Generating Maps with Location Hotspots and Leveraging Foursquare for Tourism

Justin Zheng 21.8.2019
Applied Data Science Capstone

Table of Contents

Introduction and Business Problem

Background

The purpose of this data project is to leverage the Foursquare location and venue data in order to generate maps that would be able to show hotspots within a specific neighborhood, which in this project will be Rittenhouse Square in Philadelphia, PA. Maps as generated is intended to be printed and handed to tourists for guidance to aid their exploration of the city and in particular Rittenhouse Square. The inspiration behind this project is from personal experience whereby through my work as a tutor at my university I was responsible for helping a number of visiting scholars thrive and enjoy Philadelphia. Due to many of them lacking a functioning data-plan for them to access the internet which prevents them from using any online map service such as Google Maps, my office recommended printing regular maps to help them navigate Philadelphia.

Business Problem

Foursquare data will need to be leveraged to find different important venues around the area of Rittenhouse Square. This project aims to generate relevant maps through data analysis. To generate relevant maps, I will need to filter out important categories of venues, for example a shopping mall or a convenience store. Furthermore, distance and overall radius of location search for venues will need to be individually determined through trial and error in order to generate an appropriately extensive list of venues in each category.

Interest and Relevance

The target audience for this particular data science project will be travel agencies, the hospitality industry such as hotels, motels or other forms of accommodation as well as any tourist-centric venues. Such venues often display and allow visitors to take maps with them to help them navigate in the neighborhood Although physical maps are becoming less relevant due to well-run and up to date online navigational systems like Google Maps, they still play a role especially as tourists may not always have quick access to the internet. Furthermore, the kind of cluster-based maps that I will generate in this project will allow extremely quick and simple navigation to different kinds of venues as the maps are separated by venue category. This will help alleviate confusion and the problem that many run into using apps like Google Maps, which is too many information which makes the map unreadable.

Data

Acquisition and Purpose of Data

A majority of location data was scraped directly from Foursquare. This location data included postal codes, names, longitude and latitude values and exact addresses of different venues within the Philadelphia area. For example, one of the examples I scraped of a shopping mall near Rittenhouse Square was the Liberty Place on 1625 Chestnut Street and is within the 19103 postal code area in which Rittenhouse Square is located. Example of a data-frame:

1	The Shops at Liberty Place	Shopping Mall	1625 Chestnut St	39.951919	-75.167833	19103	PA
2	The Gallery at Market East	Shopping Mall	901 Market St	39.952689	-75.158149	19107	PA

Much of the data scraped from Foursquare was omitted, however I will discuss this more under the Data Cleaning section. The purpose of the data I kept and acquired is to generate a map with clusters, hence columns such as “Neighborhood” was negligible since I had the postcode. Furthermore, the map code I ran only required longitude and latitude values hence they were the priority of my data acquisition.

Longitude and latitude values that were acquired through Foursquare were particularly used to form maps out of clustering. Postal codes acquired through Foursquare were further used to identify specific neighborhoods and to selectively clean data to only show different venues of the same post code as Rittenhouse Square.

Other data was also acquired and added to data-frames manually. The Foursquare location data was not fully accurate, hence I also utilized Google Map data and Apple Maps data to manually fill missing addresses and postal codes. This was mainly done for consistency and accuracy in order to generate the utmost updated maps.

Cleaning

The Foursquare data scraped required significant cleaning. I ran a function that extracts venue categories from the original data scraped from Foursquare then filtered each row. The end result was as follows:

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng	postalCode	state
0	On-Line Shopping Store	Shoe Store	1509 Walnut St	US	Philadelphia	United States	NaN	80	[1509 Walnut St, Philadelphia, PA 19102, Unite...	[{'label': 'display', 'lat': 39.95205126106208...	39.952051	-75.163867	19102	PA
1	The Shops at Liberty Place	Shopping Mall	1625 Chestnut St	US	Philadelphia	United States	at 16th St	378	[1625 Chestnut St (at 16th St), Philadelphia, ...	[{'label': 'display', 'lat': 39.95191851893288...	39.951919	-75.167833	19103	PA
2	The Gallery at Market East	Shopping Mall	901 Market St	US	Philadelphia	United States	at 9th St	458	[901 Market St (at 9th St), Philadelphia, PA 1...	[{'label': 'display', 'lat': 39.95268865552114...	39.952689	-75.158149	19107	PA
3	Chinatown	Neighborhood	NaN	US	Philadelphia	United States	NaN	659	[Philadelphia, PA 19107, United States]	[{'label': 'display', 'lat': 39.95548773443018...	39.955488	-75.156693	19107	PA
4	Walnut Street	Shopping Mall	NaN	US	Philadelphia	United States	NaN	447	[Philadelphia, PA 19103, United	[{'label': 'display', 'lat': 39.94973955025363	39.949740	-75.167047	19103	PA

Furthermore, many columns in the generated pandas data-frame were removed such as distance from town hall or the "cross-street". Both were negligible as they only served to help me locate the venue, which the postal code was already sufficient in providing. Furthermore, the decision to leave out "cross-street" was for clarification purposes as streets in Philadelphia run through the city hence the identification of a perpendicular street was not relevant in locating venues near Rittenhouse Square.

Also, the column names within the data also needed to be transformed for simplicity purposes. Each column name began with *"location."* hence I ran a simple for loop and split at the dot.

Trial and Error for Venue Category

It is also important to note that after data-cleaning, often the data would only contain one or two venues which is unrealistic for most venues I used in my analysis. For example, when I ran a search for "Cafe" through Foursquare and finished all cleaning, only two venues remained. This is unrealistic, hence I needed to run a search again, however this time under the category of "Coffee Shop". Running this search however required additional data-cleaning as the Foursquare data set which is scraped often contain outright errors such as:

13	Groom Barber Shop	Salon / Barbershop	1324 Locust St	US	Philadelphia	United States	btwn Juniper And Broad	510	[1324 Locust St (btwn Juniper And Broad), Phil...	[{'label': 'display', 'lat': 39.9481439748482,...	39.948144	-75.163728	NaN	19107	PA	4b632f

The data also contained many slight inaccuracies, such as listing a shopping mall which contains coffee shops as a coffee shop. This needed to be cleaned as well.

Therefore, for each different category I was required to run and scrape different category data-sets from Foursquare and through trial and error decide upon which to use for further clustering and cleaning. This was done to maintain realism and accuracy.