find hierarchal clustering model to help us find the most similar neighborhood for a specific neighborhood in order to help someone trying to decide on a new place to live without making dramatic changes to his/her routine but we find that one limiting factor in our work is the features we chose to help our model decision as it is not enough to help our model achieve the best decision possible thus it will add more feature in our future work.

References

[1]: https://cocl.us/new_york_dataset

[2]: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M