What would you do next?
Modelling Activity Transitions using the Foursquare API

Elson Serrao ELSON.SERRAO@GMAIL.COM

1. Introduction

Have you ever wondered, where do people go after watching a game at the stadium? Or if you just ate lunch what would you do next? If you were at the game and your team won you would probably end up at a bar or pub celebrating your team's victory. And after lunch, you would probably want to get some dessert. Are these just guesses, or are they backed by some statistical data?

The main goal of this project is to answer such questions, leveraging the data obtained from location sharing services like Foursquare. Each individual task that a person does could be considered as an activity. For example, watching a game at the stadium, having drinks at a bar, eating lunch, having dessert are all activities. The transition from watching a game at the stadium to having drinks at a bar or eating lunch to having dessert is considered to be an activity transition. In this project, we will try to model the activity transitions and answer the primary question: "If a person is doing some known activity currently, what is the person most likely to do next?"

Modelling activity transitions could be considered as a sub-task of modelling patterns of human mobility which are significant for traffic forecasting, urban planning, as well as epidemiological models of disease spread [Cheng et al. (2011)]. As such one of our target audiences would be governments and municipalities. Understanding human mobility also allows developers to enhance recommender systems [Noulas et al. (2011)] providing targeted recommendations to users.

2. Related Work

As compared to what research is already present in this field, this is a very small scale project. If the topic has piqued your interest, I would suggest you to go through the papers listed below. However, definitely have a look at Livehoods if the papers seem boring. They have a very good interactive web interface that help understanding several aspects of a dynamic city.

Related work in this field usually utilize the checkin footprint of users to model human mobility. However, getting checkin information using the Foursquare API is a premium call which caused several limitations. Also unless the checkin is made public, the API terms and conditions states that you cannot store the data more than a couple of hours. For this reason, I had to mostly use only the regular calls.

2.1 Data Sources and Analysis

As a use case, in this project we will try to model activity transitions in the city of San Francisco. We would first need geospatial data to get the boundaries of San Francisco. It can easily be obtained from DataSF. We get the geojson files for zip codes and San Francisco neighbourhoods.

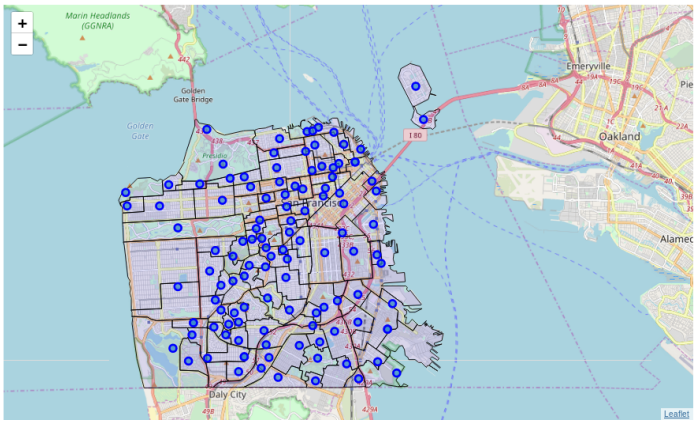

Figure 1: Neighbourhoods in San Francisco

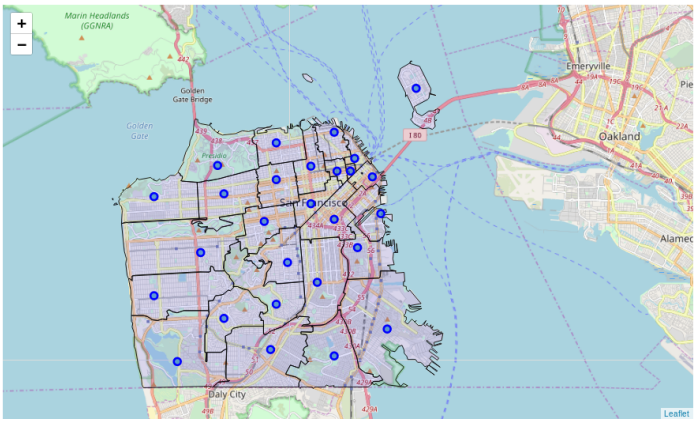

Figure 2: Postal Codes in San Francisco

We will then use Google Maps Geocoding API to get locations within the zip codes and neighbourhoods. Using these locations as seed points we will explore the venues of San Francisco using the Foursquare API. A venue is identified by the venue_id. We parse the result obtained from the explore endpoint of the Foursquare API to store the venue name, address, category, city, state country, postal code, cross_street, latitude and longitude. A venue is capable of having several categories. However, every venue has one category marked to be it's primary category. We extract only the primary category for the venue.

We find that the venues in San Francisco are distributed unevenly among the postal codes. There are some regions that have more venues than the others.

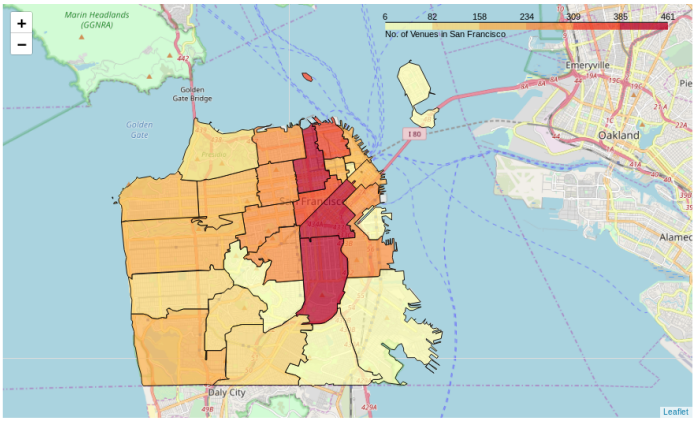

Figure 3: Density of Venues across San Francisco

We consider the primary category of the venue as the activity. The list of all the categories available in Foursquare are also obtained. Each category will be identified by its unique id. A category can have several sub-categories. Along with the id, we retrieve the name of the category, the icon that represents it and the parent category if it has any.

We find the number of Food and Shop & Service venues account for more than 50% of all the venues in San Francisco.

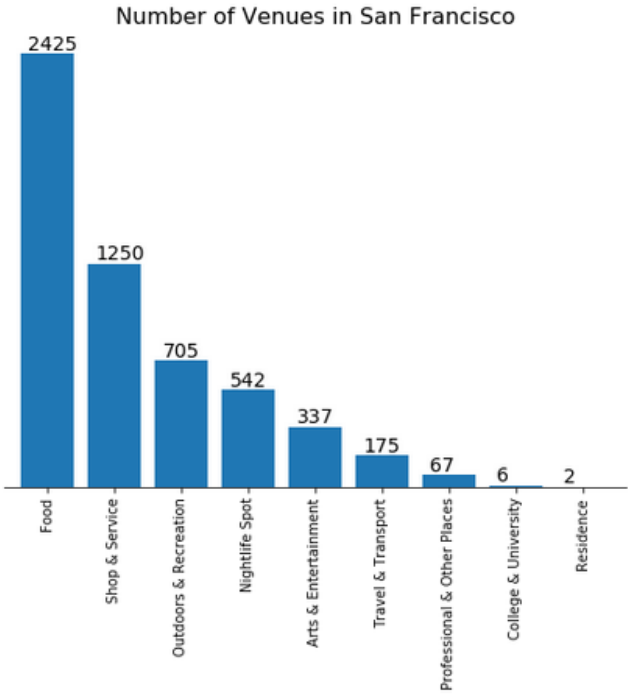

Figure 4: No. of Venues in different categories. Only the main categories as show in the plot

For each venue, we use the next venues endpoint from the API to get a list of next venues. This endpoint gives us the top 5 most probable venues to be visited next. An activity transition could be considered as a change in venue i.e. when a person moves from one venue to the next. And the activity the person preforms at the venue is the primary category of the venue. We can then correspondingly map the venue transitions to activity transitions using the category to which the venue belongs to.

For each venue transition, we then calculated the distance between them. Nearly 92.5% of all the venue transition obtained are within a radius of 1km. This further indicates that the activities that people do, are generally restricted with the region.

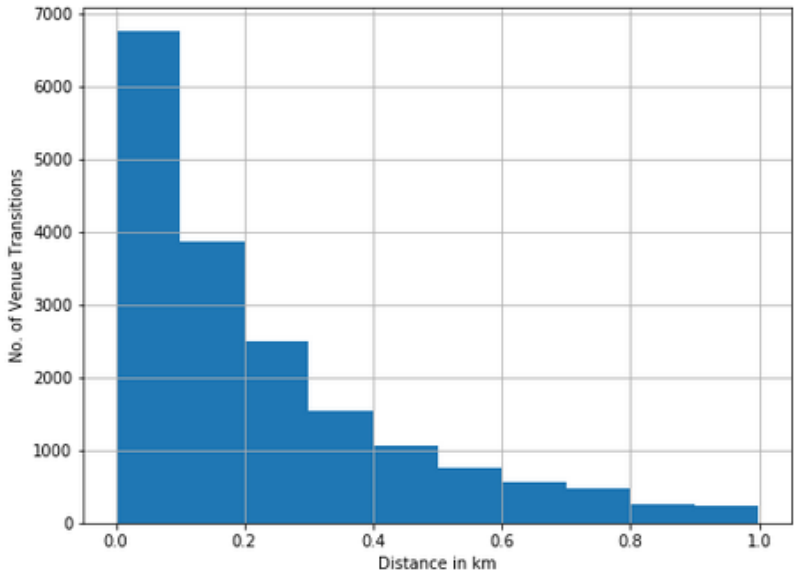

Figure 5: Histogram showing the distance travelled to the next venue. We can see that the histogram is left skewed.

For examples of the data please have a look at the notebook Data Description

3. Methodology

3.1 Problem

The below mathematical formulation of the problem to be solved is taken from Noulas et al. (2011).

The question we ask is: *If person X is engaged in activity Y (i.e. visiting a place belonging to category Y), which activity will follow next?* In particular, we calculate the transition probability $P_t(i, j)$ from category i to category j as:

$$P_t(i,j) = \frac{c_{ij}}{\sum_{k \in \mathcal{C}} c_{ik}} \tag{1}$$

where $\mathcal{C}$ denotes the set of all categories and c_{ij} the number of transitions from a place in category i to one in category j.

3.2 Markov Models

To model activity transitions we can use Markov Models. A Markov model is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

An event in our case, would be the activity that a person is performing at a venue and can be considered to be the category of the venue. To create a Markov model we need to build a transition graph. A node in this graph is the activity that is being performed. A directed link from one node to another represents the transition between those activities. The weight of the link is probability of this transition.

This graph can be represented using a transistion matrix. A transition matrix is an N x N matrix, where each row represents an activity and each column represents the next activity to be performed. An entry in this matrix represents the probability of the activity transition.

Using the data collected from the next venues endpoint of the Foursquare API, we can build such a transition matrix calculating the transition probability using Equation 1.

4. Results

To demonstrate how the Markov Model can be used to model activity transitions, we show below an example with the main categories of the Foursquare API.

As we have noted earlier the Foursquare API maintains categories at various hierarchy levels. An example of category hierarchy is shown below:

- Arts & Entertainment

  Movie Theater

  Drive-in Theater

  Indie Movie Theater

  Multiplex

We can identify a main category if it does not have any parent category. There are 10 main categories defined in the Foursquare API: Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residences , Shop & Service and Travel & Transport. For the venues in San Francisco, we do not find any venue belonging to the category 'Event'.

4.1 The Transition Matrix

The first step to build the transition matrix would be to identify the main categories of the venues for the venue transitions. We can then calculate the transition probabilities between the main categories using Equation 1. The transition matrix is shown in Figure 6.

4.2 The Transition Graph

With the help of the transition matrix, we can easily construct the transition graph as shown in Figure 7. The graph was created using the force directed graph of d3.js. A live working demo of the transition graph is available on jsfiddle

venue_category	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
Arts & Entertainment	0.231465	NaN	0.303797	0.177215	0.147378	0.019892	0.096745	0.023508
College & University	0.055556	NaN	0.388889	0.055556	0.166667	0.055556	0.277778	NaN
Food	0.098659	0.000665	0.366589	0.181909	0.125707	0.009533	0.206518	0.010420
Nightlife Spot	0.072161	NaN	0.363792	0.514091	0.017506	0.005124	0.025192	0.002135
Outdoors & Recreation	0.098407	0.000937	0.259138	0.025305	0.379569	0.028585	0.189784	0.018276
Professional & Other Places	0.082569	NaN	0.362385	0.091743	0.270642	0.022936	0.151376	0.018349
Residence	NaN	NaN	NaN	NaN	1.000000	NaN	NaN	NaN
Shop & Service	0.023245	NaN	0.315567	0.029819	0.096032	0.006340	0.522423	0.006574
Travel & Transport	0.078481	NaN	0.379747	0.075949	0.222785	0.035443	0.139241	0.068354

Figure 6: The transition matrix of the main categories in San Francisco

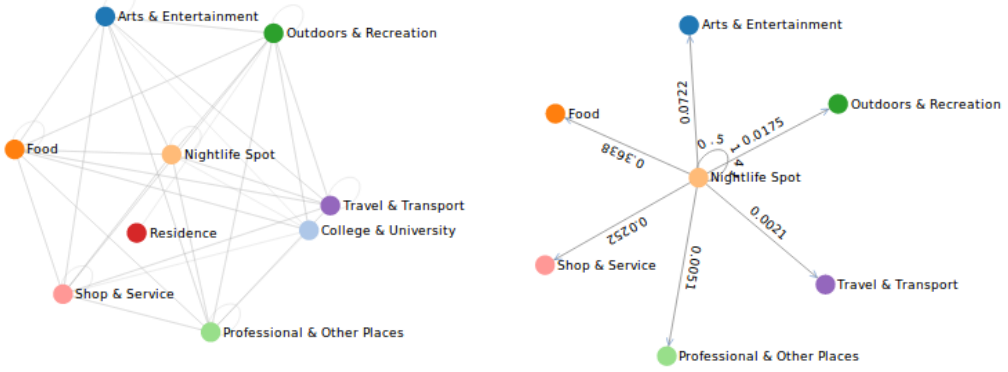

Figure 7: On the left, we see the transition graph for all the main categories without the transition probabilities and directed links. On the right, we see the corresponding transitions if the current category was 'Nightlife Spot'. This graph also shows the transition probabilities allowing us to easily identify the most probable next category, which is again a 'Nightlife Spot'.

5. Discussions

The are a total of 937 categories defined in the Foursquare API. In San Francisco alone venues are categorized into 397 different categories. Such a high dimensional transition matrix is sparse. Take for example, the transitions for Residence in the transition matrix of Figure 6. We can see that all the transitions are to Outdoors & Recreation. There could be several reasons for this like users do not checkin when they are at home. Also when we see Figure 4, there are only 2 venues marked as Residence further indicating that users do not mark their residence and checkin at their residence.

From this we can conclude that there is an unpredictability in user checkins and it is not guaranteed they will always checkin at the venues they visit. This gives us lesser

transitions to calculate accurate probabilities. However, instead of calculating the transition probabilities at a sub-category level, if we calculate it's parent level, then we can use all the transitions from child categories of the parent level to calculate the transition probabilities.

Let's demonstrate this with an example where a person is currently at a Brewery. We want to obtain top 10 activities that the person would do next. Figure 8, shows the activities that would be performed next in terms of the main categories. This is specified using the *level_agg* parameter that is set to 1. The probabilities are shown in percentage. In Figure 9 we set this parameter to 2. This shows a breakdown of the transitions with much more detail as compared to having only the main categories. Similarly, if we further increase the *level_agg* parameter, we get much detailed result (Figure 10).

```
model.get_next_activities("Brewery", n=10, level_agg=1, show_percent=True)
```

next_venue_category	Nightlife Spot	Food	Outdoors & Recreation	Arts & Entertainment	Shop & Service	Professional & Other Places
venue_category						
Brewery	52.845528	25.203252	7.317073	6.504065	6.504065	1.626016

Figure 8: Top 10 next categories after 'Brewery'. Aggregation Level: 1

```
model.get_next_activities("Brewery", n=10, level_agg=2, show_percent=True)
```

next_venue_category	Bar	Brewery	American Restaurant	Food & Drink Shop	Mexican Restaurant	Park	Dessert Shop	Stadium	Pizza Place	Gastropub
venue_category										
Brewery	31.707317	21.138211	5.691057	4.878049	4.065041	4.065041	3.252033	3.252033	2.439024	2.439024

Figure 9: Top 10 next categories after 'Brewery'. Aggregation Level: 2

```
model.get_next_activities("Brewery", n=10, level_agg=3, show_percent=True)
```

next_venue_category	Brewery	Cocktail Bar	Bar	American Restaurant	Dive Bar	Beer Bar	Park	Baseball Stadium	Mexican Restaurant	Ice Cream Shop
venue_category										
Brewery	21.138211	8.130081	7.317073	5.691057	4.878049	4.065041	4.065041	3.252033	3.252033	3.252033

Figure 10: Top 10 next categories after 'Brewery'. Aggregation Level: 3

6. Conclusion

In this project, using Markov models we were successfully able to identify the most probable activities that a person would do next. The model generated is a minimalistic, basic first order Markov model using the *next_venues* endpoint of the Foursqare API. We were also able to get the next activities at different levels of the category hierarchy defined in the Foursquare API. The results were visualized using the transition graph.

A much more detailed model can be built using the checkin data of users. Due to restrictions and data privacy laws in obtaining the checkin data of users from the Foursquare API is rather difficult and can be obtained by scrapping Twitter to search for Foursquare public checkins. As this quite a tedious task and time consuming, I opted for using the data from the *next_venues* endpoint. However, if the checkin data was available we could even build higher order Markov models. For example, a second order Markov model would answer a question like: *If a person first performed activity A and then activity B, what would the person do next?*

With the checkin information we could also tune the model to take into consideration the temporal aspect of when the activities are performed. For example, a person is more likely to visit a Bar during the evening or night rather than in the morning.

References

Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Sui. Exploring millions of footprints in location sharing services, 2011. URL `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2783`.

Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare, 2011. URL `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2831`.