Finding Prospective Business Development Routes in a City of Choice – with Data Science.

Prepared by Maksim Mislavskii for
IBM Data Science Professional Certificate capstone project on Coursera

Background

The business scene in today's city is highly competitive, diverse, and complex. Most neighborhoods are packed with venues offering a wide range of products and services in a plethora of categories, both commercial and non-profit, in an endeavor to cater to needs of residents and visitors alike.

1. Business Problem

In this environment aspiring **startups**, **business entities seeking expansion** of their operation, and **government agencies or NGOs** caring to provide better service to the public, all face a tough challenge of making decision on **what new venues are best opened in which areas of the city**.

2. Proposed Solution

To explore existing business scenes in each of the city neighborhoods to find similarities and distinctions and, based on that data, build an automated system to identify prospective penetration routes. The key objectives of such data science driven project can be summarized as follows:

3. Objectives

- ✓ Generate exploratory description of city neighborhoods based on the business scene in each of them. Highlight similarities and distinctions by means of clustering;
- ✓ Suggest preferable categories for opening a new venue in any given neighborhood;
- ✓ Suggest preferable neighborhoods for opening a new venue of given category.

4. The Data. Acquisition, Processing, Analysis

To meet the above objectives, the following data will be required:

1. The **list of neighborhoods** - that is, the low-level administrative divisions - in the city of choice.
   - ➢ If not provided by Client, can be obtained by web scraping using an algorithm in Python programming language from sources like Wikipedia or city administration website. For example, if the target city were Bangkok, the capital of the Kingdom of Thailand, the lowest level administrative division would be subdistrict, or "khwaeng", and there is a page titled "Khwaeng" on Wikipedia from which a list of all 180 neighborhoods across the city's 50 districts can be extracted
   - ➢ In case the primary source doesn't include geographical coordinates, the list will have to be geocoded, for example, using the Geopy library for Python.
2. The **list of existing venues** in each neighborhood
   - ➢ Can be acquired via API of a geoinformation system operator like Google or Foursquare. For example, a "venues/search" endpoint request to Foursquare API returns a list of venues known to the service within a given radius around a geographical location point (latitude and longitude).

Obtained data will be persisted in an SQLite database according to a model developed on MySQL Workbench and subject to K-Means clustering leveraging the capabilities of Scikit-Learn package and Pandas dataframes with visualization in Folium maps / Matplotlib charts to discover similarities and distinctions between the city neighborhoods in terms of the venue scene and generate recommendations as per the project objectives using an algorithm similar to those employed by existing media or goods recommender systems envisaging **locality** itself **as a prospective customer**.

5. The Outcome

An automated venuescape penetration recommender in form of a Django-powered web application designed to suggest most-wanted categories per location or best-prospect locations to open a venue in a given category.