CS210 Data Analysis Project

Location Prediction

Rana Taşkın

28387

Instructor: Onur Varol

metin, yazı tipi, grafik, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Introduction**

In an increasingly connected world, understanding human mobility patterns has become vital for various applications, from personalized recommendations to urban planning. This project delves into the intriguing realm of location prediction, seeking to uncover predictable patterns in individual location visits. By analyzing historical location data and employing machine learning techniques, we aim to answer the fundamental question: Can we accurately forecast future location visits? This exploration not only provides insights into human mobility but also holds the potential for practical applications that enhance user experiences and optimize resource allocation.

**Motivation**

The motivation behind this project is to explore and understand the predictability of individual location visits based on historical location data. We aim to answer the question: Can we forecast future location visits with significant accuracy? This project has several practical applications, including personalized location-based recommendations, urban planning, and resource allocation.

**Hypothesis**

The hypothesis of this project posits that individuals' patterns of visiting specific locations display a degree of predictability. These patterns, captured from historical location data, can serve as a foundation for forecasting future visits with a notable level of accuracy. In essence, it seeks to uncover the regularities in people's movements and use these insights to make informed predictions about their future location visits.

“ Individuals exhibit predictable patterns in their location visits, and these patterns can be utilized to forecast future visits with significant accuracy."

**Data Source**

The data for this project was sourced from Snapchat, a popular social media platform known for its location-sharing features. To collect the necessary data, the user utilized Snapchat's location history feature, which records location data along with timestamps. This data, stored in the my Snapchat account, was exported and subsequently used for analysis. It provides a valuable source of information regarding the user's past location visits and forms the basis for exploring the predictability of location patterns.

**Data Collection**

The data collection and preprocessing phase of this project involved several key steps to prepare the Snapchat location history data for analysis. Initially, the user exported their location history data from the Snapchat application, resulting in a JSON-formatted file. The following are the detailed steps undertaken to preprocess the data:

1. Data Extraction: The JSON file containing location history was parsed to extract the relevant information, including timestamps, latitude, longitude, and location accuracy.

2. Timestamp Conversion: The timestamps were converted from their original format ("YYYY/MM/DD HH:MM:SS") to a standardized datetime format to enable time-based analysis.

3. Data Cleaning: The data was thoroughly cleaned to address issues such as missing values and inconsistent formats. Rows with missing or incomplete data were either dropped or imputed as appropriate.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

These preprocessing steps were vital in ensuring the data was ready for analysis, enabling the exploration of location-based patterns and the development of predictive models.

**Exploratory Data Analysis (EDA):**

In data analysis part I used folium and matploit.pylot. Folium was so efficient to analyze the data because I could see the data on a real world map which helped me a lot to analyze and better understand the data. **ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

*Plot for Location Frequencies by Hour of Day*

**öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin, diyagram, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

*Plot for Location Frequencies by Day of Week*

**harita, atlas, metin içeren bir resim

Açıklama otomatik olarak oluşturulduharita, metin, atlas içeren bir resim

Açıklama otomatik olarak oluşturuldu**

*Real world maps integrated with the data created with folium*

**Feature Engineering:**

Feature engineering is a crucial step in the data analysis process, where new features are created or existing ones are modified to improve the model's predictive power. In this project, feature engineering played a significant role in capturing meaningful patterns in the data to enhance the accuracy of predicting future location visits.

Temporal Features: Several temporal features were engineered, including "hour of the day" and "day of the week." These features helped the model understand how visit patterns varied throughout the day and week, capturing daily and weekly routines.

Location-Based Features: Location-based features were generated by calculating the distances between each location point and key places like home and work. These distances provided valuable insights into travel behavior and proximity to significant locations.

Frequency Features: The "visit count" feature was used to represent how frequently an individual visited specific locations. These counts were important for identifying frequently visited places and understanding the popularity of certain locations.

Cluster Features: Clustering techniques were applied to group similar latitude and longitude coordinates into clusters. These clusters served as categorical feature

**Model Selection:**

In this project, the choice of an appropriate model was a crucial step in testing the hypothesis that individuals exhibit predictable patterns in their location visits. The selection of the right model depended on the nature of the problem and the available data. After careful consideration, a classification approach was chosen as it aligned well with the research question and the nature of the target variable.

**Why Classification?**

The hypothesis revolved around predicting an individual's next visited location, which is inherently a classification problem. Locations can be categorized into different classes or labels, making it suitable for classification models.

**Hypothesis Testing:**

Hypothesis testing is a statistical method used to determine whether there is a significant relationship or difference between two or more variables in a dataset. It is a critical step in scientific research to assess the validity of a proposed hypothesis based on empirical evidence.

Null Hypothesis (H0):

The null hypothesis, denoted as H0, represents the default assumption that there is no significant effect or relationship in the data. In the context of your project, the null hypothesis would state that "Individuals do not exhibit predictable patterns in their location visits." Essentially, it assumes that there is no underlying structure or predictability in an individual's choice of locations.

Alternative Hypothesis (Ha):

The alternative hypothesis, denoted as Ha, is the opposite of the null hypothesis. It posits that there is a significant effect or relationship in the data. In your project, the alternative hypothesis would state that "Individuals exhibit predictable patterns in their location visits, and these patterns can be utilized to forecast future visits with significant accuracy." This hypothesis suggests that there are discernible patterns in an individual's location visits, and these patterns can be leveraged to make accurate predictions about their future visits.

metin, ekran görüntüsü, yazılım, multimedya yazılımı içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Model Development and Future Prediction:**For model development I used random grid search and its best parameters for a new model then made predictions using it.

**Findings:**

During the course of this project, I gained valuable insights into my own location visit patterns and habits. It was fascinating to observe the predictability in my visits to specific locations, which confirmed the existence of patterns I had never noticed before. I should spend more time outside rather than my home ☺

**Limitations and Future Work:**

While this project provided valuable insights, it also had its limitations. One limitation is the lack of visit duration data, which could have improved the accuracy of predictions. Additionally, the dataset's granularity could be improved for more precise location-based predictions. In the future, I plan to explore more advanced machine learning techniques and consider integrating additional data sources to enhance the accuracy of location visit forecasts.

**Future Plans:**

Looking ahead, I aim to refine the model and potentially develop a mobile application that can provide location visit forecasts and recommendations based on historical data. This project has opened up possibilities for creating personalized location-based services that can benefit not only me but also others interested in optimizing their daily routines based on location patterns.