Location prediction of the eastern European cuisine restaurant in the Toronto area

Data

Author
Vladimir Dikić

Keywords
Location prediction
Toronto income and demographics analysis
Clustering
Classification

Report issued : 02-21-2020

2. Data acquisition and processing

2.1 Data acquisition

The necessary data and the sources are described in the list below:

- Neighbourhoods, Borough and postal code data:
    - The information is obtained from canadapost.ca website, the grouped data is taken from [1]. This data is primarily needed to access the geographical data and Foursquare API. Tables were scraped using the BeautifulSoup toolbox
- Neighbourhood geo data:
    - The coordinates data for each neighbourhood was obtained using geocoder-arcgis method on postal code information from the Toronto neighbourhood data
- Population, average income and demographic data:
    - The demographics data was obtained from the [2]. This will be used to determine the best-suited neighbourhood in terms of number and income of target customer group (i.e. immigrants from Eastern Europe). Tables were scraped using the BeautifulSoup toolbox
- Venue data:
    - Foursquare API is used to obtain the venues in Toronto area. This will be used to determine the amount of competition in each neighbourhood

2.2 Data processing

Initial step of data processing was scrapping the tables from [1] and [2] using the BeautifulSoup toolbox. Subsequently raw data was pre-cleaned and placed in the corresponding dataframes. Data was additionally cleaned removing duplicates, NaN values and transforming numbers of string type into real numbers

Dataframe obtained from data [1] will be referred to as “postal code dataframe”, and dataframe [2] will be referred to as “demographics dataframe”.

The main difficulties in merging the postal code and demographics data came from the significant difference in neighbourhood naming. This consisted of differently grouped

neighbourhoods in both data sets (e.g. neighbourhood X and Y in data set 1 are grouped together while in the data set 2, X and Y are separate, and the other way around). In addition, the level of detail was different in the data sets, where table 1 can contain one neighbourhood but split into west and east side, while the table 2 only accounts for the entire area (e.g. Steeles vs South and East Steeles).

First, it was necessary to separate all the grouped neighbourhoods. This was made slightly more difficult by inconsistent grouping symbols (e.g. "\", "-", ",", "/"). The loop was utilized to account for all the aforementioned symbols, where in order to retain the other data (e.g. postal code while separating neighbourhoods), grouped data was separated by using split, expand and stack function.

Second, in some cases the data in the demographic data set only accounted for whole neighbourhoods, while postal code data set also accounted for geographical areas of the neighbourhood, as was previously mentioned. This was solved by appending the demographic dataframe with the disaggregated neighbourhoods. Each row of the postal code data frame was tested if it contains the neighbourhood that includes the name of non-joinable neighbourhood from the demographics dataframe (e.g. East Steeles contains Steeles). Then, the demographics data frame was appended using the following rules:

- Neighbourhood is split into segments that correspond to number of segments in postal code dataframe. Names are equalized
- Population data is separated in even amounts depending on the number of segments
- Demographics and the average income are assumed to stay the same for each separated segment

Demographics of the area were determined by the majority spoken second language in the neighbourhood (outside of English). Subsequently, these language groups were assigned to geographical / cultural groups. These assignments do not yield in a precise segmentation (e.g. grouping Persian and Somali or Filipino and Mandarin), but will allow for easier clustering in order to find the best location for an eastern European restaurant.

Table 2-1 Segmentation based on geographical area and cultural similarity

Geo/Cultural group	Language group
South Asian / Indian	[Tamil, Gujarati, Bengali, Urdu, Punjabi]
East Asian	[Filipino, Cantonese, Mandarin, Korean]
West Asian and African	[Persian, Somali]
Unspecified	[Unspecified]
Eastern European/Slavic	[Russian, Bulgarian, Ukrainian, Polish]
South European / Romance	[Greek, French, Spanish, Portuguese, Italian]

Literature

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

[2] https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods