CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

By: Thomas Sinka

12/22/2020

INTRODUCTION

The objective of this project is to analyze the cultural makeup of restaurants in Toronto and see how they can be grouped together and classified in different neighborhoods.

This report could be useful for people looking to start their own restaurant business and would like to know which areas a particular type of cultural venue would do well in judging by surrounding establishments

Hopefully, the findings will be clear enough for someone to look at the data and say, “This area does well for Chinese food”.

DATA

The information utilized in this project is made up of two parts:

1. Neighborhood and postal code information that was obtained through Wikipedia, and corresponding geographic coordinates datamined through the Geocoder API
2. Venue data obtained through the Foursquare API

The Foursquare service provides a wide range of venue information such as ratings, reviews, photos, but the relevant service for this problem is the ability to search for venues that fall in a specified category.

METHODOLOGY

OBTAINING DATA

Using the Toronto neighborhood location data gathered through Wikipedia, the Foursquare Search API endpoint was queried for venues categorized as restaurants within a 1km radius of each group of neighborhoods (grouped by postal code coverage)

PREPROCESSING

Steps were taken to filter out generic and non-culture-specific venues:

1. Venues that did not contain “Restaurant” in their category name were removed
2. Venues categorized as “Fast Food” were removed
3. Venues categorized as just “Restaurant” were removed
4. Irrelevant categories such as “Vegetarian” and “Comfort Food” were also removed

Categories that did not have an explicit culture in their name were recategorized. Example: Arepa Restaurants were recategorized to South American Restaurants.

Categorizes that did not appear frequently enough to be significant in clustering were merged up to broader categories. Example: Brazilian Restaurant appeared one time, so it was recategorized as South American.

ANALYSIS

It is possible for groups of neighborhoods to have overlaps in venues as a search radius of 1km was applied to each location.

The total count of venues across all neighborhoods was 521 before dropping duplicates, and 490 afterwards.

The frequency of each cultural venue in Toronto was tallied and a bar graph showing the top 10 was produced.

A list of each (group of) neighborhoods and their top 3 venue types and frequencies was produced

CLUSTERING

The “K-nearest neighbors” algorithm was used to cluster neighborhoods based on their similarity to each other in terms of cultural makeup. “One-hot encoding” was applied to each venue to convert them into a binary matrix, and then the mean of each category occurrence in each neighborhood was taken. This step essentially converts the neighborhood venue frequency data to a form the machine learning algorithm can “understand” and use to cluster neighborhoods by similarity.

The “Elbow method” was used to calculate the optimal ‘k’ value, which was k=7.

CLASSIFICATION

A list of the neighborhoods in each cluster was produced along with their top three venue categories in order to manually label the clusters based on their defining characteristics.

VISUALIZATION

An interactive map of Toronto was created with clickable markers for each group of neighborhoods, which were color-coded based on their class. Clicking the markers list the neighborhood's names along with the class identifier.

RESULTS

The KNN algorithm did a decent job at clustering neighborhoods based on similarity and clear distinctions could be made in 39.3% of the neighborhoods. The remaining 60.7% were too mixed to make any obvious distinction and were grouped into a single cluster.

Cluster #	Culture Label	Area %
0	Chinese	7.87
1	Caribbean	4.49
2	Mixed	60.67
3	Spanish/Vietnamese	3.37
4	Korean	3.37
5	Japanese	11.24
6	Middle Eastern	8.99

On an individual basis, each neighborhood in the mixed cluster could be analyzed to give insight on which cultural restaurants are the most prevalent in order to make informed business decisions

DISCUSSION

The concern throughout this project was the large number of dimensions in the venue data: each possible venue category represented another dimension of complexity. Steps were taken to reduce the number of variables as much as possible but further steps could be taken. See: https://en.wikipedia.org/wiki/Dimensionality_reduction

A limitation in this project was the unavailability of neighborhood coordinates, so the next best option was to use postal code coordinates and group all the neighborhoods by postal code. The ability to cluster on individual neighborhoods would have probably increased the distinctions in the neighborhoods falling in the mixed cluster significantly.

CONCLUSION

Overall, the information gathered was sufficient to give the reader of the Power Point presentation an idea of the cultural makeup of neighborhoods in Toronto based on the restaurants present in those areas.