Travel Partner:
Route Planning in Vancouver with OSM Data

CMPT 353 Final Project Report

Caijie Zhao 301354256
Huiyi Zou 301355563
Yao Tong 301336982

1 Introduction

Vancouver is one of the most popular tourist destinations in North America. It is a travel-friendly city with a mix of cultures and beautiful natural scenery, which attracts lots of tourists. There are many concerns from them about how to have a wonderful trip to a strange city. The plan for travelling routes and trying local cuisine can mainly determine the travelling quality when travellers come to a new city. Based on their concerns, our group has designed a complete guideline for them. There are three parts of the project, recommending interesting routes for tourists before the trip, finding the nearest restaurants during travelling with the user's current location, and presenting the possible amenities that might visit before by inputting photos with locations.

2 Material

2.1 Datasets

Datasets are OSM data in Vancouver collected from the OpenStreetMap project, providing detailed information on amenities. All components in the project have extracted the desired information from OSM datasets, and some parts achieved more data to complete their own goal.

For planning routes, the program gains more tourist attractions along with rates and reviews by using google maps and places API. For sceneries from OSM, place API is applied to find its place id and then find its reviews based on ids. The additional data is achieved by specifying a centre point in Vancouver and returns all attractions with place id within a 50000m radius. Besides, our group has collected many geotagged images that can present the tour to figure out visited amenities for the after-trip part.

2.2 Cleaning Data

Planning for interesting routes

Text and data cleaning is applied to the collected dataset. For cleaning OSM data, we have filled some missing names based on their wikidata tag and selected the tourist attraction. There are some attraction names in OSM that represent the same location, like" Stanley Park - bicycle" and "Stanley Park-parking", so we have changed those data to the same names. For additional data, sceneries outside of Vancouver are excluded.

The second step is to clean reviews. We have removed URLs and punctuations from the texts before splitting them. The stopwords are removed, spell checking and lemmatization are applied. We have analyzed sceneries with a rating greater than 4.6.

Nearest Restaurants

Inside the OSM dataset, based on some amenities, the first step is to choose all the restaurants. Then, we use the latitude and longitude in the dataset to calculate the distance between the restaurant and the current location and choose all the values less than 300m.

According to the taste of travellers, we are also interested in the distribution of different cuisine restaurants in Vancouver. Using the cuisine tags in the dataset, we calculate the number of restaurants in Vancouver that offer the same food.

For the third part, we focus on the distribution of chain restaurants and independently-owned places. For all raw data, the wikidata tags can help us to accurately extract restaurants with the same name.

Possible Amenities Visited

For this part, we ignore the amenities without names, while the OSM file given has many missing values of the name field. Avoiding to throw away too many records, we try to fill the name attribute. There are "official_name", "operator", "wikidata" and "wikipedia" tags in OSM data, which help us find their name. After filling the name, we bring about 800 records back to life.

The location and date taken information from Exif data embedded in JPEG images provided by the user are extracted. To generate the user's route in order, we format datetime objects for photo taken time.

3 Methods

3.1 Planning for interesting routes

Sentiment Analysis: Polarity and Subjectivity

We have computed the polarity and subjectivity scores to quantify the review. The polarity is floating within [-1,1] to reflect if the command with positive, neutral or negative attitudes toward attractions. The overall sentiment is inferred as positive, neutral, or negative based on the sign of the polarity. The range of subjectivity is in the [0, 1] where 0 represents an objective command. The scores are computed by unsupervised lexicon-based approaches from the TextBlob library [1].

Based on those scores, we have designed a function to define a new rate (1). The reviews have higher rates when they contain more positive and objective information. The popularity depends on the preference and the number of visitors. The heat score is computed by (2).

The new rate and heat functions are:

rate= original rating *5+polarity*3+subjectivity*-2 (1)

heat=original rating*5+the number of ratings *3 (2)

Where the original rating and the number of ratings from datasets.

Shortest Time Routes Analysis

After selecting the top-5 attractions, we have computed the most efficient routes for travelling to all selected locations exactly once from the origin and returning to the original point, which is a Traveling Salesman Program (TSP). We have calculated the transportation time for each pair of locations via Google Map API and utilized a backtracking method to solve TSP.

The idea of backtracking is to search by depth-first. When the algorithm searches a node from the original point, the algorithm keeps searching down if the node may be a part of the solution. Otherwise, it returns to its ancestor node and tries other paths. Since the program should return an optimal solution to meet our goal, the algorithm expands all solution space trees and chooses the shortest routes [2].

3.2 Nearest Restaurants

Haversine Algorithm

It is used to calculate the distance between 2 points by using latitude and longitude.

$$a = sin^2\left(\frac{\Delta\varphi}{2}\right) + cos\varphi 1 \cdot cos\varphi 2 \cdot sin^2(\frac{\Delta\lambda}{2}) \quad (3)$$

$$c = atan2(\sqrt{a}, \sqrt{1-a}) \quad (4)$$

$$d = 2Rc \quad (5)$$

Where φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km).

3.3 Possible Amenities Visited

Nearby Amenities Analysis

After generating image information, it is easy to form the user's route by the time taken data. Then we find amenities in OSM, which are along the tour of any two connected nodes in an area with a 100-meter radius, that the user might visit, through checking whether the distance from an amenity to the route formed by two connected nodes is less than 100 meters. The amenities close to the nodes are found by applying the Haversine algorithm directly. For other amenities' locations, we calculate the distances between the nodes and amenities, and now it becomes a math problem - finding the height of a triangle with known three sides, which can be solved by Heron's formula:

$$Area = \sqrt{s(s-a)(s-b)(s-c)} \quad (6)$$

Where a, b, c are lengths of a triangle, and $s = \frac{a+b+c}{2}$ is the semi-perimeter of the triangle [3].

Assuming the distance between two connected nodes is a, since $Area = \frac{ha}{2}$, then the distance $h = \frac{2Area}{a}$. We can choose amenities with small heights. Nevertheless, if this triangle is obtuse, the result may be wrong - the amenity is close to the line formed by two connected nodes but far away from the nodes. Thus, the triangle must have $|b^2 - c^2| \leq a^2$.

4 Results

4.1 Planning for interesting routes

Figure 1, zoom-in map, and figure 2, zoom-out map, show the recommended route from the original location (49.281579, -122.996366) after sentiment and shortest route analysis. Based on the rate calculated from (1), the shortest route of recommended tourist attractions is Origin->Whytecliff Park| West Vancouver->Stanley Park->The Orpheum->Spanish Banks Beach Park->Golden Ears Provincial Park->Origin. The predicted transportation time is 226 minutes.

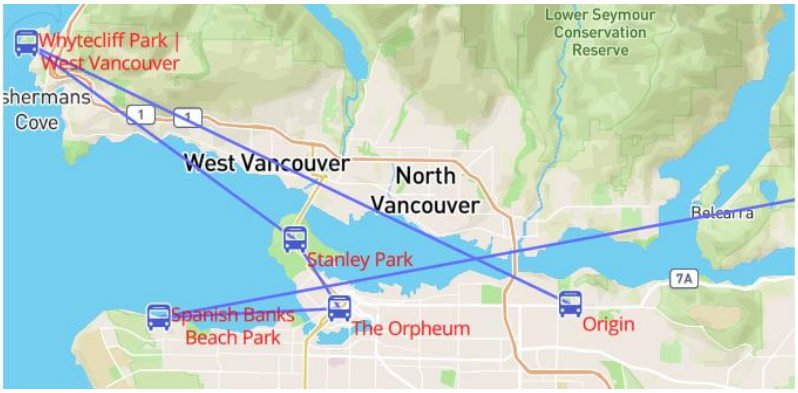

Fig.1. an Interesting Route from the Origin (49.281579, -122.996366) (Zoom In)

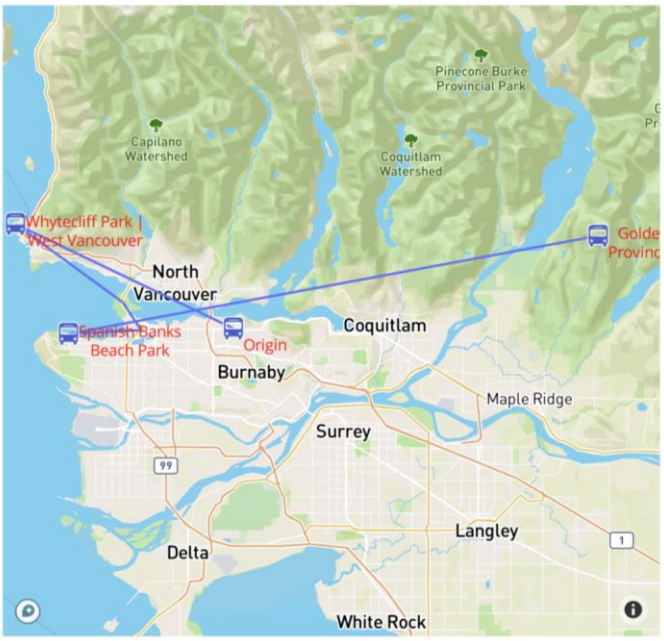

Fig.2. an Interesting Route from the Origin (49.281579, -122.996366) (Zoom out)

Figures 3 and 4 demonstrate the word frequency of reviews from top-5 and last-5 locations, respectively. Since sentiment analysis processes the tourist attractions with original ratings higher than 4.6, most adjectives are positive in both wordclouds. There are some differences between them. For top-5 locations, the wordcloud contains more positive descriptive words with higher frequency, such as ‘great’, ‘beautiful’ and ‘good’, than the last-5 wordcloud. The last-5 wordcloud has more nouns and verbs compared to the top-5, but reviews contain a high frequency of word ‘great’. The ‘great’ is the highest frequency adjective in both top-5 and last-5 reviews.

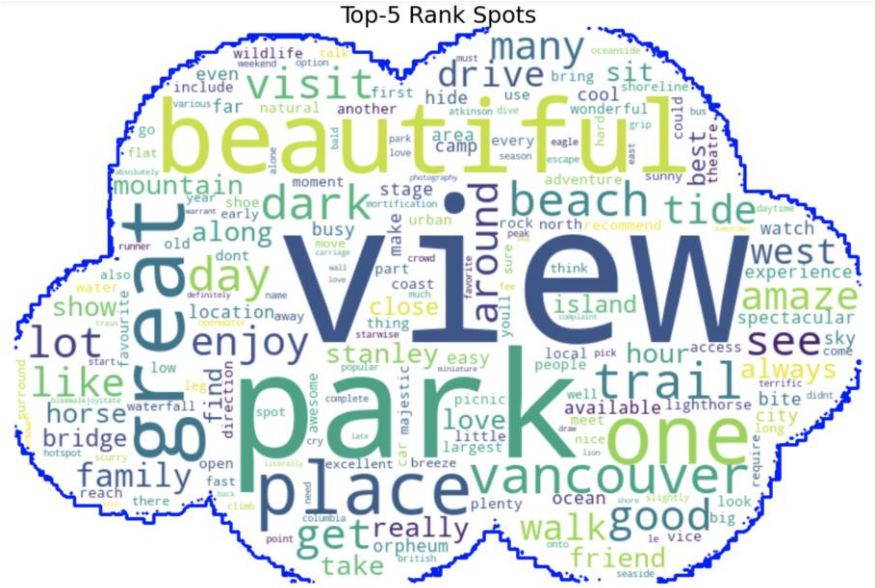

Fig.3. A WordCloud for Reviews from Top-5 Attractions

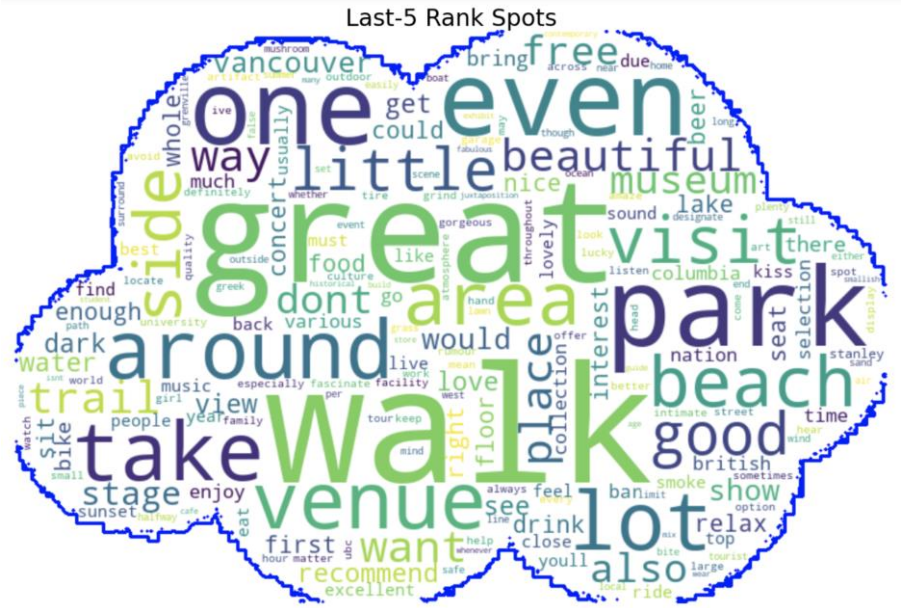

Fig.4. A WordCloud for Reviews from Last-5 Attractions

4.2 Nearest Restaurants

Figure 5 shows all restaurants within 300 meters from the starting point, latitude = 49.282761666666666, and longitude = -123.12364166666666.

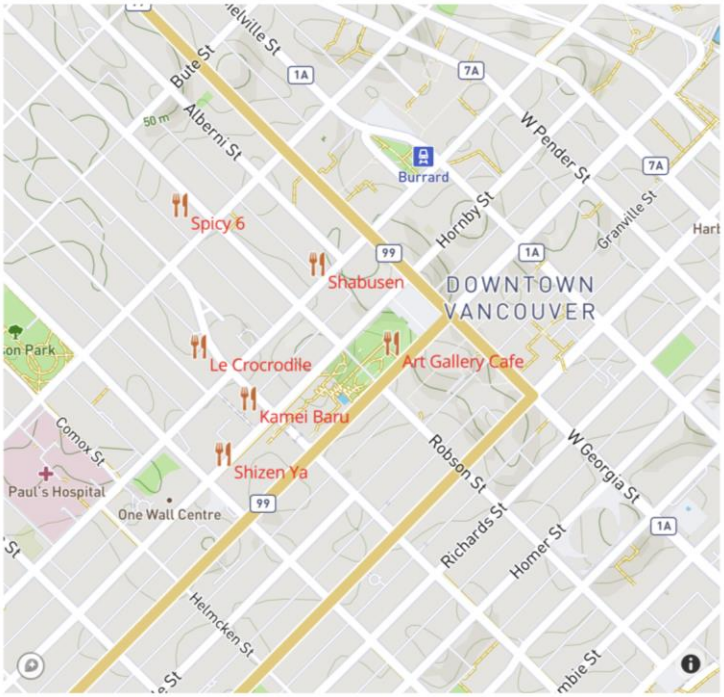

Fig.5. Nearest Restaurants around the origin

Figure 6 and 7 displays the distribution of different cuisine in Vancouver. After filtering the data in OSM, we find the number of Chinese restaurants is the largest, and most of them are located in Downtown. Besides, most other cuisines are also concentrated in the city centre, and some are scattered in the other business districts.

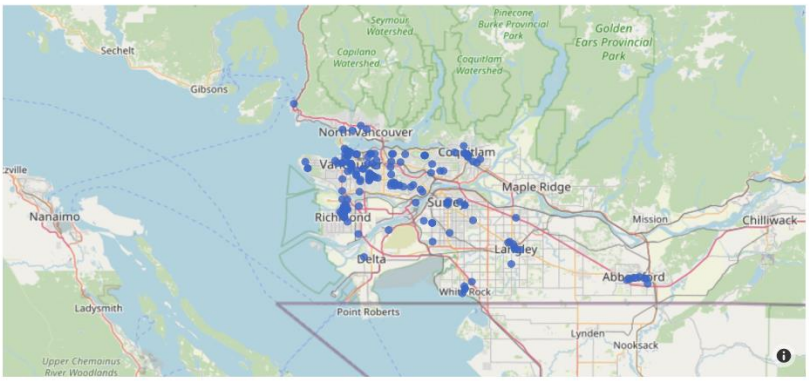

Fig.6. The Most Restaurant with Cuisine: Chinese

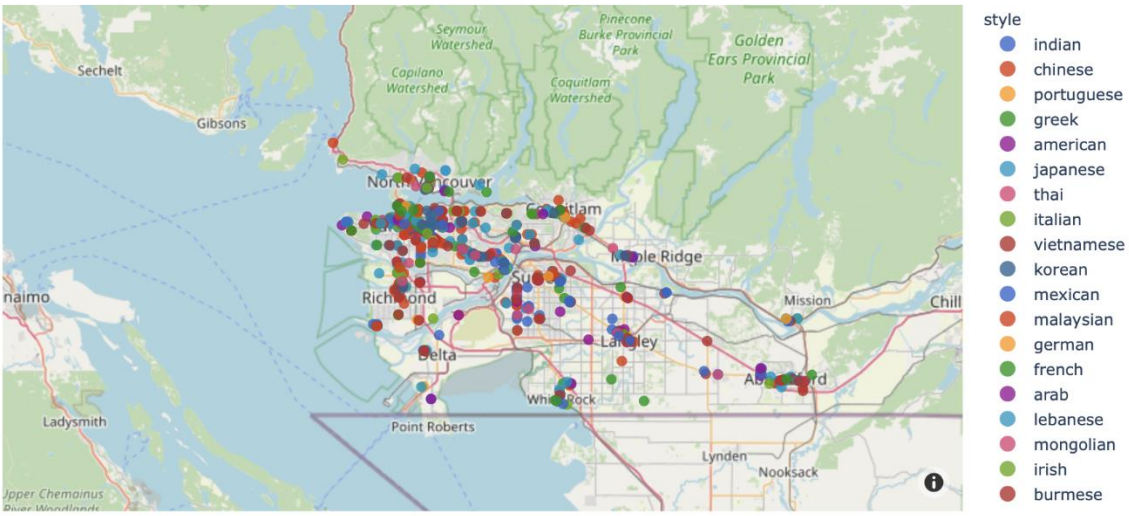

Fig.7. The Different Cuisine Restaurants in Vancouver

Comparing Figure 8 and 9, we find that there is a large number of independently-owned restaurants in Vancouver. Both chain restaurants and non-chain restaurants are located in commercial districts, and their distribution becomes more and more scattered as the distance from the city centre increases.

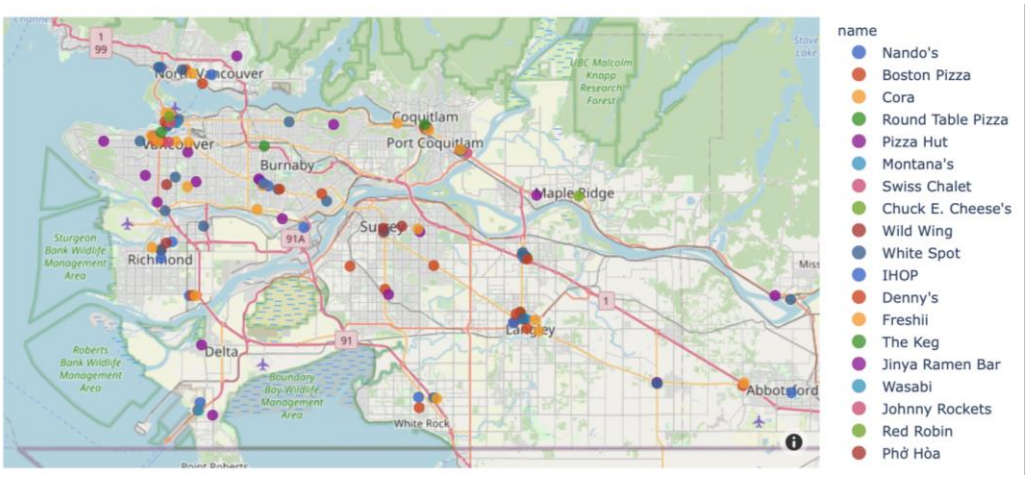

Fig.8. The Chain Restaurant in Vancouver

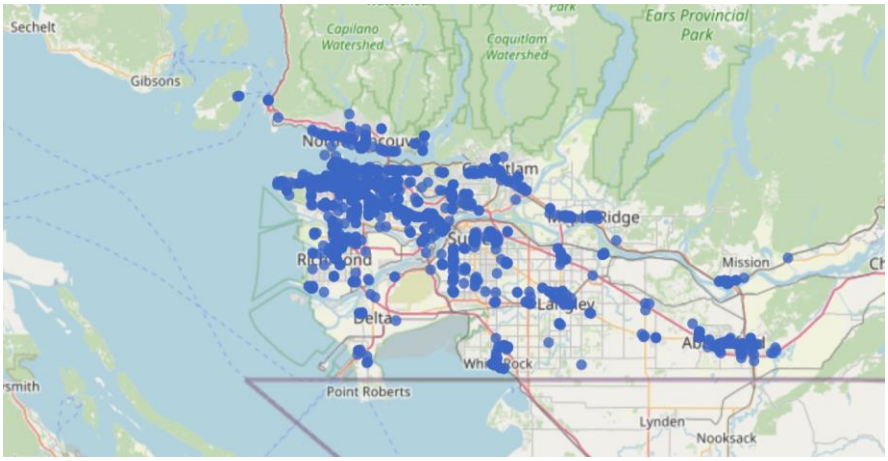

Fig.9. The Non-Chain Restaurant in Vancouver

4.3 Possible Amenities Visited

Figure 10 has all amenities from provided OSM data shown on a street map. The amenities are dense in the Downtown Vancouver area, which is the business district. The number of amenities is associated with economic development.

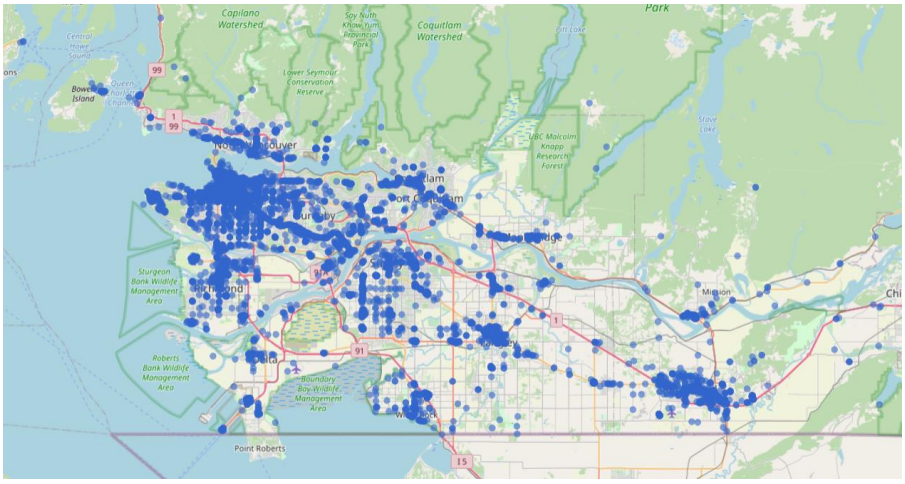

Fig.10. OSM Location in Vancouver

We collect 7 photos taken in Downtown Vancouver on one day with location data and want to figure out the amenities nearby. Figure 11 shows the locations where the photos were taken.

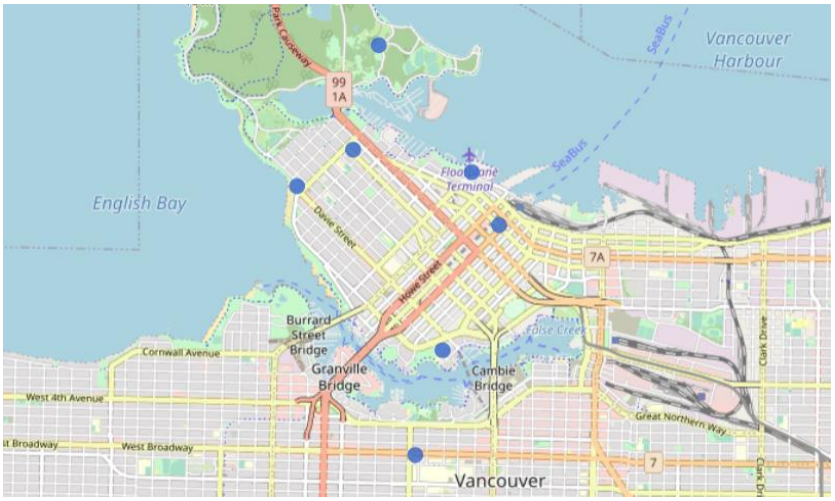

Fig.11. Locations in Input Image

Figure 12 and 13 illustrates the route of the user's trip based on the images' information with their nearest amenities where the user might take photos and any amenities along the user's route in an area with a 100-meter radius. If the distances between the nodes are not close enough, the result may not be accurate - people do not walk in a straight line.

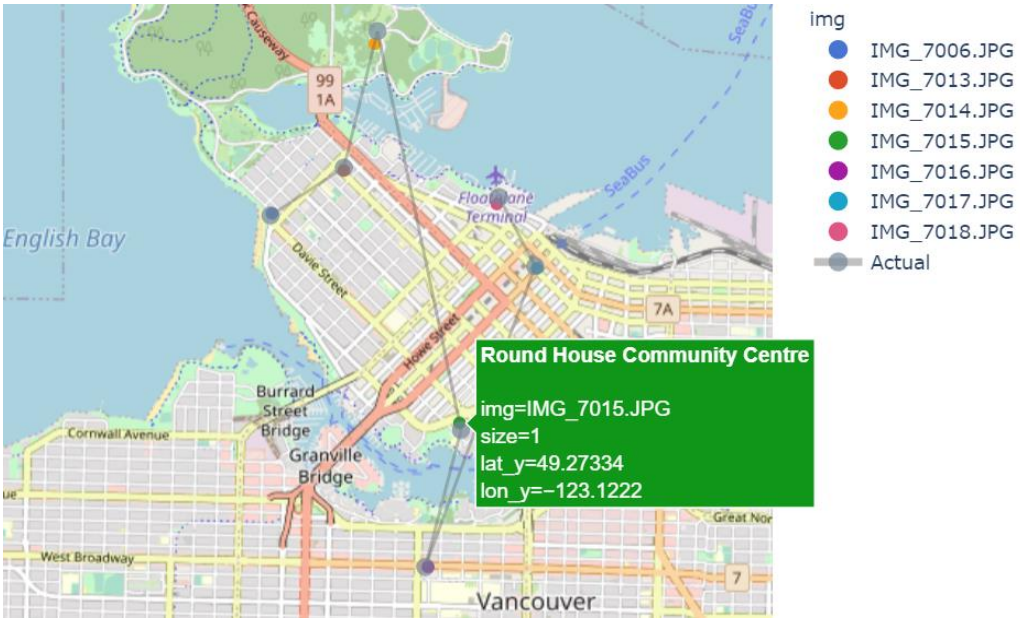

Fig.12. The User's Route with Nearest Amenity in Provided OSM

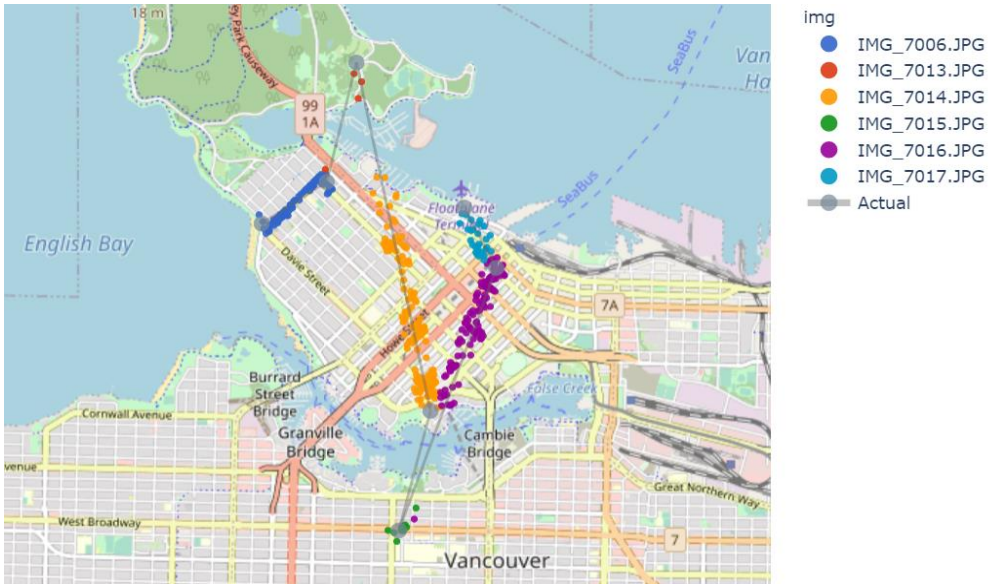

Fig.13. The User's Route Generated from Images with Near Amenities

5 Conclusions

Our project provides high-quality but cost-effective travel plans for tourists to Vancouver. Users can input the GPS coordinates of their departure and get their own recommended route depending on ratings or popularity. When on the trip, the program can get the users' position and show restaurants nearby. If the users take photos when travelling, they will know any possible places they might pass.

6 Limitations

The OSM dataset lacks lots of important data. Most of the data in the label have different attributes, and in most cases, data attributes are irregular, which is difficult to filter the data.

When we are filtering the restaurant data, we find data values are similar but not the same. Therefore, when performing data integration, there are repeated classifications that will interfere with data analysis. For addressing this issue, we should achieve extra unique information about each restaurant through API or by finding more datasets.

We get the latitude and longitude from the photos we took, and we found the latitude and longitude provided by the OSM dataset has a big gap with our photos' data. Since the accuracy decreases as

the distance of the picture is taken, we need tighter data for latitude and longitude. We need to optimize our algorithm to achieve more efficiency and accuracy. And it is possible to improve our program to analyze GPX data, which will produce more reliable and accurate results.

7 Accomplishments

Worked in a group of three to implement a python application for providing travelling guidelines

Caijie Zhao

- Achieved data about tourist attractions via API and cleaned text and data used for faster and meaningful analysis.
- Gained insight from tourists' reviews through sentiment analysis and defined new functions to compute rates used to decide top-rate sceneries.
- Utilized backtracking algorithm to find the shortest path for travelling all recommended locations once to give visitors guidelines about trips.
- Designed maps for presenting the optimal tour and visualized word distributions in top-5 and last-5 tourist attractions.

Yao Tong

- Detected travelers' ip addresses and used haversine function to select the restaurant closest to travelers.
- Appropriated screening to integrate rough data to facilitate data analysis.
- Extracted the data to analyze the distribution of cuisine in Vancouver and chose the restaurant with the largest number.
- Filtered the wikidata tag to choose all chain restaurants and non-chains and visualized their density in the graph.

Huiyi Zou

- Retrieved Wikidata entry via Wikidata API from wiki tags and filled missing values of name for amenities.
- Extracted Exif data from images via GPSPhoto and Exifread module in Python for generating user's route by their locations.
- Calculated distances between amenities and paths by applying Heron's formula to select amenities along route in area with 100-meter radius as possible visited places.
- Generated street maps with paths and nearby amenities for analysis and visualization results.
- Refactored code for consistency and replaced duplicative code with functions.

8. References

[1] ‘Emotion and Sentiment Analysis: A Practitioner’s Guide to NLP,’ *KDnuggets*, 2018. [online]. Available: https://www.kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html#:~:text=The%20key%20aspect%20of%20sentiment,sign%20of%20the%20polarity%20score. [Accessed: 11-Dec-2020].

[2] ‘Backtracking method and branch and bound method help you solve traveling salesman problems (TSP),’ *copyfuture*, 01-Jul-2020. [online]. Available: https://copyfuture.com/blogs-details/202007010819540158 0fd9la5l55wut2.[Accessed: 11-Dec-2020].

[3] ‘Heron's Formula,’ *MathisFun.com*, 2018. [online]. Available: https://www.mathsisfun.com/geometry/herons-formula.html. [Accessed: 11-Dec-2020].