Project 2: Inferring Significant Locations

Author: Bianca Linares

Date: April 1, 2025

Course: CS 4501/6501

GitHub Repository:

https://github.com/blinares-cs/Project-2_Inferring-Significant-Locations/tree/main

Introduction

As a college student, I typically move between a small number of familiar places, such as my apartment, classes, and local stores. These repeated patterns provided an opportunity to uncover insights about the most significant places in my life using location data. Therefore, a project was done to investigate whether data collected through Google Maps Timeline can be used to identify and label frequently visited locations. I hypothesized that if frequent locations are identified through clustering of the data, then the Google Places API will label them accurately. The dataset included 259 visited points recorded between February 5, 2025, and March 25, 2025. After filtering out visits lasting less than five minutes, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm was used to group nearby points into clusters. The centroid of each cluster was calculated, and the Google Places API was used to find the name of the location of each centroid. These labels were then manually reviewed and compared to known locations based on personal experiences to assess labeling accuracy. Through my project, I hoped to evaluate the effectiveness of combining clustering with the Google Places API to extract meaningful insights from my personal location history.

Data Collection and Processing

The dataset for this project was collected using Google Takeout, which provided a file of personal Google Maps Timeline data in JSON format. The analysis focused specifically on the date range from February 5, 2025, to March 25, 2025. Each location was extracted from the exported file to find the data about visit times, geographic coordinates, and other important data for each recorded stop. The start and end timestamps were extracted from each visit entry, and the duration of each visit was computed in minutes by taking the difference between the end and start times. Geographic coordinates were extracted from the "placeLocation" field, which contained a location string formatted as "geo: latitude, longitude." These coordinates were split into latitude and longitude values for use in clustering. In addition to the timestamps and coordinates, other data was extracted for analysis, including the place ID, semantic type, hierarchy level of the visit, the visit's overall probability score, and the probability assigned to the top candidate location. This additional data was included to provide insight into how confidently Google identified each visit. Once the visits were extracted, a data frame was created to store all relevant information in a structured format. The dataset was filtered to retain only visits longer than five minutes to focus the analysis on meaningful stops. This filtering step helped remove brief or accidentally visited points that were unlikely to represent personally significant locations. After filtering, 240 visits remained and were used as the input for the clustering. The latitude and longitude data values were later converted to radians so the Haversine distance metric in the DBSCAN clustering algorithm could be used. By selecting only

the visits with valid duration and filtering out shorter visits, the dataset was better prepared to extract meaningful locations.

Methodology

After filtering the data, clustering was done using the DBSCAN algorithm to identify the frequently visited locations. DBSCAN was selected because it does not require a predefined number of clusters and can group data points based on density. The search radius was set to approximately 100 meters, and a minimum of two points was required to form a cluster. Each group of nearby visit points was assigned to a cluster, but points that did not belong to any group were categorized as noise. The centroid was then calculated by averaging the latitude and longitude values of all visits within the cluster to assign place names to each cluster. These centroid coordinates were shared with the Google Places API, and the API returned the name and associated types for the top result near the cluster center. This information was stored as the representative label for each cluster and was used to understand the type of location it represented. Then, every cluster was reviewed and labeled manually based on personal knowledge of the places visited. Labeling accuracy was determined by comparing the label to the actual location represented by the cluster center. A label was considered correct if it referred to a place that was regularly visited, even if the assigned label was not the exact name of the building. The Google Places API often returned street addresses instead of recognizable place names. For example, a cluster located directly on Rice Hall was labeled with its street address rather than the building name, but it was still marked as correct due to the clear match in location. A label was considered partially correct if it referred to a nearby or adjacent place that was close to the true location but not the actual destination. For instance, if a cluster was

centered near my apartment complex but labeled as a neighboring street, it was categorized as partial. A label was considered incorrect if it pointed to a location that had never been visited or to a place that was incorrectly identified as a stop after actions such as walking or driving. Each cluster was summarized with its location name, place type, average duration, total number of visits, and centroid coordinates. These summaries were merged with the manually assigned accuracy categories to evaluate overall performance. A bar chart was generated to visualize the top 10 most visited clusters. Additional charts were created to display the distribution of accuracy across all clusters.

Results

The results of the methodology were the following figures that visualize the clusters, the performance of the labeling method, and the most highly visited places.

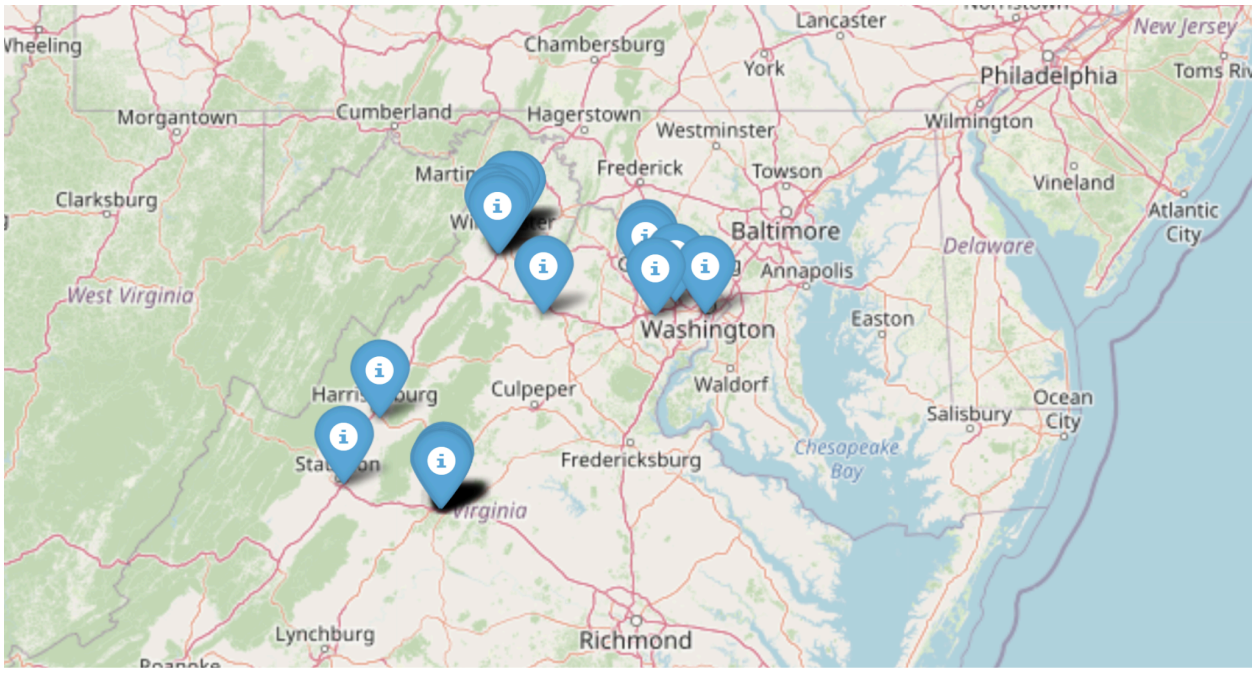

Figure 1: Map of Cluster Centroids

The spatial spread of the identified centroids is shown in Figure 1, where each marker represents a location that was visited repeatedly during the data collection period. The clustering process successfully showed high-density areas in Charlottesville, Winchester, and Northern Virginia, which align with commonly traveled routes and locations of personal importance.

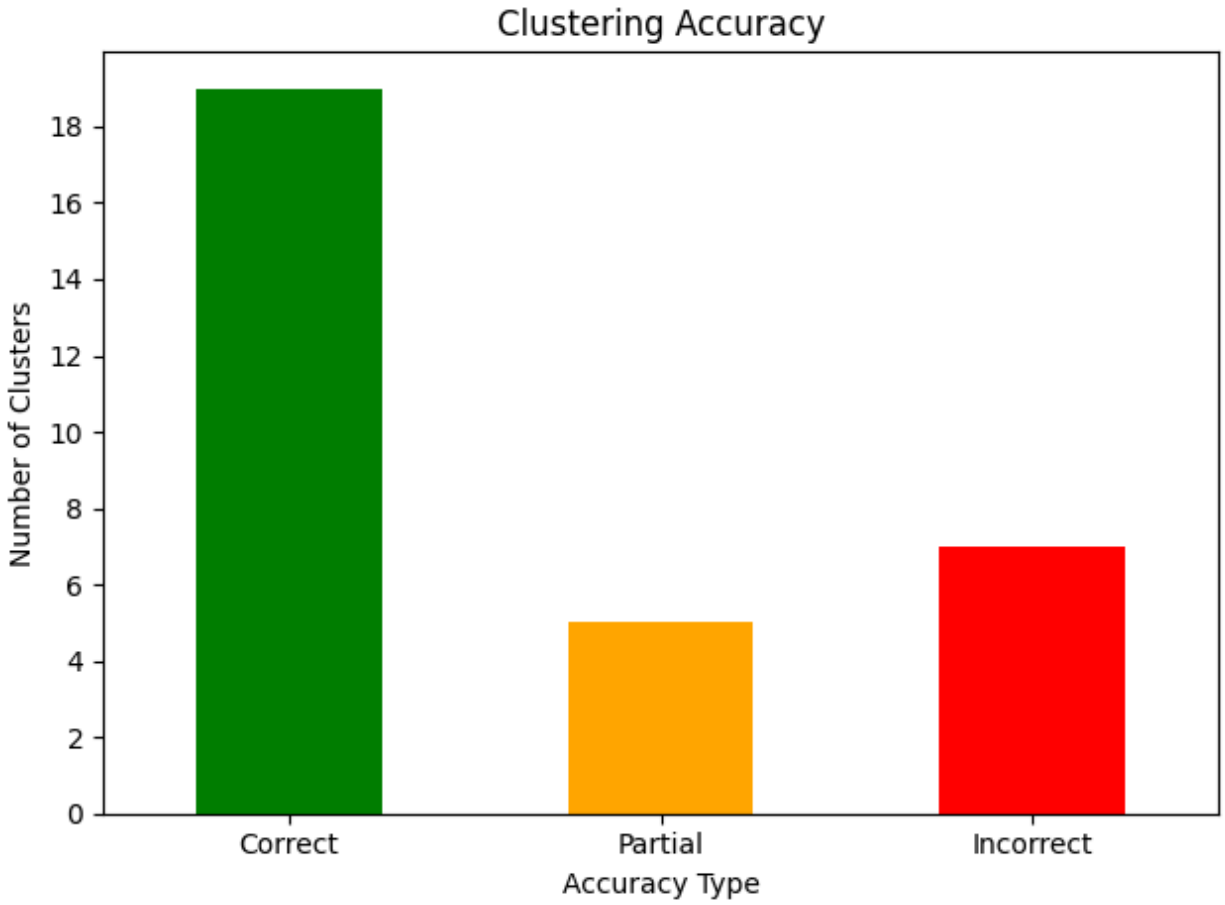

Figure 2: Labeling Accuracy by Cluster Category

Labeling accuracy was determined by manually evaluating each cluster's assigned label against the true location of the cluster center. A label was considered correct if it matched the exact location, partial if it referred to a nearby but not identical place, and incorrect if it represented a location that was not visited. As shown in Figure 2, 19 clusters were correctly labeled, 5 were partially correct, and 7 were incorrect.

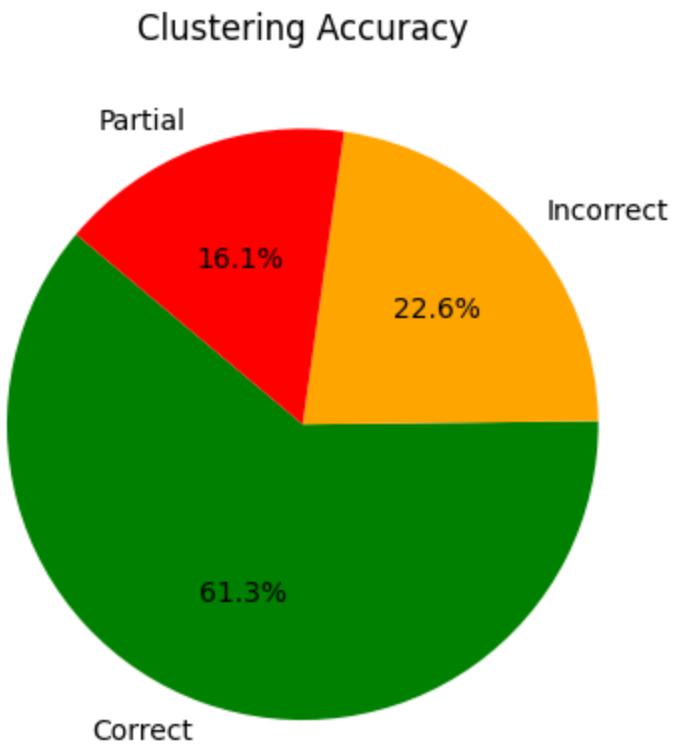

Figure 3: Proportion of Labeling Accuracy Across Clusters

The overall distribution of accuracy is further shown in Figure 3. Correct labels accounted for 61.29 percent of the total, partial labels for 16.13 percent, and incorrect labels for 22.58 percent. While the majority of results were accurate, the presence of partially and incorrectly labeled clusters highlights the limitations of using the Google Places API for precise place recognition, especially in areas where buildings or public spaces are close.

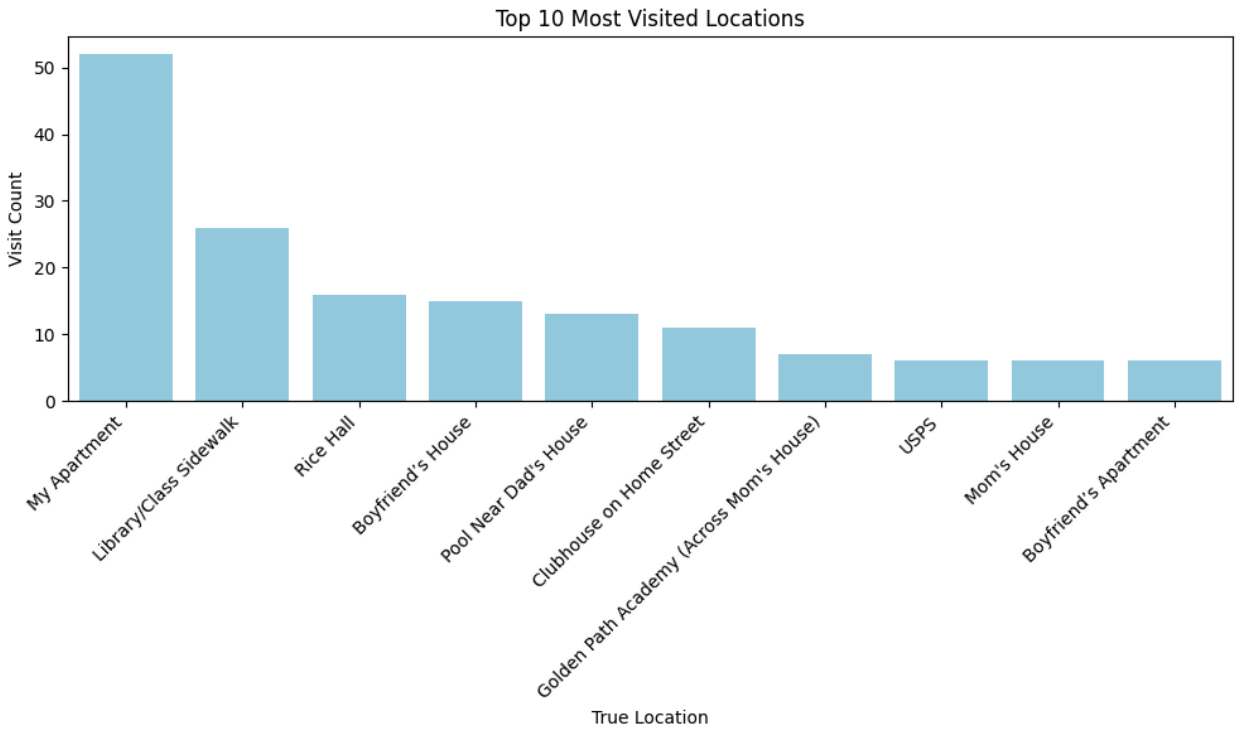

Figure 4: Top 10 Most Frequently Visited Locations

In addition to labeling accuracy, the number of visits per cluster was analyzed to identify the most frequently visited places. Figure 4 displays the top ten locations by visit count. “My Apartment” was the most visited location, followed by “Library/Class Sidewalk,” “Rice Hall,” and “Boyfriend’s House”. These clusters show a combination of personal, academic, and other public spaces that were consistently present in my daily routine.

Overall, these results demonstrate that clustering based on personal location data can effectively uncover significant locations and that public APIs can provide reasonably accurate labels when combined with manual validation. The method was successful in identifying both place frequency and meaning, although some inaccuracies suggest the need for further refinement in fully automated labeling systems.

Limitations

The results demonstrate that clustering personal GPS data can help identify frequently visited locations. However, despite its strengths, the approach had several important limitations. One major limitation was the reliance on the Google Places API for automatic labeling. While the API was generally accurate for well-known or public locations such as grocery stores, it was less reliable in areas where there is a large combination of places such as apartments and stores. In several cases, the API returned a nearby location that was technically correct in terms of proximity but did not reflect the actual destination visited most often, which resulted in a label being categorized as only partially correct. For example, when the cluster center was located at my apartment, the API sometimes returned the name of a nearby road. In other instances, particularly in less densely mapped areas, the API returned locations that had not been visited at all, leading to incorrect labels. Another limitation relates to how labeling accuracy was evaluated. While each cluster was reviewed manually to determine whether the assigned label matched the true location, this process is subjective. The decision to classify a label as partial rather than correct or incorrect depends on context and familiarity with the area. Someone unfamiliar with the region might interpret the API's label differently. The clustering process itself also introduces some challenges. DBSCAN is sensitive to the choice of parameters, such as the search radius and the minimum number of points required to form a cluster. A slightly different radius could lead to clusters being split or merged differently, which would affect both the labeling process and the interpretation of visit frequency. Finally, the dataset was limited by the duration of data collection. Data was collected under two months at the start of the year, so the findings are not necessarily generalizable to other parts of the year. While there were some

limitations, this analysis still offered meaningful insights into how personal location data can reveal significant places in daily life.

Future Work

Future improvements to this project could involve exploring alternative geolocation services that could improve label accuracy by offering more complete or location-specific databases. In addition, expanding the study to include a longer period of time would reveal trends that emerge across different times of the year. As the project scales, developing a more objective and repeatable method for evaluating labeling accuracy will also be important. While manual review based on personal knowledge was effective for this study, future work could benefit from a more systematic approach to validating place predictions.

Conclusion

This project demonstrated that personal data when processed through clustering and combined with labeling, can effectively find meaningful patterns in visited locations. By applying DBSCAN and using the Google Places API to label each cluster, it was possible to visualize key places that are a part of my daily routine. The results partially supported my initial hypothesis that if frequent locations are identified through clustering, then the Google Places API will label them accurately. Manual evaluation of labeling accuracy led to the finding that while many predictions were correct, others were imprecise or incorrect, demonstrating the limitations of relying solely on the Google Places API for place recognition.