Capstone Project - The Battle of Neighbourhoods – Warsaw

Applied Data Science Capstone
by IBM/Coursera

Author: Aman Agrawal

February 29, 2020

Table of contents

1. Introduction and Data description

Introduction/Business Problem

Prices of flats in Poland go up faster than inflation, according to the report of money.pl website. Rising apartment prices on the market effectively obscure another problem - the increase in rental prices. This trend affects students, young workers without their own flats or economic immigrants. According to the analysis of experts at Rynekpotny.pl, the increases reached even 23%. Despite everything, life in Warsaw tempts many young people.

The capital is mainly attracting to itself those who are focused on making dizzying careers or artists and creative people. The heart of the city is the City Centre, which is vibrant with life at any time of day or night. It is one of eighteen districts, but each of them has different advantages. In this scenario, machine learning tools should be used to assist people coming to Warsaw to

make wise and effective decisions. As a result, the business problem is:

- **How can we help people moving to the capital to choose the right location to rent an flat in Warsaw?**

In order to solve this business problem, we intend to merge Warsaw districts into a cluster in order to recommend facilities. We will recommend facilities according to the amenities and necessary equipment of the surrounding facilities such as: Café, Bus Station, Pizza Place.

Data

To consider the objective stated above, we can list the below data sources used for the analysis.

- Districts of Toronto Wikipedia page was scraped to pull out the necessary information;
- Coordinate data for each Districts of Toronto obtained through Nominatim search engine for OpenStreetMap data;

In order to investigate and target recommended locations in different locations depending on the presence of facilities and necessary objects, we will access the data through the FourSquare API and arrange it as a data frame for visualization. By combining data about districts in Toronto and data about amenities and essential facilities surrounding such properties from the FourSquare API, we will be able to recommend an appropriate location.

2. Methodology

The methodology in this project consists of two parts:

- **Data Understanding & Data Preparation**: Process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. Once the data has been collected, it must be transformed into a useable

subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases.

- **Data Preparation & Data Exploration**: Process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. It is necessary to visualise the districts of Toronto.
- **Modelling**: To help people find similar neighbourhoods in the safest borough we will be clustering similar neighbourhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster neighbourhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighbourhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighbourhood.

Data Understanding & Data Preparation

Scrape the Wikipedia page and gathering data into a Pandas dataframe

To start with our analysis, we used the Beautiful Soup package to transform the data in the table on the Wikipedia page into the below pandas data frame. The data frame will consist of three columns: Distric, and Neighbourhood. Subsequently, we transform the data into a pandas data frame.

	District	Neighborhood
0	Bemowo	Bemowo Lotnisko
1	Bemowo	Boernerowo
2	Bemowo	Chrzanów
3	Bemowo	Fort Bema
4	Bemowo	Fort Radiowo

Data cleaning

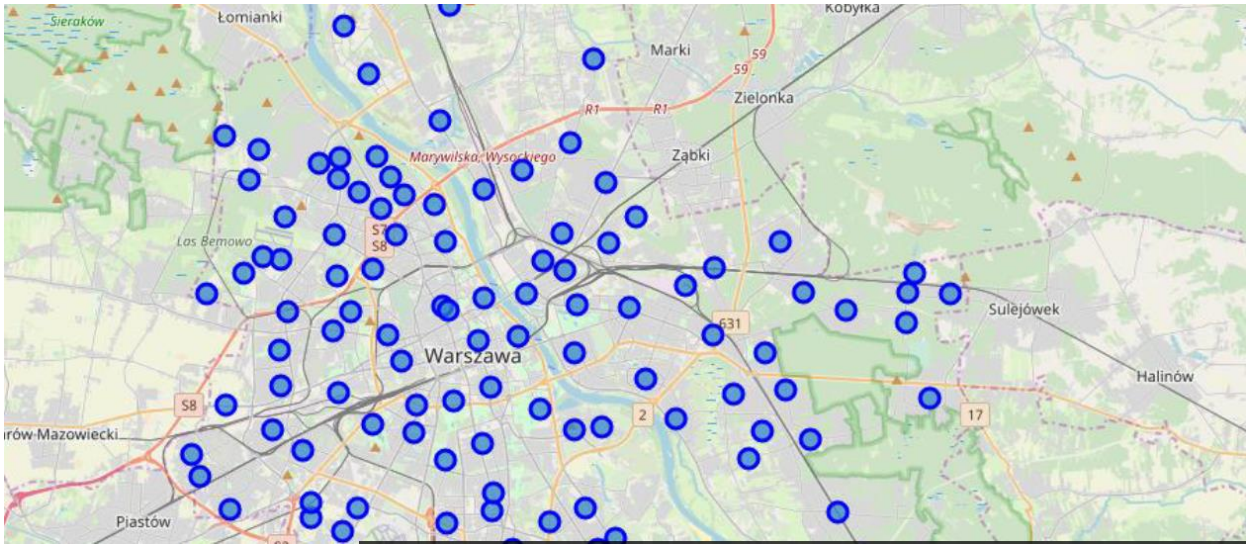
Only cells that have a district assigned to them will be processed. There may be more than one district in one of the postal codes. If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough.

Use geopy library to get the latitude and longitude values of Warsaw Localities

fter we have built a dataframe of Warsaw localities along with the district name and neighbourhood name, in order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighbourhood. It possible to export data to a csv file for easier loading later..

	District	Neighborhood	Latitude	Longitude
0	Bemowo	Bemowo Lotnisko	52.261261	20.910737
1	Bemowo	Boernerowo	52.262390	20.901451
2	Bemowo	Chrzanów	52.216759	20.882969
3	Bemowo	Fort Bema	52.256562	20.938620
4	Bemowo	Fort Radiowo	52.257211	20.891900

Generating a map of Warszawa and plotting the Neighbourhood data on it



Modelling

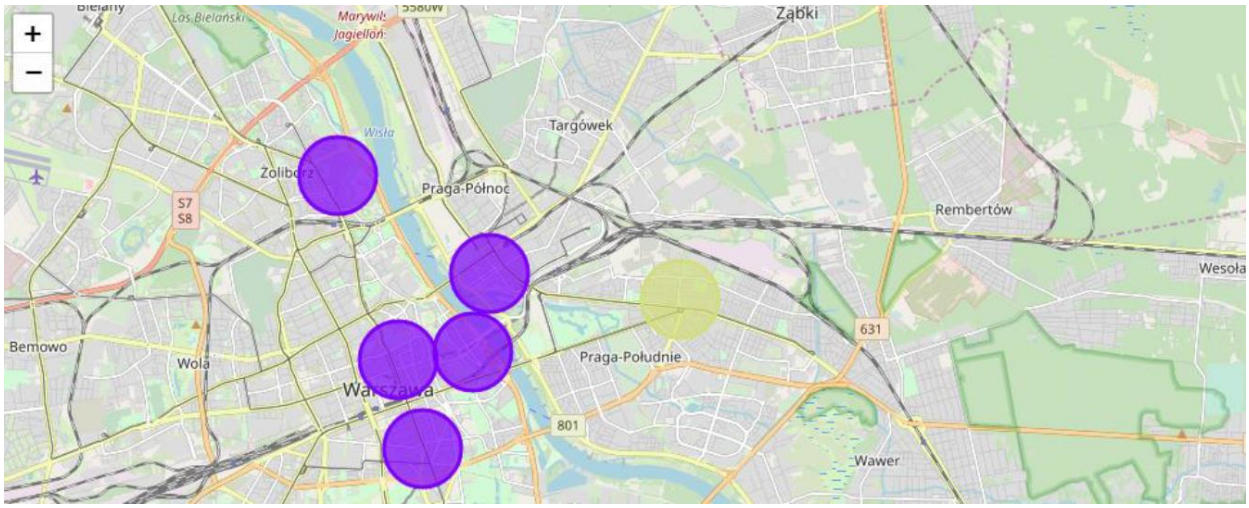
- Finding all the cafes within a 500-meter radius of each neighbourhood.
- Perform one hot encoding on the venues data.
- Grouping the venues by the neighbourhood and calculating their mean.
- Performing a K-means clustering (Defining K = 4)

Finding all the cafes within a 500-meter radius of each neighbourhood.

Foursquare is the most trusted, independent location data platform for understanding how people move through the real world. We have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots for each neighbourhoods of Toronto. The recommended location needs to have many eating and shopping venues nearby. Convenient public transport is also required.

All modelling process you find notebook [link]

Analysis



Examine Cluster 0

```
warsaw_merged.loc[warsaw_merged['Cluster Labels'] == 0, warsaw_merged.columns[[1] + list(range(5, warsaw_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
43	Stary Mokotów	Bakery	Café	Ice Cream Shop	Italian Restaurant	Convenience Store	Coffee Shop	Dessert Shop	Pizza Place	Movie Theater	Eastern European Restaurant

Examine Cluster 1

```
warsaw_merged.loc[warsaw_merged['Cluster Labels'] == 1, warsaw_merged.columns[[1] + list(range(5, warsaw_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
59	Stara Praga	Diner	Restaurant	Hotel	Coffee Shop	Middle Eastern Restaurant	Road	Public Art	Plaza	Park	Movie Theater
66	Powiśle	Pizza Place	Café	Eastern European Restaurant	Asian Restaurant	Pub	Polish Restaurant	Bar	Italian Restaurant	Science Museum	Restaurant
69	Śródmieście Północne	Nightclub	Coffee Shop	Cocktail Bar	Café	Hotel	Italian Restaurant	Restaurant	Beer Bar	Polish Restaurant	Greek Restaurant
70	Śródmieście Południowe	Café	Vegetarian / Vegan Restaurant	Coffee Shop	Cocktail Bar	Italian Restaurant	Sushi Restaurant	Hostel	Bistro	Plaza	Hotel
141	Stary Żoliborz	Café	Thai Restaurant	Polish Restaurant	Coffee Shop	Plaza	Burger Joint	Restaurant	Public Art	Playground	Breakfast Spot

Examine Cluster 2

```
warsaw_merged.loc[warsaw_merged['Cluster Labels'] == 2, warsaw_merged.columns[[1] + list(range(5, warsaw_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
89	Natolin	Sushi Restaurant	Restaurant	Park	Coffee Shop	Indian Restaurant	Italian Restaurant	Sandwich Place	Café	Convenience Store	General Entertainment

Examine Cluster 3

```
warsaw_merged.loc[warsaw_merged['Cluster Labels'] == 3, warsaw_merged.columns[[1] + list(range(5, warsaw_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
53	Grochów	Café	Dessert Shop	Bus Station	Supermarket	Pizza Place	Fast Food Restaurant	Flea Market	Restaurant	Coffee Shop	Mexican Restaurant

Results and Discussion

I think it is no surprise that all these districts are very centrally located in the circular layout of Warsaw. Locations meeting the criteria of popular places would usually be in central locations in many cities around the world. From this visualization it is clear that on a practical level, without data on the basis of which decisions could be made, the circle of 103 locations is very large. We have significantly narrowed the search area from 8 potential districts to 5, which should respond to the business problem.

Moreover, FourSquare is not popular in Warsaw, the data maybe out-dated or unreliable, the report should gather more data from other location data source such as Google Place API.

Conclusion

Different applications of this analysis are available based on a different methodology and possibly different data sources. The stakeholder problem has been resolved. The stakeholder wants to find the best place to live in Warsaw, and the "best location" factors are based on the number of places in the food, cafe and park category around the location. Machine learning technique based on content filtering is the most appropriate method to solve the problem. Eight destination locations may not be a good choice, but I can quickly choose other locations and issue a recommendation again.