Personalized Travel Destination

Capstone Project Report

From: John Harrop

Date: 2020-04-09

Introduction

A hypothetical, but realistic problem is being investigated for the Capstone Project in the IBM Data Science Specialization on Coursera. This report is presented in conjunction with the supporting work that was conducted using Python in a Jupyter Notebook. A business problem is presented and then translated into a science/engineering question that a research and development team can address. Data and methods are selected to address this problem discussed by an interpretation and recommendation based on the findings.

Business Problem

A travel agency is interested in developing a niche market identifying less well-known destinations that fit with individual client's profiles. The concept would be used both to target marketing to their clients but would also be available to their agents when discussing options with clients. The company feels that there is a huge amount of information available on the Internet about under utilized, lower profile destinations. Based on feedback from their clients they feel that making these kind of connections could provide a significant competitive edge for their business.

Following this boardroom level discussion we have suggested that the scope of this initial test be to determine if destinations can be differentiated using open data sources. A positive answer to this question would be an important green light for further development. This characterization would need to systematically group destinations in a way that corresponds to clients likes (or dislikes). This is a complex problem that could benefit from a number of sources with different kinds of data (venues, topography, climate, etc.). An initial test of the concept has been proposed using only venue data.

Data

A set of locations were selected that represent a number of different regions and various size communities. Several characterization samples were collected for each location. Locations were chosen from regions in western Canada, UK and Ireland and Argentina. A fourth location is being considered. Each region is sampled in several locations that include cities, towns and villages. For cities, multiple sampling centres

are used but this is not possible for smaller locations. A mixture of well-known destinations that attract tourists, industrial or commercial areas and smaller "off the beaten track" are included in the study.

Geographical coordinates (longitude and latitude) were determined for each location for input into the venue search engine. In addition, elevation, temperature (minimum, maximum and average), monthly rainfall (minimum, maximum and average) and the population of the entire community (not just the venue search radius) was also determined. These are considered simple proxies for topography and climate. Due to the small number of locations in this study these parameters were determined manually from Wikipedia.