IBM APPLIED DATA SCIENCE CAPSTONE PROJECT
Week 4 – proposal report

Identifying the restaurant cuisine trends

By: Hwei Ong

Introduction

The restaurant industry is vibrant in many cities, where changes in demographics and lifestyle are driving the demand for such services. As the economy increases household income, there is potential for increased consumer spending. Nonetheless, with housing, healthcare and common household expenditures consuming majority of this spending, the consumers tend to be conservative in their spending for dine-ins and take-outs. It is not uncommon for food trends to change and drive consumer spending on certain foods. Too many restaurants serving the same cuisine in one neighborhood cluster would mean a highly competitive environment for a newcomer. While a high density of restaurants may signify a strong customer base, it may also mean higher business expenses due to higher rents in a popular area. Starting a new restaurant business requires significant capital and planning upfront. Moreover, it would be more difficult to obtain approval to open a new restaurant from the city for an area that has a concentrated number of restaurants due to congestion concerns. As such, it would be ideal to be able to research the restaurant industry of a city's neighborhoods in order to gain some insights as to what is currently available.

Business Problem / Target Audience

The business problem is how to identify restaurant cuisine trends in the city. The idea is to cluster neighborhoods by cuisine type and gain an understanding on the density of restaurants within the neighborhoods.
The target audience would be:

- investors and/or restauranteurs (i.e. commercial business planning) who are interested in opening a new restaurant;
- policy wonks involved in urban planning as such analytics can guide the zoning of different areas for development; and
- property developers who are interesting in knowing what new property type would be suitable for an area (e.g. a mixed-use residential building vs. a commercial building).

Data Sources and Methods Used

The Canadian cities used in this analysis are Toronto, ON and Calgary, AB. As such, the following data are required:

- A list of neighborhoods and their respective latitude and longitude coordinates for each city – the data required in order to obtain data on the types of venues found in each neighborhood, and for subsequent plotting on the city map during analysis.
- A list of venues found in each neighborhood – the data is required for the machine learning algorithm used to perform clustering on the neighborhoods.

The following Wikipedia pages contain the list of neighborhoods for each city specified below:

- Toronto (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Calgary (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T)

The following Python packages will be used in this project:

- Beautiful Soup will be used to design a web scraper that can obtain the information contained on the respective Wikipedia pages listed above.
- The requests and json libraries will be used to work with the data downloaded using Beautiful Soup.
- The numpy library is a dependency in some functions defined within the neighborhood, primarily use to define a range for looping.
- Pandas will be used to structure and clean the downloaded data.
- Geocoder will be used to obtain the coordinates (latitude and longitude) of each neighborhood.
- The unsupervised machine learning algorithm, k-means, from Scikit-learn will be used to cluster the neighborhoods.
- The kneed library will be used to identify the elbow point, which signified the optimal number of clusters for *k*-means.
- Folium will be used to visualize the neighborhoods and clusters for each city.
- The seaborn and matplotlib libraries will be used to visualize the data.

Additionally, the Foursquare API will be used to query for venue data using the coordinates of each neighborhood. The venue data contains information on the different types of business found in each neighborhood. From these, we can gain insights into the density of a neighborhood, as well as evaluate the potential customer base and competitors.

Below is a map showing the locations of the different cities covered in this project.