実装要件

1. Data Preparation and Groundtruthing Requirements

The goal of this stage is to generate a reliable set of training pairs $\langle m, v(m) \rangle$, where m is a detected GPS stop and $v(m)$ is the truly visited venue (ground truth).

Requirement	Technical Specification	Source Reference
Input Data	Raw GPS trajectories (location, timestamp) and corresponding Foursquare check-in data.	
Trajectory Denoising	Filter GPS points where the distance and speed relative to the predecessor exceed two thresholds: **500 meters** and **50 m/s**.	
Stop Detection (ϵ, τ)	Identify stops m where successive points fall within a small region (ϵ) for a long enough time (τ). Optimal parameters derived via grid search: **$\epsilon = 500$ meters** and **$\tau = 8$ minutes**.	
Data Fusion (Matching)	Match stops to check-ins based on temporal (β) and spatial (α) proximity. Requirements: distance $\le \mathbf{500}$ **meters** (α) and check-in time $\le \mathbf{30}$ **minutes** (β) *after* the stop.	
Train/Test Split	Use a fixed **80% of the matched check-ins for training** and 20% for evaluation.	

2. Model Definition and Feature Engineering

The model uses a Discrete Choice framework, where the probability of choosing a venue v at a stop m is proportional to its score $s(v, m)$.

Requirement	Technical Specification	Source Reference
Candidate Venue Set (V_m)	The set of alternatives V_m for any stop m includes all venues within **500 meters** of the stop	

Requirement	Technical Specification	Source Reference
	location.	
Feature 1: Distance (D)	Calculate the L2 distance between the stop location $m.l$ and the venue location $v.l$.	
Feature 2: Rank (R)	Calculate the number of alternative candidate venues v' in V_m that are **closer** to the stop m than v is.	
Feature Discretization	Both Distance (0 to 500m) and Rank (0 up to max rank) must be discretized into **40 evenly spaced values** (buckets).	
Scoring Function	The optimal score function is the **product** of learned functions (Φ) for distance and rank (the $\Phi(D) + \Phi(R)$ model). $$s(v,m) = \Phi(D) \cdot \Phi(R)$$	

3. Optimization and Learning Algorithm

The model parameters (the coefficients for the 40 distance buckets and 40 rank buckets) are learned by maximizing the Log Likelihood (LL).

Requirement	Technical Specification	Source Reference
Objective Function	Maximize the **Log Likelihood (LL)** over the training data C: $$LL = \sum_C \log s(v(m),m) - \sum_C \log \left(\sum_{v' \in V_m} s(v',m) \right)$$	
Optimization Method	**Gradient Ascent** is used to fit the parameters. The optimization is performed on the functions Φ_D and Φ_R.	
Parameter Update	The learning process relies on calculating the partial derivatives of LL with respect to each parameter $\Phi_D[d_i]$ (or $\Phi_R[r_i]$). The gradient reflects the **discrepancy** between the observed frequency and the expected frequency (based on current weights) for that feature bucket.	
Implementation Note	Although the optimization problem is complex, the LL function is concave in the logarithm of the scores, which allows for efficient optimization using natural **multiplicative update steps** (similar to DBP).	

4. Evaluation Metrics

Model performance should be evaluated on the 20% test dataset using standard Information Retrieval metrics.

Metric	Definition	Source Reference
Log Likelihood (LL)	The maximum LL achieved on the test set.	
NDCG@k	Normalized Discounted Cumulative Gain, evaluating ranking effectiveness at top k positions (e.g., $k=1, 2, 5, 10, 20$).	
MAP	Mean Average Precision, calculating the mean of the Average Precision for each stop.	