The Battle of Neighborhoods | Business Proposal | Introduction

Introduction:

This project aims at creating a tool that can be used to explore facilities in and around neighbourhoods in a way that allows residents to make informed decisions and select the most optimal location that best caters to their needs.

An example of such need could be smaller cities and towns in various provinces of Canada. Canada has had a long and successful immigration policy that allows regulating the country's population. From its onset in 1870, the influx of immigrants has been tailored to grow the population, settle the land, and provide labour and financial capital for the economy. However, in a democratic country, such as Canada, both the newcomers and the existing residents have a right to move around and select a location with the housing prices, good schools, medical facilities and job opportunities that suit their needs and financial means.

This project would be highly beneficial for those who seek a systematic, algorithm-based approach for comparison and selection of the desired parameters, e.g. access to public or catholic schools, supermarkets, shopping malls, movie theaters, hospitals, or religious communities. This selection can be made using analytical tools rather than basic internet searches, reliance on real estate agencies or word of mouth that could be biased or misleading, whether intentionally or not. Studies such as [1] show that people need to think critically about claims and compare their options using a systematic concept because simply trusting the sources of information may not be a sufficient basis for making the best choice:

- Competing interests can result in misleading claims.
- Personal experiences or anecdotes alone are an unreliable basis for most claims.
- Opinions of acquittances, real-estate agents or newspaper/social media publications are not solely a reliable basis for making informed-decisions.
- Endorsements by community leaders or other respectable individuals do not guarantee that comparisons have been fair.

Objective:
The main objective of this project is to propose an approach for selection of an optimal community in a new city for someone looking to relocate. The individual using this proposed tool would be able to compile, sort and filter parameters of interest, for example:

1. A list of houses/residencies that can be sorted in terms of housing prices in an ascending or descending order, or
2. A list of restaurants that cater particular type of cuisine that can be sorted in terms of location, prices, rating and reviews.

The problem for the Canadian province of Ontario is that most immigrants to the province make their home in the GTA, specifically 77 per cent, according to an August report from the Conference Board of Canada [2]. However, as their populations age and young adults move away, small cities and towns across Canada are increasingly looking to immigration as a way to rejuvenate their workforce and expand their tax base. But many struggle to find reliable data and make informed decision to move to smaller rural areas due to lack of data and unavailability of unbiased selection tools. The city of Waterloo, Ontario will be used in this project to demonstrate the proposed approach as opposed to Toronto.

Specific Objectives:
Specifically, this project looks at a community of Russian-speaking expatriates who are looking to re-settle in the Greater Toronto Area (GTA) or in other smaller neighbouring communities in the province of Ontario. This project will propose a tool that can be used to perform searches and comparisons between various neighbourhoods in order to determine a community with specific desired parameters, e.g., access to grocery stores and marketplaces where Russian-produced goods and food supplies are available.

Additional Parameters:
Additional parameters could include ease of access to children's daycares and after-school activities where Russian language is taught. Another important factor could be availability of various medical facilities (e.g., Russian-speaking family doctors, dentists, phycologists and other medical specialists) for Russian-speaking patients and access to nursing homes where services in Russian language are available for seniors. Restaurants offering traditional Russian cuisine is another important factor that may sway a customer's decision to relocate to a particular community.

Data:

The data used in this project was acquired mostly from Wikipedia.com and Toronto.ca websites. In order to be used in the proposed approach it was formatted and restructured into a csv. file.
The files used in the project can be found on github repository as well as referenced in [3]. The project data was transferred to a local drive and compiled to dataframes as can be seen from the project code.

Based on the 2011 Canadian census data, most of Russian-native speakers in Toronto (GTA) reside in the geographical area of York. Per [4], there are more than 35,000 people in this area who list Russian as their mother tongue.

Next, Foursquare approach was used for data segmentation and clustering as follows:

Foursquare API:
This project uses Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business. Important to point out that 100% data capture accuracy cannot be achieved due to inherent data source omissions and inaccuracies.

Methodology:

The approach in this project is based on the methodology discussed in Week 3 lab and is as follows:

- First, neighbourhood postal codes were converted into the equivalent latitude and longitude values.
- Next, Foursquare API was used to explore neighborhoods in both cities, Toronto and Waterloo. The data for Toronto was limited only to those corresponding to the York region.
- Next, a function to determine common venue categories in each neighbourhood was created and executed.
- Next, the neighbourhoods were clustered according to the venue features.

- Subsequently, K-means clustering algorithm was used to complete the analysis. Folium library was utilized to visualize the emerging clusters for neighborhoods in Toronto and Waterloo.

These steps are described in more detail below:

Work Flow:
Using credentials of Foursquare API, features of near-by places of the neighborhoods would be mined. Due to http request limitations, the number of places per neighborhood parameter would be reasonably set to 50, the radius parameter would be set to 500 for Toronto, and to 2000 for the city of Waterloo.

Clustering Approach:
To compare the similarities of the two cities, it was decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in a big city like Waterloo and Toronto. To be able to do that, it is necessary to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

Libraries Used to Develop the Project:
- Pandas: For creating and manipulating dataframes.
- Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
- Scikit Learn: For importing k-means clustering.
- JSON: Library to handle JSON files.
- XML: To separate data from presentation and XML stores data in plain text format.
- Geocoder: To retrieve Location Data.
- Beautiful Soup and Requests: To scrap and library to handle http requests.
- Matplotlib: Python Plotting Module.

Results:
Using the proposed methodology, it was determined that North York area in Toronto and Waterloo-Central area in the city of Waterloo both showed the highest number of neighbourhoods in a cluster. Shown below are the code and the geo-spacial graphs for the project data. Tables for each cluster with the venue categories are also shown here.

YOUR AMAZING ANALYSIS SKILLS AND AWSOME CODING GOES HERE

Discussion:
Based on the cluster data for the two cities of interest, the classification for each cluster was done with calculation of venue categories (i.e., the most common category). Analysis of each cluster does not produce conclusive one-to-one results when Foursquare - Most Common Venue data is used. Thus, the following assumption had to be made for each cluster:
- Cluster 1: North York: Mix (i.e., more than one (1) category)
- Cluster 2: North York: Parks
- Cluster 3: North York: Paper / Office Supplies Stores
- Cluster 1: Waterloo: Mix (i.e., more than one (1) category)
- Cluster 2: Waterloo: Parks
- Cluster 3: Waterloo: Restaurants

Next, a gap in determination of specific districts and establishing correlations between common venues was identified. A systematic, quantitative approach of venues recorded in Foursquare was needed in order to advance onto the next stage. This was further exacerbated by the fact that not all cities have similar common venues despite obvious similarities. Therefore, an additional step of applying a suitable extraction and special integration method was performed.

Conclusion:

The work performed in this Capstone was focused on design and application of a new systematic tool that can be used to perform quantitative analysis of data between different cities. Two large Canadian cities in the province of Ontario, specifically Toronto and Waterloo, were selected to demonstrate the proposed approach. In this project, the specific parameters of interest were the availability of Russian language-based services, such as restaurants offering authentic Russian-style cuisine and service. The results indicate that the chosen methodology was successful in identification of neighbourhoods with similar demographic composition and venue categories. It showed high promise for those who seek a systematic, algorithm-based approach for comparison and selection of the desired parameters and can be applied to any of the desired features, e.g., access to public or catholic schools, supermarkets, shopping malls, movie theaters, hospitals, or religious communities.

Future studies can focus on establishing relations withing the acquired data, as well as additional exploration of tools and methods for further optimization of the results.

References:

[1] Jeffrey K. Aronson, “Key concepts for making informed choices”, A Nature Research Journal, 12 August 2019

[2] https://www.conferenceboard.ca/e-library/abstract.aspx?did=10342&AspxAutoDetectCookieSupport=1

[3] https://github.com/jnasimi/Coursera_Capstone

[4] https://www.toronto.com/news-story/5588878-russian-in-toronto-10-neighbourhoods-where-you-re-likely-to-hear-it/