IBM Data Science Professional Certificate

Applied Data Science Capstone

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

March 24, 2019

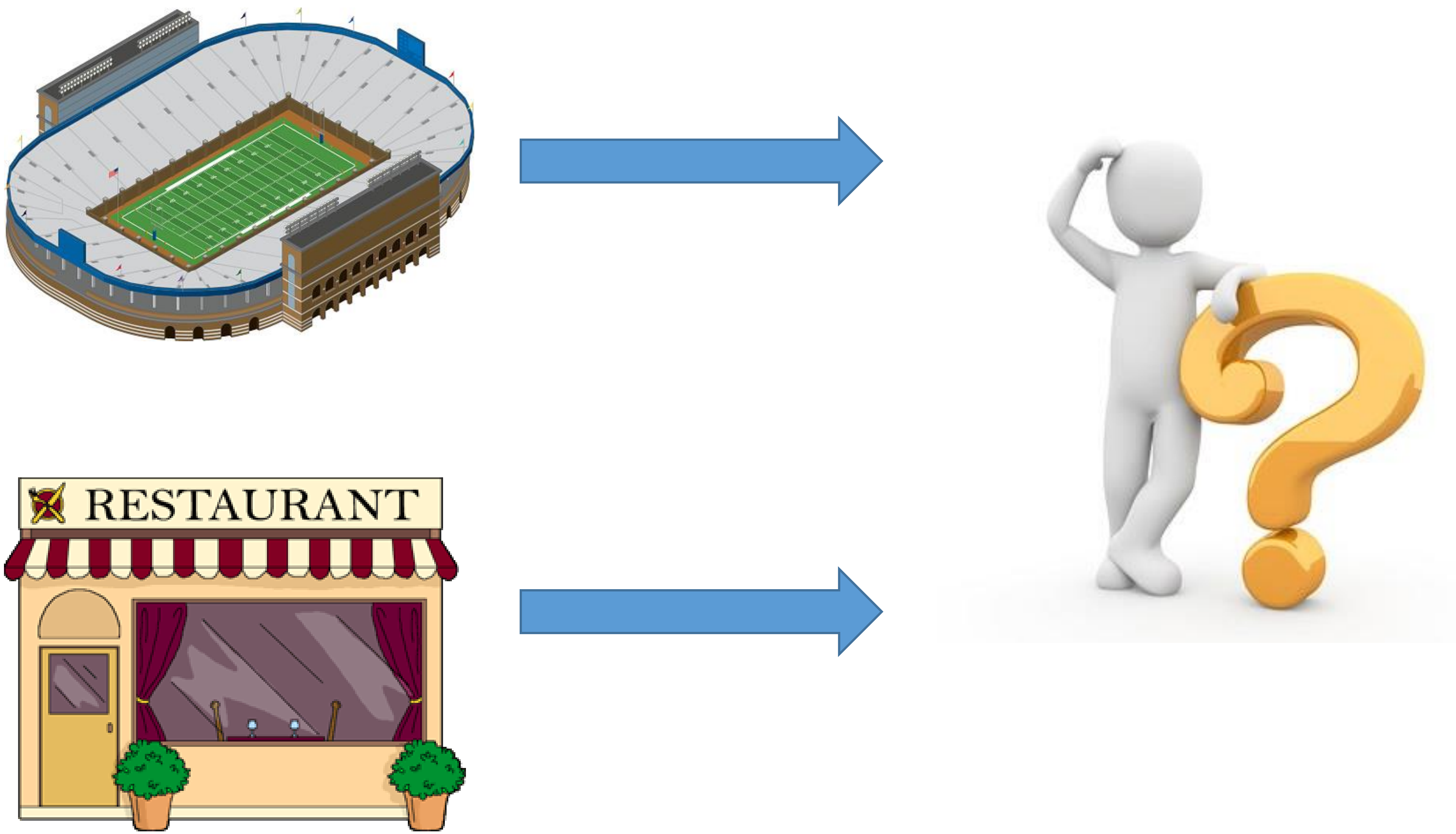
Elson Serrao

Outline

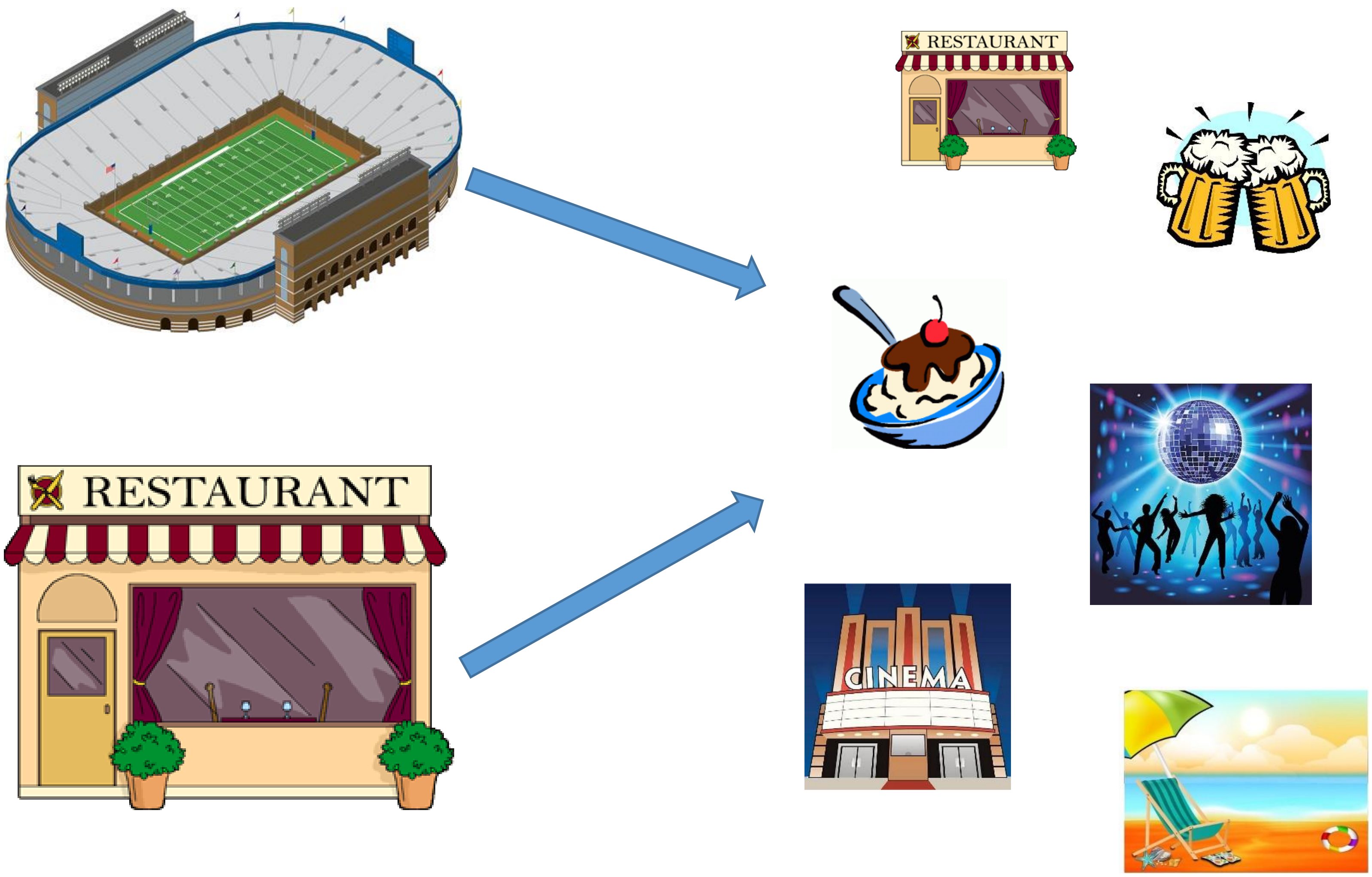
- Introduction
- Related Work
- Data Sources and Analysis
- Methodology
- Results and Discussions
- Conclusions

Introduction

Introduction



Introduction



Introduction

Modelling Activity Transitions answers the question:
If a person is engaged in an activity, which activity would follow next?

Significance of modelling activity transitions in:

- Traffic Forecasting
- Urban Planning
- Epidemiological models of disease spread
- Recommender systems

In this project, we use the Foursquare API to obtain data to model activity transitions.

We use the city of San Francisco for our use case.

The model built is a first order Markov model that provides the most probable next activity the person would do.

Related Work

Related Work

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

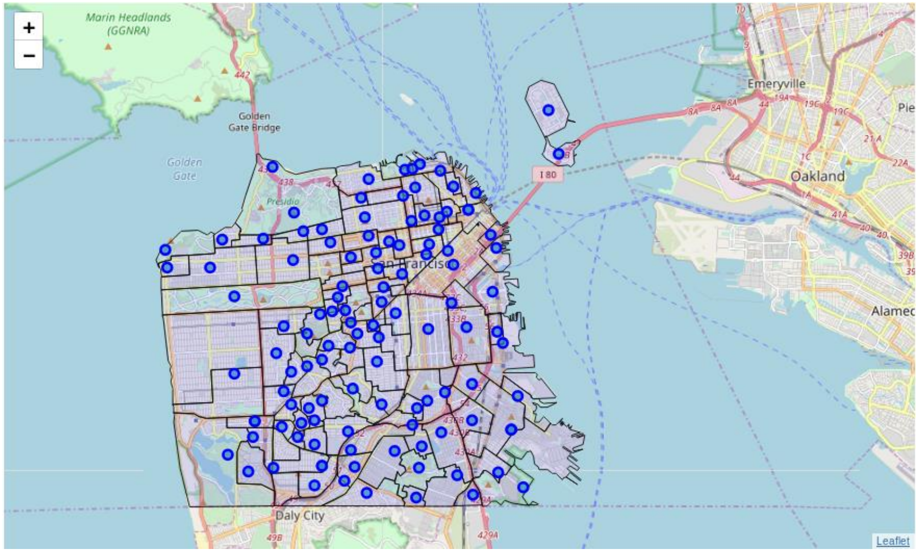
Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare, 2011.
URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2831

Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. The 6th International AAAI Conference on Weblogs and Social Media. Dublin, Ireland, June 2012.
URL http://www.livehoods.org/

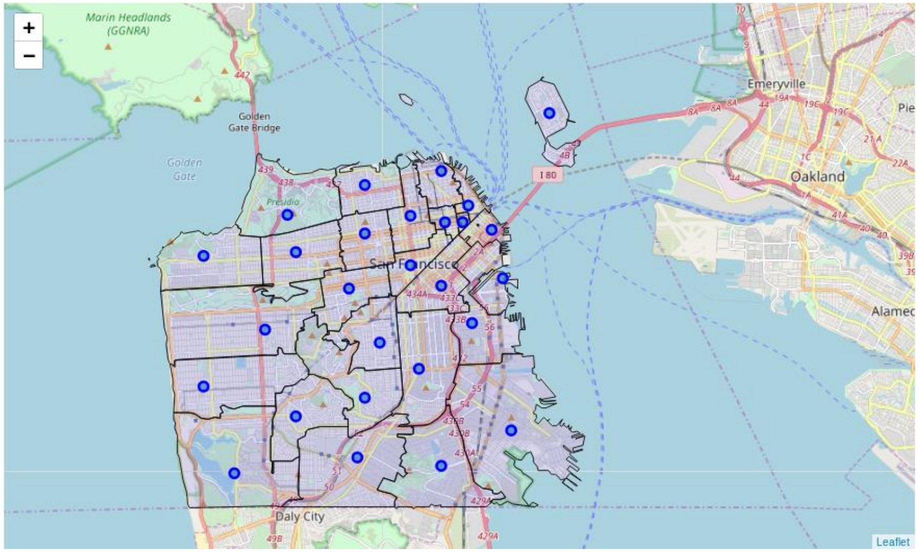
Data Sources and Analysis

City of San Francisco

- Geospatial data for San Francisco obtained from DataSF
- We get the geojson files for
    - Neighbourhoods
    - Postal Codes
- San Francisco has 116 neighbourhoods and 27 postal codes



Neighbourhoods in San Francisco



Postal codes in San Francisco

The Foursquare API

Foursquare offers hosted technology and data to build context-smart, location-aware apps.

It provides an API to access the data about venues, users, photos, tips, checkins and lists.

We mainly use the following endpoints to extract data from Foursquare.

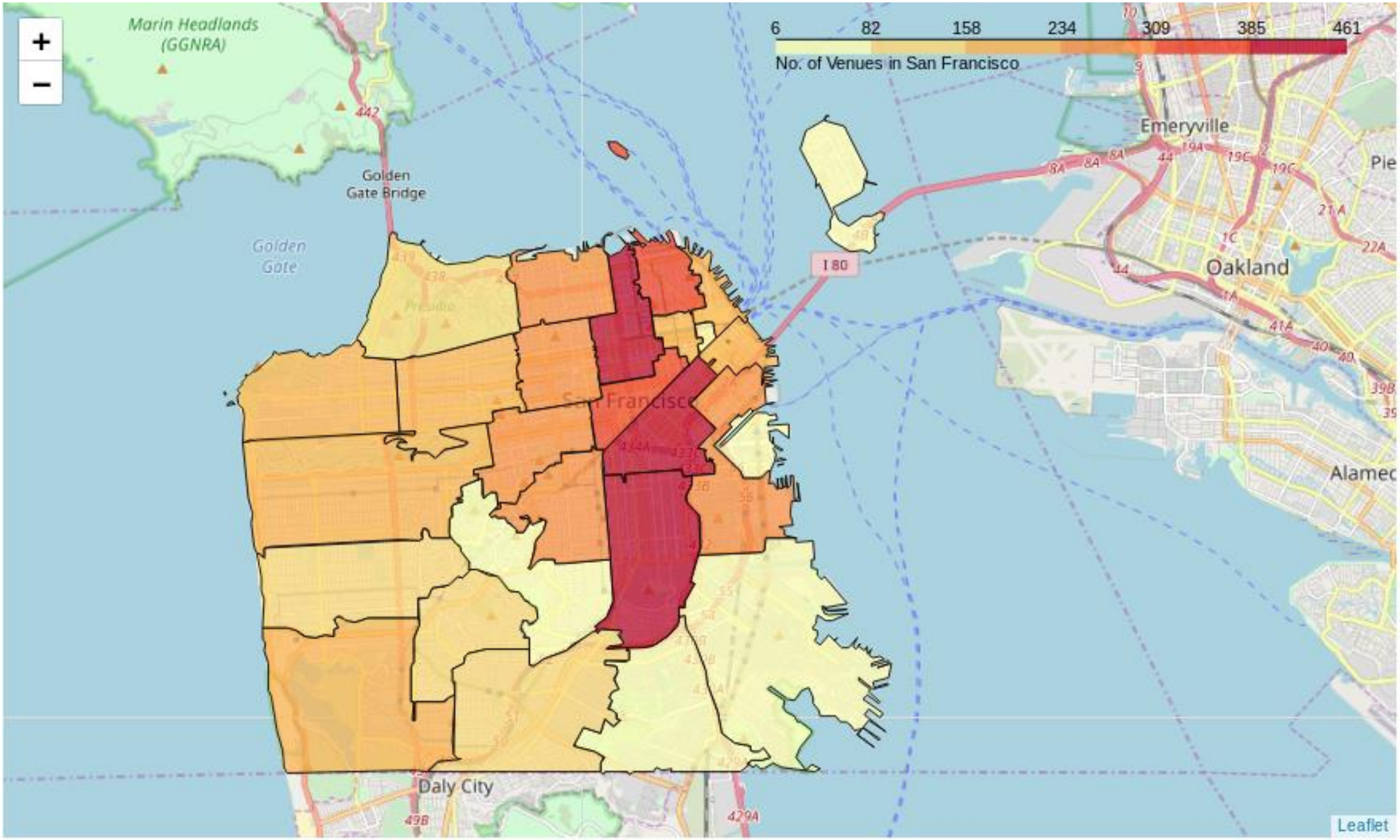
1. The **explore** endpoint: Returns a list of recommended venues near the current location.

2. The **nextvenues** endpoint: Returns venues that people often check in to after the current venue. Upto 5 venues are returned in each query, and results are sorted by how many people have visited that venue after the current one.

We use the explore endpoint to get venues in San Francisco.

We use the nextvenues endpoint to get the transitions of people from one venue to the next.

Distribution of Venues across San Francisco



Venue Categories

Foursquare defines 937 categories and sub-categories

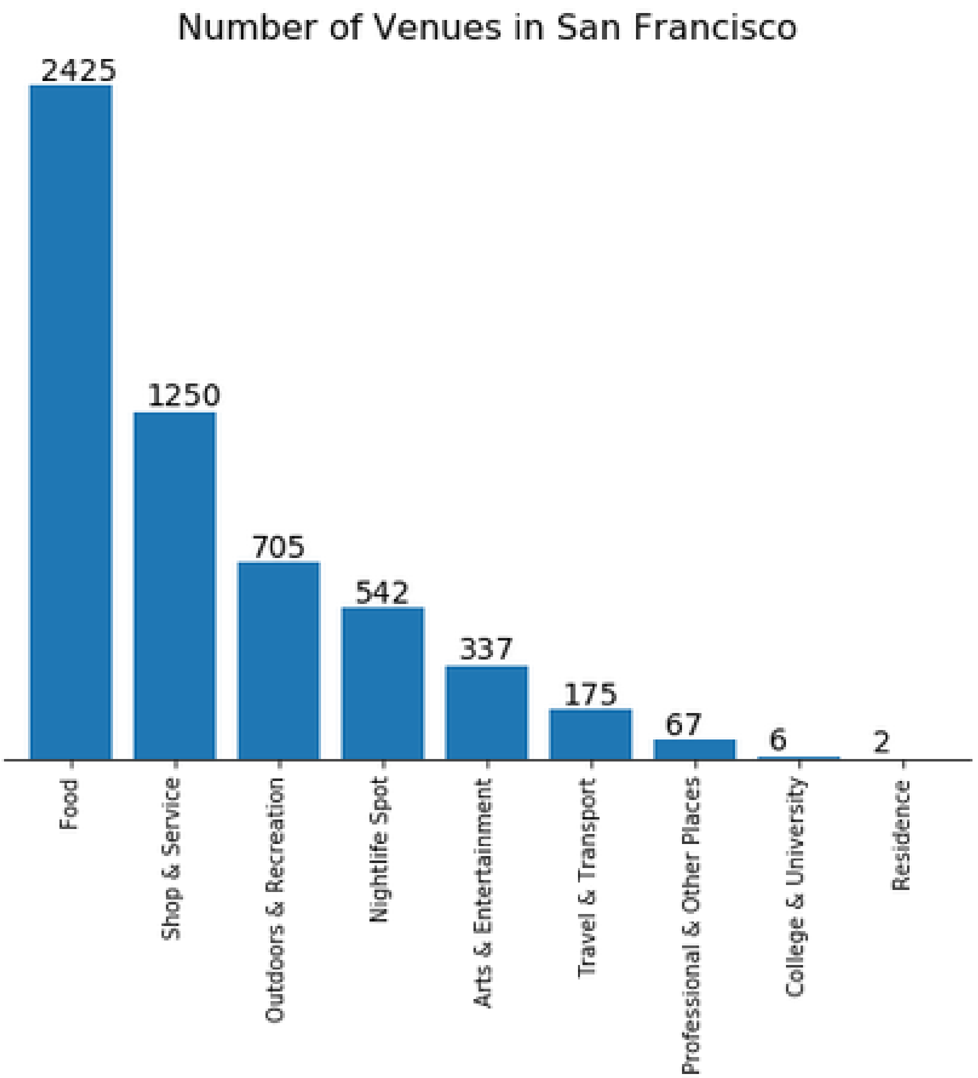
There are 10 main categories

Example of category hierarchy:

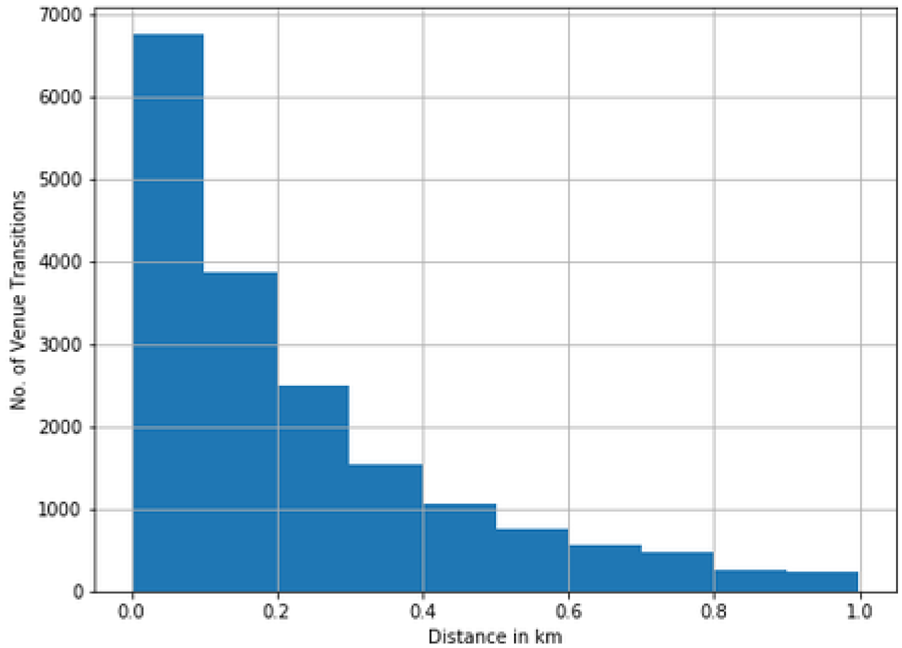
- Arts & Entertainment
  - Movie Theater
    - Drive-in Theater
    - Indie Movie Theater
    - Multiplex

A venue can have multiple categories. However, every venue has a primary category.

In San Francisco, there are 397 different categories among all the venues.



Distance to the Next Venue



Nearly 92.5% of all venue transitions are within a radius of 1km

The activities that people do, are generally restricted within the region

Methodology

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 and is also known as a first order Markov model.

In our case: Event => Activity => Category of the Venue

The Transition Graph:

- Node => Activity
- Links => Transition between activities
- Weight of the link => Transition Probability

The Transition Matrix:

- Row Index: Activity
- Column: Next Activity
- Cell value: Transition Probability

Transition Probability:

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

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

Results and Discusscions

The Transition Matrix

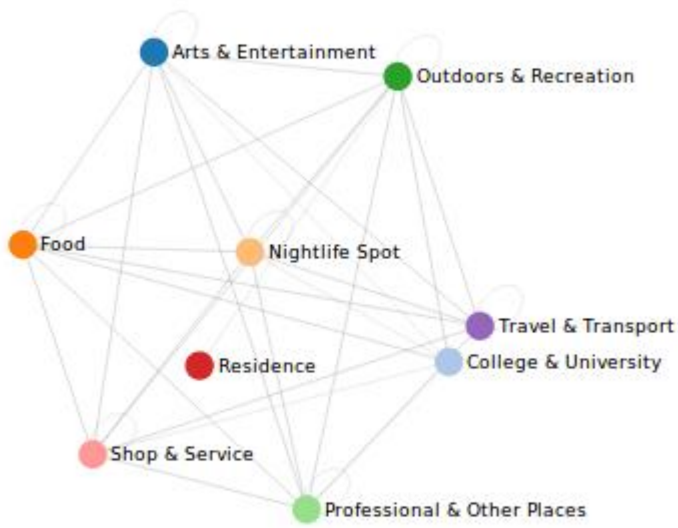
The transition matrix for only the main categories is as shown below.

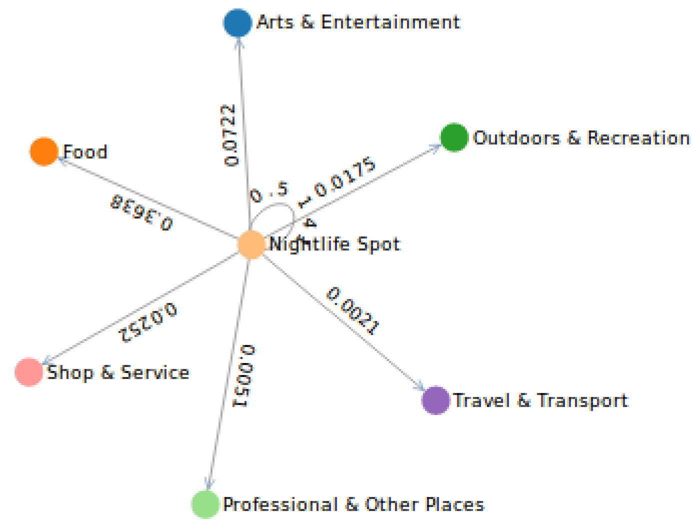
The row index identifies the current activity, while the columns are the next activity

Each value in this matrix is the transition probability.

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

The Transition Graph





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'.

Category Hierarchy

Top 10 activity transitions for 'Brewery'

```
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

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

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

Aggregation Level: 3

Conclusion

Conclusion

We modelled activity transitions using a first order Markov model with the venue transition data from the Foursquare API.

We saw how to use the transition matrix and transition graph to obtain the most probable next activities.

The model built is a very simplistic model.

A much more detailed model can be built using the check in data of users

- Higher order Markov models
- Temporal aspect of when activities are performed

Thank You