Predicting population size based on abundance of different venues

Franziska Metge

June 15, 2020

1 Introduction

Problem

We live in a world of travelers but we also live in a world of individuals with divers interests. It would be preposterous to assume that everybody enjoys exactly the same cities. Some people might enjoy cities with a good night life, while others want to travel to cities with a diverse cultural landscape.

Background

Traveling greatly promotes personal growth, hence I try to travel somewhere new every year. By know I have traveled to most European capitals and other major cities. Each year it becomes harder and harder to pick a city and I have also encountered cities that I did not enjoy visiting. One major reason was that they were too crowded with tourists (I understand irony of the problem). For my next travel destination I would like to select a less popular city, while ensuring that the city is similar to what I have enjoyed before.

Solution

Therefore I will develop a program which is able to find cities a user might find enjoyable based on an input city and venue landscape. I will implement the code in an Ipython notebook and will make use of the python libraries introduced in the Machine Learning Course such as `numpy`, `pandas`, and `sklearn`. The program will return as many similar cities as desired. Besides a table listing the similar cities, the program will return a heat-map visualizing the similarity of these cities by venue occurrence. A scatter-plot using the cities coordinates will also be returned. This material will enable the user to confidently select their next travel destination.

Target audience

My idea is intended to be used by people who are looking for a new travel destination spanning further than all the known classics like Paris or New York. A user will be able to make the most of the program if they had already traveled to a considerable amount of cities and know what they appreciated.

2 Data

I will use the population data made publicly available by the United Nations (Demographic Yearbook – 2018[1]). I will use `numpy` and `pandas` libraries to clean the data. An example of this data is given in table 1.

Secondly, I will use the geopy library to acquire the coordinates for all cities. I will separate cities into different classes/groups based on their population size, i.e. < 0.5 Mio, $0.5 - 1$ Mio, $1 - 5$ Mio, $5 - 10$ Mio, $10 - 20$ Mio, and > 20 Mio (see Table: 2).

[1] `https://unstats.un.org/unsd/demographic-social/products/dyb/dyb_2018/`

City	Population Size	Country
Adrar	200834.0	Algeria
Ain Defla	450280.0	Algeria
Ain Temouchent	299341.0	Algeria
ALGIERS (EL DJAZAIR)	2712944.0	Algeria
Annaba	442230.0	Algeria
Batna	768444.0	Algeria
Béchar	236213.0	Algeria
Bejaïa	559981.0	Algeria
Beskra (Biskra)	563245.0	Algeria

Table 1: Cleaned population data

City	Population Size	Country	Latitude	Longitude	population_bin
Adrar	200834.0	Algeria	27.9458867	-0.1992938330258469	1
Ain Defla	450280.0	Algeria	36.15868425	2.084281730358365	1
Ain Temouchent	299341.0	Algeria	35.26665705	-1.149927622407504	1
ALGIERS (EL DJAZAIR)	2712944.0	Algeria	36.7753606	3.0601882	3
Annaba	442230.0	Algeria	36.8982165	7.7549272	1
Batna	768444.0	Algeria	35.3384291	5.731545299000572	2
Béchar	236213.0	Algeria	31.62298095	-1.914198993519679	1
Bejaïa	559981.0	Algeria	36.7511783	5.0643687	2
Beskra (Biskra)	563245.0	Algeria	34.7845635	5.812435334419206	2

Table 2: Cities with coordinates

I will use the `Foursquare API` to look for all venues within a 5km radius of the city's center. In order to deal with the limited amount of requests that can be made to foursquare (see Table: 3).

City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adrar	27.9458867	-0.1992938330258469				
Ain Defla	36.15868425	2.084281730358365				
Ain Temouchent	35.26665705	-1.149927622407504	Fast food Le Loft	35.2949989986954	-1.137600108318426	Fast Food Restaurant
ALGIERS (EL DJAZAIR)	36.7753606	3.0601882	Restaurant Le Thyrolien	36.77518773893406	3.058731268449381	BBQ Joint
ALGIERS (EL DJAZAIR)	36.7753606	3.0601882	CARACOYA	36.76667223648845	3.053610267518587	French Restaurant
ALGIERS (EL DJAZAIR)	36.7753606	3.0601882	"TNA ""Théâtre National d'Alger"""	36.78097827704275	3.060508018126319	Theater
ALGIERS (EL DJAZAIR)	36.7753606	3.0601882	Didouche Mourad	36.76557038107158	3.051074029384855	Plaza
ALGIERS (EL DJAZAIR)	36.7753606	3.0601882	Tantonville	36.780824	3.06031	Café
ALGIERS (EL DJAZAIR)	36.7753606	3.0601882	Musée d'Art Moderne Algérie	36.77720262790217	3.058272841808561	Art Museum

Table 3: Foursquare results

The data will be stored in a table with one row per city containing the cities name, country, coordinates, population size, population size category, number of venues from different categories in one-hot encoding (see Table 4).

City	Population Size	Country	Latitude	Longitude	population_bin	ATM	...	Fast Food Restaurant	...	Zoo Exhibit
Ain Temouchent	299341.0	Algeria	35.26665705	-1.149927622407504	1	0		1		0
ALGIERS (EL DJAZAIR)	2712944.0	Algeria	36.7753606	3.0601882	3	0		0		0
Annaba	442230.0	Algeria	36.8982165	7.7549272	1	0		0		0
Bejaïa	559981.0	Algeria	36.7511783	5.0643687	2	0		0		0
Bordj Bou Arreridj	422986.0	Algeria	36.095506	4.661100173631754	1	0	...	0	...	0
El Bayadh	192958.0	Algeria	33.63785225	1.012203911250456	1	0		0		0
Guelma	363716.0	Algeria	36.3491635	7.409498952760461	1	0		0		0
Jijel	391096.0	Algeria	36.8167305	5.771494	1	0		0		0
Laghouat	371204.0	Algeria	33.8063518	2.8808616	1	0		0		0

Table 4: Foursquare results

Table 4 will be the input table used for my program. Based on the columns **ATM** to **Zoo Exhibit** the function will calculate a similarity score between all cities using the function `pdist` from the package `sklearn`.