Applied Data Science (IBM) Capstone Project :- The Battle Of Neighbourhoods

Submission Report

Content

1. Introduction and business problem
2. Data Gathering, Cleansing and Exploratory Data Analysis

1) Introduction and business problem

Objective:

The idea of this machine learning project is to build a similar neighbourhood recommendation system.

Target Audience:

The Business Problem is related to many People who do often switch companies for a better opportunity. But while switching, most of the times people have to change the current Neighbourhood/City/Country. They always wish that they get all same required things like services, enjoyments, clubs, restaurants, hangout places etc in the new Neighbourhood/City/Country. So Is there a way we can recommend them best neighbourhoods near their new office.

Introduction / Business Problem :

So Here in this project, we are going to recommend the best and same type of neighbourhoods as their current neighbourhood to a user in terms of service, search for the potential explanation of why a neighbourhood is popular, the cause of complaints in another neighbourhood, or anything else related to neighbourhoods.

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 Gathering, Cleansing and Exploratory Data Analysis

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) Web Scraping

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 Cleansing

Before we do any kind of experiments the Data needs to be cleaned.The Data cleansing is done using Python Pandas.

In Data cleansing part ,

- Remove unwanted / missing rows.
- Fill missing values.
- Format data in a particular way that computers can understand.

3) Get Location Data

For Geonencoding, We used Geoencoder python library to get latitude and longitude for each neighborhood.The geocoder will call ArcGIS World Geocoding Service which is a REST API provided by ESRI.

4) Get venues list of all 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.