Applied Data-Science (Machine Learning)

Capstone Project on Neighborhoods

Submission Report

Content

1) Introduction and business problem

Objective:

The idea of this machine learning project is to create a neighborhood comparison model.

The model should support you to compare neighborhoods of different cities and find the most suitable place for yourself. So for example you are about to relocate to another city and would like to find a flat in a similar neighborhood.

In a way it is similar to a recommendation tool, which recommends similar neighborhoods in other cities.

Target Audience:

This tool can be very useful if you're moving to a different city and not sure what neighborhood to choose. So in the tool you would be able to enter the name of the neighborhood of your current city and city you would like to move or relocate. The tool will provide a list of similar neighborhoods in the new city and show the location on a map.

So the target audience can be :

1. People relocating to another city and looking for similar neighborhoods

2. Business people who are on a business trip and want to stay within a certain neighborhood for some time

3. Owners of shops that want to open a shop in another city within a special area

Success Criteria:

The objective of this project is not to develop the application itself – this could be a simple web application. The idea is rather to provide the algorithm/ model behind the application.

In order to keep it more simple we take the example, that somebody would like to move from Toronto (e.g. neighborhood : St. James Town in Toronto) to London or New York and would like to find similar neighborhoods in London or New York.

Success criteria of the project are :

- define common cluster/class values for similar neighborhoods in London / New York
- deliver optimized model for these classes
- provide a list of similar neighborhoods within the chosen cities
- show the recommended neighborhood on a map

2 Data understanding and data preparation

In order to be able to segment and compare different cities we need borough and neighborhood data from these cities as well as latitude and longitude (coordinates) of each neighborhood.

1) Extract data from the websites

We decided to compare neighborhoods between Toronto, London and New York. Therefore we need to get the relevant data for all those cities.

We get the basic neighborhood data from this websites :

https://en.wikipedia.org/wiki/List_of_areas_of_London

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

https://geo.nyu.edu/catalog/nyu_2451_34572

The data can be extracted from the websites with the web scraping tool "Beautiful Soup".

Beautiful Soup is a Python package for parsing HTML and XML documents (incl. having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML.

Beautiful Soup can be installed using the Python package manager pip or the anaconda package manager.

2) Data cleaning and pre-processing

Date need to be cleaned (e.g. remove whitespaces). Data cleaning will be done using Python Panda.

A big part of the pre-processing is encoding. This means representing each piece of data in a way that the computer can understand it, hence the name encode. There are different

ways of encoding such as Label Encoding or One Hot Encoding. One Hot Encoding will also be used in this project.

3) Geocoding

Geocoding is needed to get latitude and longitude for each neighborhood and display these on a map.

Geocoding can be done with Python geocoder library. The geocoder will call ArcGIS World Geocoding Service which is a REST API provided by ESRI.

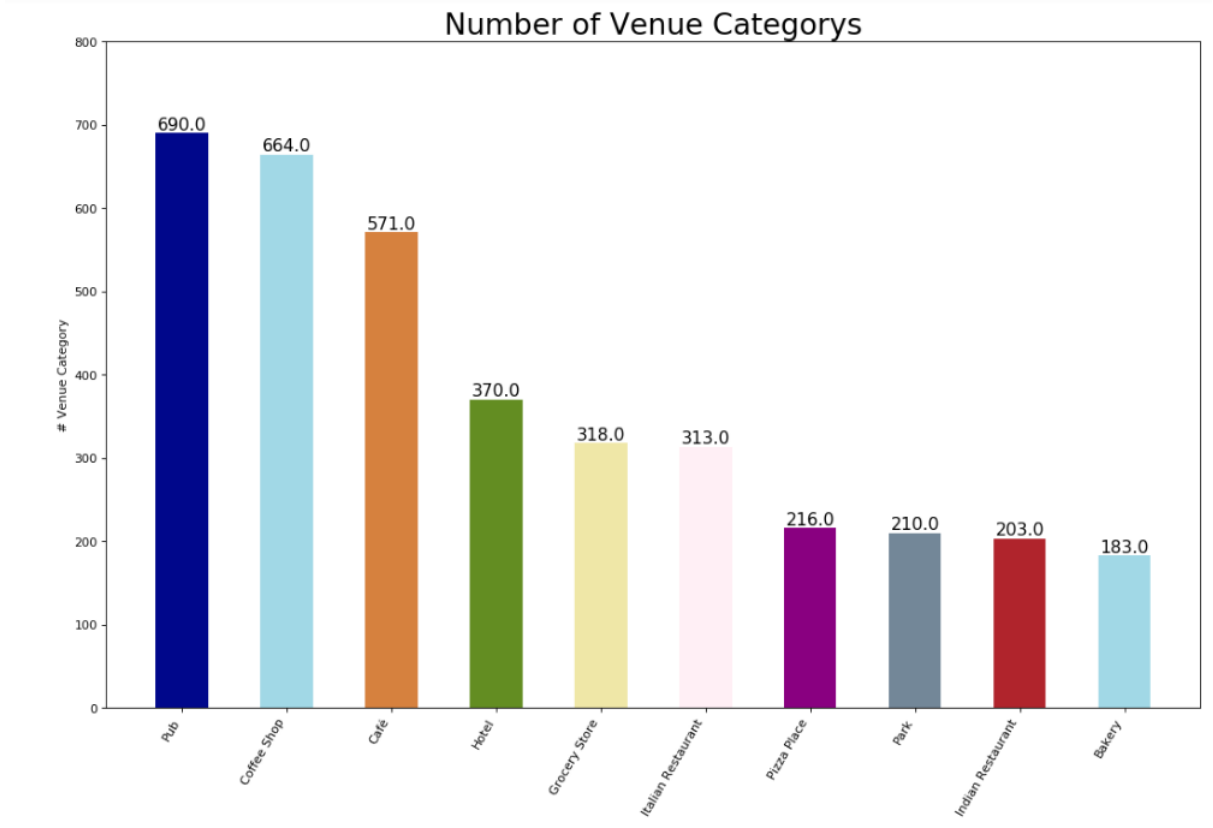
4) Get venue data for the neighborhoods for clustering

Foursquare API is used to get venue data for each neighborhood in London and Toronto.

Foursquare API can give the full details about a venue including location, tips, and categories. Important for this project are mainly the categories of venues (e.g. Hotels, Bars, Coffee Shops).

For this the explore function will be used to finally get the most common venue categories in each neighborhood.

The most important venues of London (for this project) are shown below (according to their frequency) :



In order to improve accuracy of the model we will do feature (venue) selection. This is the process of choosing the most relevant features in the data. In feature selection, we remove features to help the model generalize better to new data and create a more interpretable model.

Here is an excerpt of the features, which have only a very small impact in our use case :

Train Station	-0.031855
Grocery Store	-0.026305
Asian Restaurant	-0.016174
Tapas Restaurant	-0.005104
Indian Restaurant	0.009746
Greek Restaurant	0.012736
Japanese Restaurant	0.027825