Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Table of contents

Introduction: Business Problem

In this project we will try to find an optimal location for a hotel. Specifically, this report will be targeted to stakeholders interested in opening an **hotel** in **Beijing**, China.

Since there are lots of hotels in Beijing we will try to detect **locations that are not already crowded with hotels**. We would also prefer locations **as close to city center as possible**, assuming that first two conditions are met.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decission are:

- number of existing hotels in the neighborhood
- number of and distance to hotels in the neighborhood, if any
- distance of neighborhood from city center

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

- centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using **Google Maps API reverse geocoding**
- number of hotels and their location in every neighborhood will be obtained using **Foursquare API**
- coordinate of Beijing center will be obtained using **Google Maps API geocoding** of well known Beijing location (the Forbidden City)

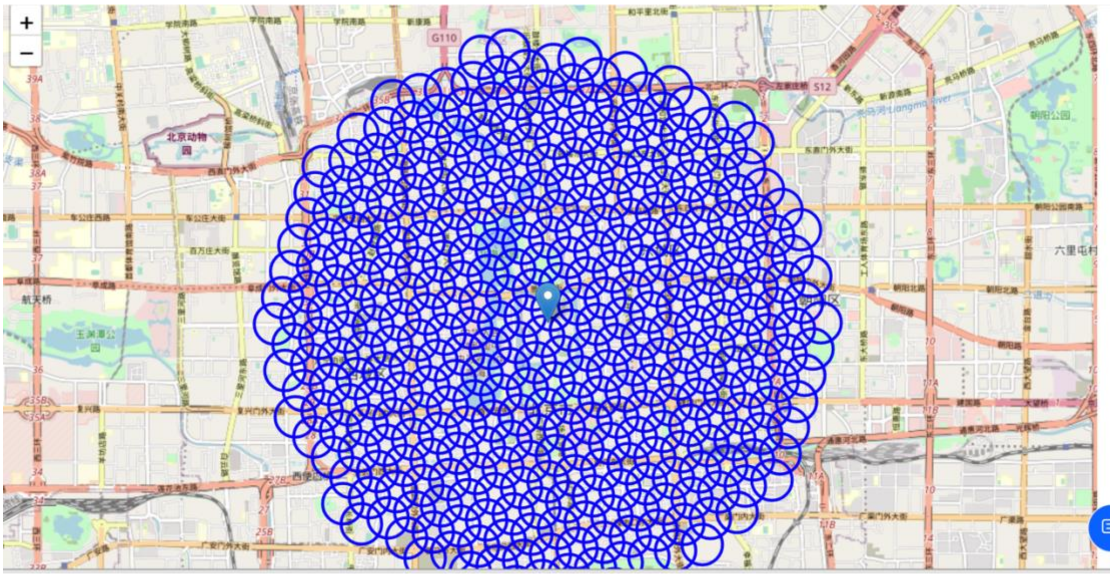
Neighborhood Candidates

Let's create latitude & longitude coordinates for centroids of our candidate neighborhoods. We will create a grid of cells covering our area of interest which is aprox. 12x12 killometers centered around Beijing city center.

Let's first find the latitude & longitude of Beijing city center, using specific, well known address and Google Maps geocoding API.
Now let's create a grid of area candidates, equaly spaced, centered around city center and within ~6km from Beijing Center. Our neighborhoods will be defined as circular areas with a radius of 300 meters, so our neighborhood centers will be 600 meters apart.

To accurately calculate distances we need to create our grid of locations in Cartesian 2D coordinate system which allows us to calculate distances in meters (not in latitude/longitude degrees). Then we'll project those coordinates back to latitude/longitude degrees to be shown on Folium map. So let's create functions to convert between WGS84 spherical coordinate system (latitude/longitude degrees) and UTM Cartesian coordinate system (X/Y coordinates in meters).
Let's create a **hexagonal grid of cells**: we offset every other row, and adjust vertical row spacing so that **every cell center is equally distant from all it's neighbors**.

Let's visualize the data we have so far: city center location and candidate neighborhood centers:



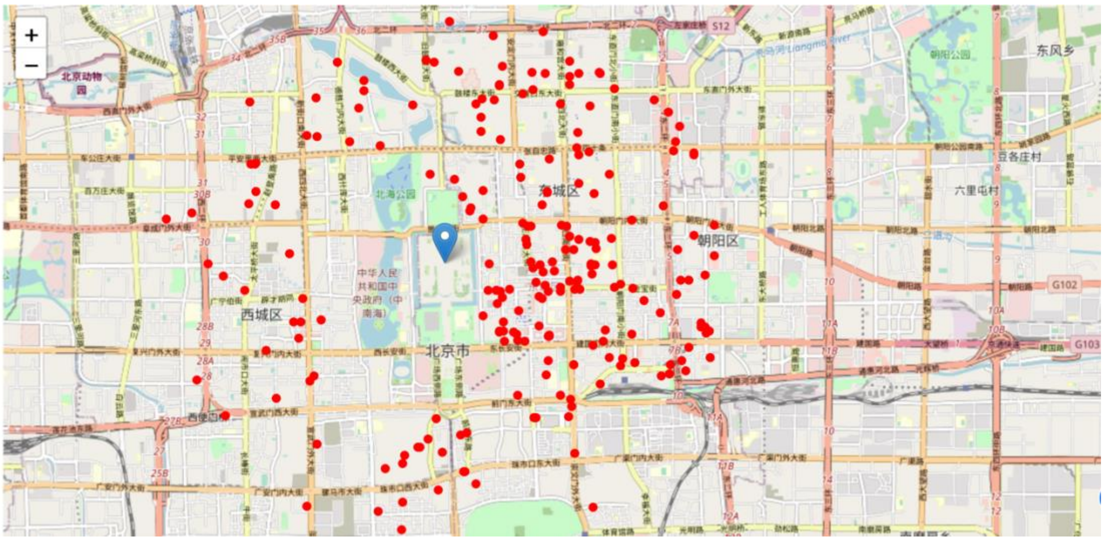
OK, we now have the coordinates of centers of neighborhoods/areas to be evaluated, equally spaced (distance from every point to it's neighbors is exactly the same) and within ~6km from the Forbidden City.

Let's now use Google Maps API to get approximate addresses of those locations.

Foursquare

Now that we have our location candidates, let's use Foursquare API to get info on hotels in each neighborhood.
We're interested in venues in 'hotel' category. Total: 224 Hotels data loaded.

Let's now see all the collected hotels in our area of interest on map.



Looking good. So now we have all the hotels in area within few kilometers from the forbidden city. We also know which hotels exactly are in vicinity of every neighborhood candidate center.

This concludes the data gathering phase - we're now ready to use this data for analysis to produce the report on optimal locations for a new hotel!

Methodology

In this project we will direct our efforts on detecting areas of Beijing that have low hotel density. We will limit our analysis to area ~6km around city center.

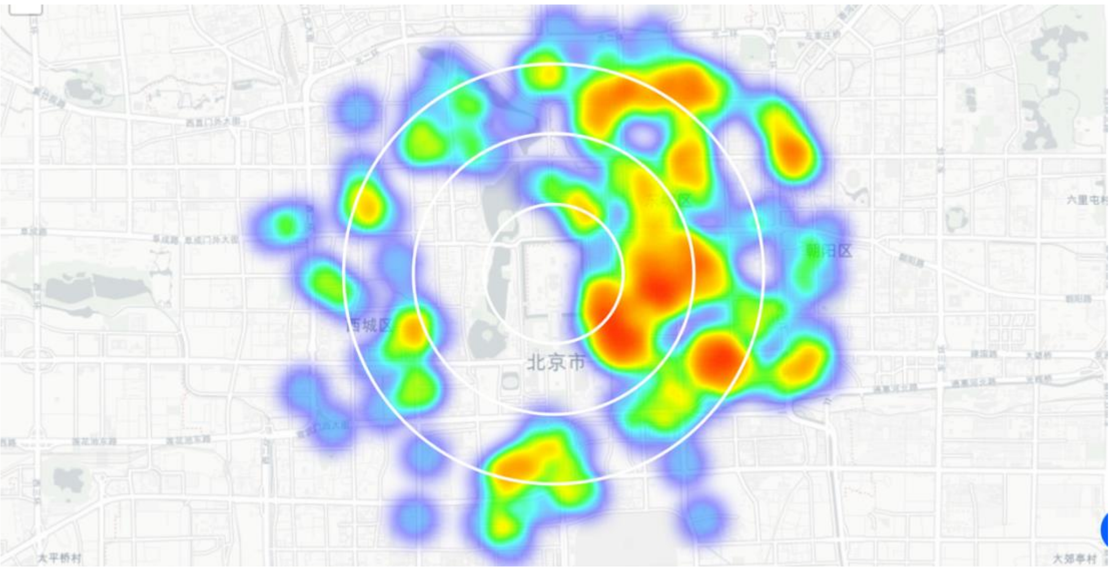
In first step we have collected the required **data: location and type (category) of every hotel within 6km from Beijing center**(the Forbidden City).

Second step in our analysis will be calculation and exploration of '**hotel density**' across different areas of Beijing - we will use **heatmaps** to identify a few promising areas close to center with low number of hotels in general and focus our attention on those areas.

In third and final step we will focus on most promising areas and within those create **clusters of locations that meet some basic requirements** established in discussion with stakeholders: we will take into consideration locations with **no more than two hotels in radius of 250 meters**. We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

Analysis

Let's perform some basic explanatory data analysis and derive some additional info from our raw data. First let's count the **number of hotels in every area candidate**:



Looks like a few pockets of low hotels density closest to city center can be found **west from the forbidden city**.
This map is not so 'hot' but it also indicates higher density of existing hotels directly east from the Forbidden City, with closest pockets of **low hotel density positioned south-west from city center**.

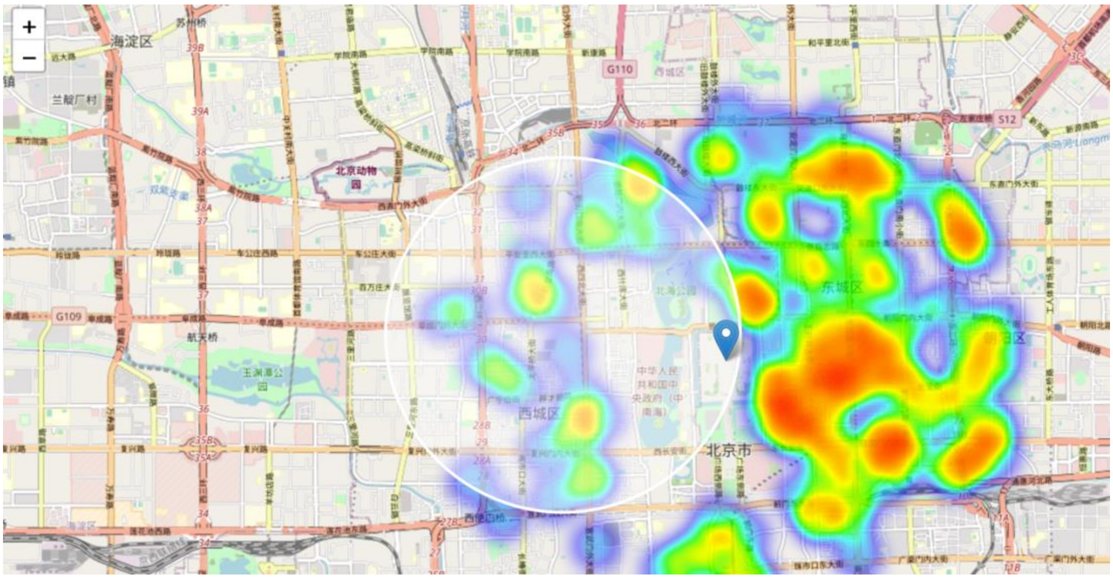
Based on this we will now focus our analysis on areas *south-west from Beijing center* - we will move the center of our area of interest and reduce it's size to have a radius of **2.5km**. This places our location candidates mostly in boroughs **Xicheng**.

Xicheng

Analysis of popular travel guides and web sites often mention Xicheng : Come to Xicheng and enjoy its history, museums, and opera.

Explore the historic area of Xicheng—museums, temples, and sights from Xidan Shopping Center to Capital Museum.

Let's define new, more narrow region of interest, which will include low-hotel-count parts of Xicheng closest to the Forbidden City.



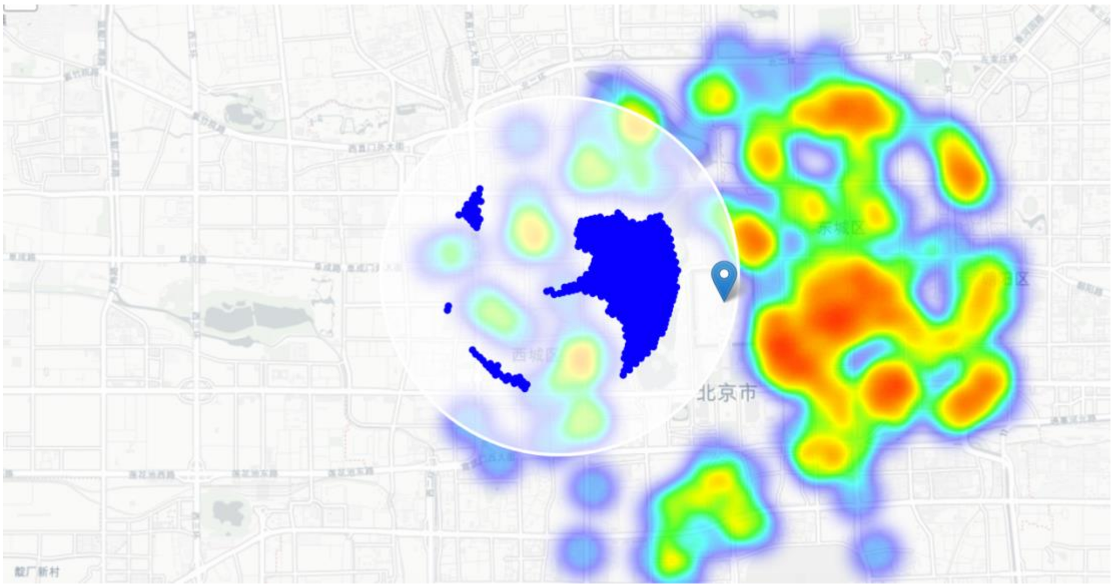
Not bad - this nicely covers all the pockets of low hotel density in Xicheng closest to Beijing center.

Let's also create new, more dense grid of location candidates restricted to our new region of interest (let's make our location candidates 100m appart).

```
2261 candidate neighborhood centers generated.
```

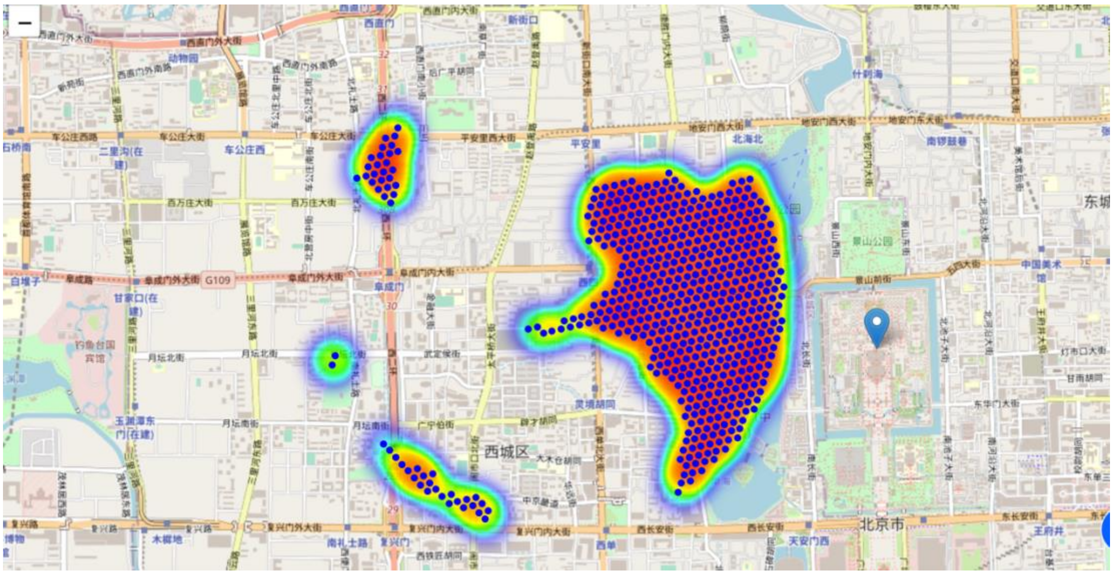
OK. Now let's calculate two most important things for each location candidate: **number of hotels in vicinity** (we'll use radius of **1000 meters**) .

OK. Let us now **filter** those locations: we're interested only in **locations with no more than one hotel in radius of 250 meters**.



Looking good. We now have a bunch of locations fairly close to the forbidden city , and we know that each of those locations has no more than one hotel in radius of 1000m. Any of those locations is a potential candidate for a new hotel, at least based on nearby competition.

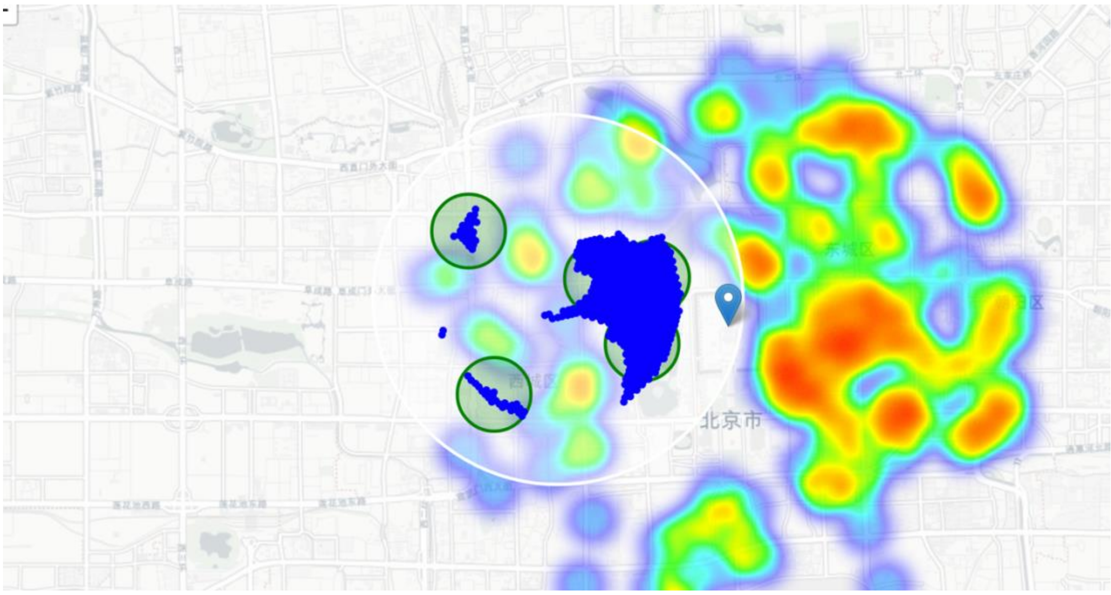
Let's now show those good locations in a form of heatmap:



Looking good. What we have now is a clear indication of zones with low number of hotels in vicinity.

Let us now **cluster** those locations to create **centers of zones containing good locations**. Those zones, their centers and addresses will be the final result of our analysis.

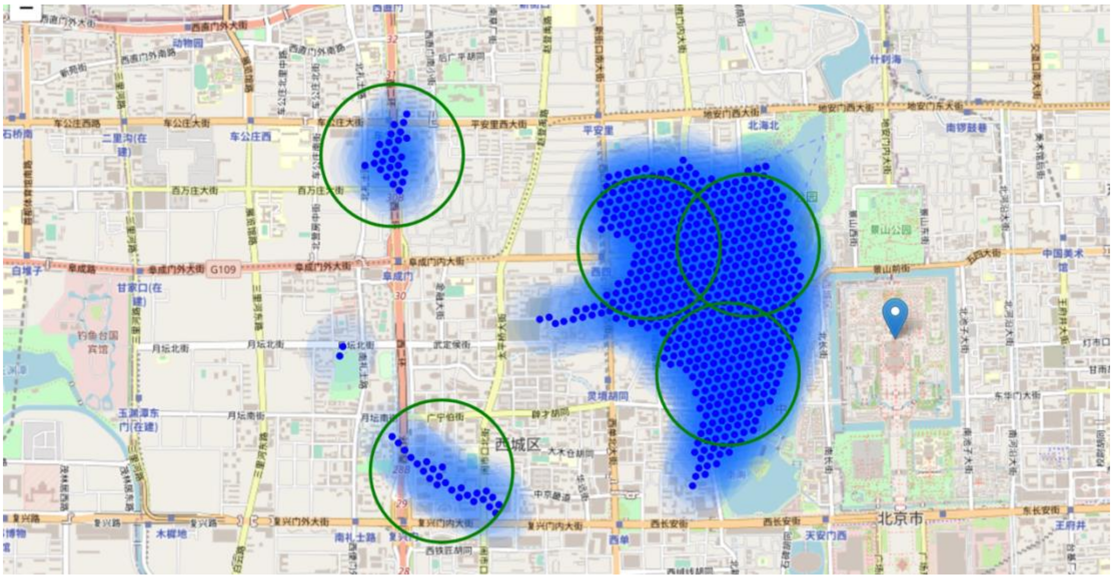
In [120]:



Not bad - our clusters represent groupings of most of the candidate locations and cluster centers are placed nicely in the middle of the zones 'rich' with location candidates.

Addresses of those cluster centers will be a good starting point for exploring the neighborhoods to find the best possible location based on neighborhood specifics.

Let's see those zones on a city map without heatmap, using shaded areas to indicate our clusters:

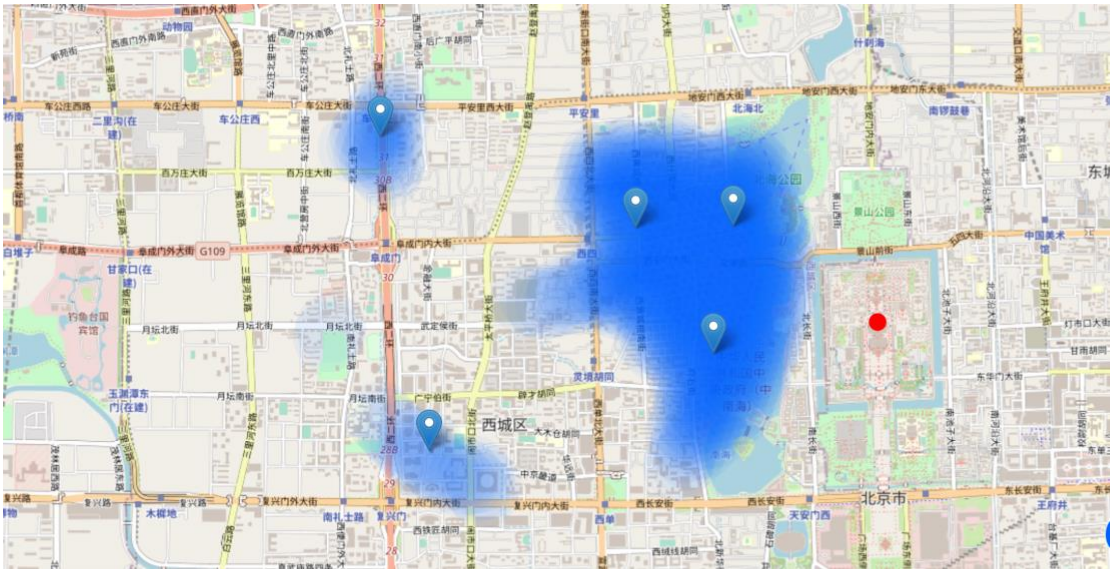


Finaly, let's **reverse geocode those candidate area centers to get the addresses** which can be presented to stakeholders.

```
北海公园, 文津街, 什刹海, 西城区, 北京市, 100032, China 中国          =>
 1.8km from the forbidden city
西黄城根北街, 什刹海, 西城区, 北京市, 100032, China 中国          =>
2.8km from the forbidden city
```

```
中华人民共和国中央政府（中南海），大宴乐胡同，西城区，北京市，100032，China 中国 => 1.8km from the forbidden city
北京市公安局公共交通管理局，1，阜成门北大街，西城区，北京市，100032，China 中国 => 5.7km from the forbidden city
都城隍庙后殿，金融大街，西城区，北京市，100032，China 中国          =>
 5.0km from the forbidden city
```

This concludes our analysis. We have created 5 addresses representing centers of zones containing locations with low number of hotels , all zones being fairly close to city center (all less than 6km from the forbidden city). Although zones are shown on map with a radius of ~500 meters (green circles), their shape is actually very irregular and their centers/addresses should be considered only as a starting point for exploring area neighborhoods in search for potential hotel locations. Most of the zones are located in Xicheng boroughs, which we have identified as interesting due to being popular with tourists, fairly close to city center and well connected by public transport.



Results and Discussion

Our analysis shows that although there is a great number of hotels in Beijing (~200 in our initial area of interest which was 12x12km around the Forbidden City), there are pockets of low hotel density fairly close to city center. Highest concentration of hotels was detected east from the forbidden city, so we focused our attention to areas west, corresponding to Xicheng borough. After directing our attention to this more narrow area of interest (covering approx. 5x5km west from the forbidden city) we first created a dense grid of location candidates (spaced 100m appart); those locations were then filtered so that those with more than one hotel in radius of 1000m were removed.

Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all this is 5 zones containing largest number of potential new hotel locations based on number of and distance to existing venues - hotel. This, of course, does not imply that those zones are actually optimal locations for a new hotel! Purpose of this analysis was to only provide info on areas close to Beijing center but not crowded with existing hotel - it is entirely possible that there is a very good reason for small number of hotels in any of those areas, reasons which would make them unsuitable for a new hotel regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

Conclusion

Purpose of this project was to identify Beijing areas close to center with low number of hotels in order to aid stakeholders in narrowing down the search for optimal location for a new hotel. By calculating hotel density distribution from Foursquare data we have first identified general boroughs that justify further analysis (Xicheng), and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby hotels. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decission on optimal hotel location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.