Clustering Amsterdam Neighborhoods Buurt or Wijk

Capstone Project - The Battle of Neighborhoods

Yi Li, liyi201809@gmail.com

1. Introduction

Being motivated by the projects of exploring Toronto and New York neighborhoods in this course, I decided to leverage Foursquare location data to explore my city, Amsterdam in the Netherlands. In Amsterdam, we can basically classify all neighborhoods by two regional types, Buurt and Wijk in Dutch. The former means a typical neighborhood or a localized community within a city, and the latter stands for a more residential area with a lot of old and/or historical family houses. In this project, I would like to investigate whether these two types of region have a similar distribution of venues, or backwardly, can we cluster all neighborhoods based on their venues in order to identify the region types of each of them.

2. Data Source

Statistics Netherlands CBS, a Dutch governmental institution that gathers statistical information about the Netherlands, published a dataset of some basic figures for the districts and neighborhoods of Amsterdam in 2016 [1]. More specifically, the CSV file consist of 583 feature columns and 4019 instance rows. Those figures included important data that I need in this project, such as Neighborhood, Region Type, Population, Latitude and Longitude.

3. Methodology

3.1. Data preparation

Given this Amsterdam dataset, we can easily find the feature *region_type* is provided to indicate this neighborhood is Buurt or Wijk, which can be treated as our ground truth labels in the clustering step later. Worth to mention, a few of neighborhoods which were defined as both Buurt and Wijk, will be simply removed from the whole dataset to avoid making a binary clustering task more complicated.

To retrieve the venue's information of each neighborhood, we can go over the coordinates (latitude and longitude) of them and apply the Foursquare API, however, it will take a quite long time to process all 4019 neighborhoods. As a result, we will only focus on those neighborhoods with a relatively large population. To investigate the distributions of inhabitants of these neighborhoods, a histogram can be plotted as below.

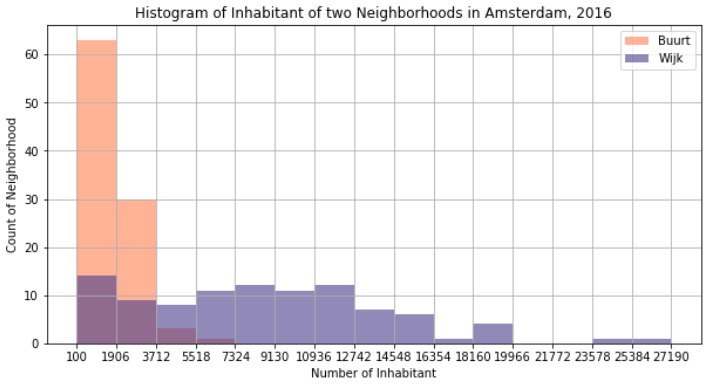

Figure 1 Histogram of Buurt and Wijk Neighborhoods in Amsterdam

From the histogram we can notice the inhabitant distribution of two types of neighborhoods are quite different, which means it can be an important feature for machine learning classification task. However, the goal of this project is investigating whether we can distinguish the neighborhood types by clustering their venues, thus, we will drop the inhabitant feature after using it to choose more populated neighborhoods. To ensure a balanced data, we finally chose those neighborhoods with more than 2500 inhabitants.

3.2. Foursquare Venues data

By passing the coordinate of each neighborhood to the Foursquare API, we can retrieve the corresponding venues information, which was specified to be the top 100 venues within a radius of 500 meters. Then, we can do one-hot-key encoding on the category of venues, summing them up and taking the average to get the frequencies table. This table is ready for us implement the k-Means clustering of *sklearn* library.

3.3. k-Means Clustering

Since we want to know whether the neighborhood is Buurt or Wijk, the parameter k of the clustering should be set to be 2. After implementing the clustering algorithm, we will get a list of labels. We will compare this list with our ground truth label mentioned in 3.1.

4. Results

After going through the methodology, we described in the above sections, we got some results. The first figure visualizes the Buurt vs Wijk neighborhood distribution on the map of Amsterdam. There are 138 neighborhoods have more than 2500 inhabitants, meanwhile, 138 of them are Buurt and 77 are Wijks. The second picture shows the two clusters derived by the k-Means clustering.

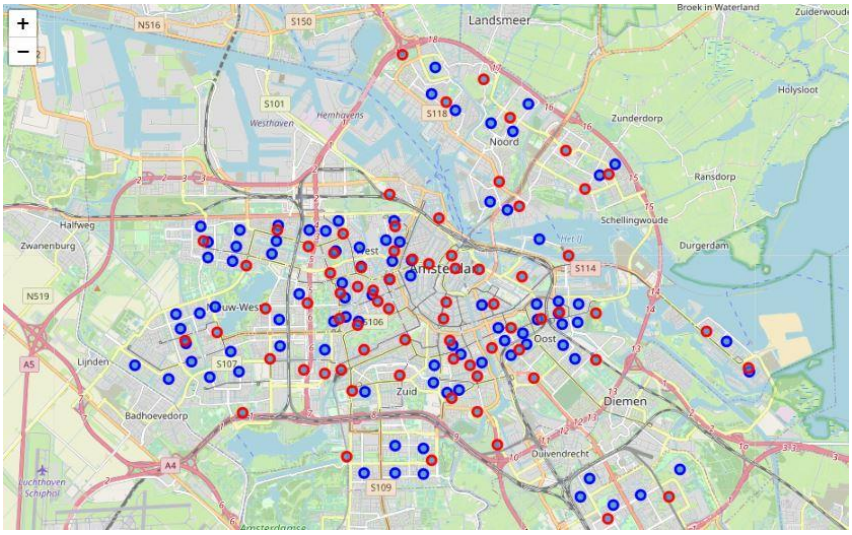

Figure 2 Distribution of Buurt and Wijk in Amsterdam

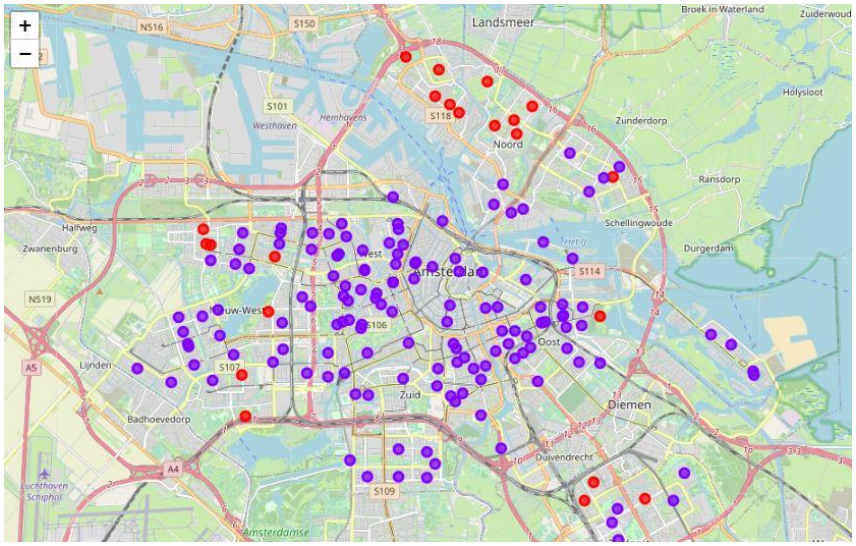

Figure 3 Two clusters of neighborhoods based on their venues

Also, we make use of the evaluation functions from *sklearn.metrics* to compare our clustered labels with the ground truth labels (region_type), the derived accuracy is just around 53% and F1 score(which can be interpreted as a weighted average of the precision and recall, where a score reaches its best value at 1 and worst score at 0 [2]) is around 0.612.

5. Discussion

Based on the results we derived in the above sections, we can know that the correlation between region types and the venues of a neighborhood is not really strong. From the map, we

can see the Buurt and Wijk are pretty mixed together with each other, however, the clusters derived based on venue information are clearly separated.

The accuracy score and the F1 score also shows that the clustering labels are not well match the ground-truth labels, which means the clusters can hardly be used to indicate the region types of a neighborhood.

From the histogram we already known that, there are some other features might be more indicative for check whether a neighborhood is Buurt or Wijk. More inferential statistical model can be built to help this; however, this is out of the scope of this projects.

6. Conclusion

In this project, we investigated the possibility to predict the region type (Buurt or Wijk) of a neighborhood in Amsterdam, based on its venue information. The data source is provided by Statistics Netherlands CBS. In the process of exploration, we made use of *geopy* and *folium* library to visualize the geographical data, applied foursquare API to retrieve the venue information. In the end, we implemented k-means clustering techniques to clustering neighborhood data with venues features. The results showed that Buurt and Wijk are geographically mixed, the distributions of their venue categories are also quite similar. It is hard to know/cluster the region types of a neighborhood based on its venue information.

[1] https://claircitydata.cbs.nl/dataset/districts-and-neighbourhoods-amsterdam

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html