Address Clustering Optimization Through Advanced ML and DL Techniques

Aditya Mehta (1226588222)
Andi Lian (1219538359)
Manav Parmar (1229897281)
Soumya Gupta (1229081362)
Vanshaj Gupta (1228730141)
Xinyuan Wang (1230936943)

Agenda

- Problem Definition
- Dataset Description
- State of the Art Methods and Algorithms
- Research and Development Plans
- Design of Experiments and Evaluation Plan
- Project Plan
- References

Problem Definition

- The project is dedicated to overcoming the challenges of address clustering in densely populated areas like India, where non-standardized and incomplete address data often impedes efficient delivery and logistical operations.
- By leveraging advanced Machine Learning (ML) and Deep Learning (DL) techniques, it aims to systematically organize addresses into cohesive clusters based on proximity.
- This initiative is crucial for enhancing delivery route optimization, reducing transit times, and improving address verification processes, ultimately facilitating more effective and cost-efficient logistics solutions in complex urban environments.

Data Set Description

- We will utilize the DeepParse Address Data, sourced from GRAAL-Research's GitHub repository. This dataset comprises 26000 instances, each with four main features.
- To facilitate our analysis, we plan to divide this dataset into training and testing subsets using an 80:20 split, although this ratio may be adjusted based on further research insights.The preprocessing steps will include feature transformation and cleaning, crucial for preparing the dataset for the clustering algorithms we intend to implement.
- This comprehensive dataset will serve as the foundation for our exploration of advanced machine learning and deep learning techniques in address clustering optimization.

State-of-the-Art Methods & Algorithms

- K-Means clustering: Divides data into k clusters by minimizing intra-cluster distances and maximizing inter-cluster distances.
- DBSCAN: Identifies clusters based on density, capable of discovering clusters of arbitrary shape and distinguishing noise.
- Self-organizing maps: Though not a traditional clustering method, SOMs can be used to visualize and cluster high-dimensional data in a lower-dimensional space, preserving topological properties.
- Hierarchical clustering: Builds a tree of clusters by continuously merging or splitting data points, revealing data structure at multiple scales.

State-of-the-Art Methods & Algorithms

- Neural networks: While primarily used for prediction, neural networks can be applied to clustering by learning representations of data that can then be clustered by traditional methods.
- Convolutional neural networks: Specialized in processing structured grid data such as images, using convolution operations to capture spatial hierarchies.
- Recurrent neural networks (LSTM): Designed for sequential data, capable of learning long-term dependencies across time steps.
- Transformer models (BERT): Transformers can generate embeddings for text data that coil dbe clustered to identify patterns or group similar texts.

Research & Development Plan

1. Introduction and Reference Collecting

   To establish a foundation for the project by collecting information we need and reviewing existing research related to clustering, machine learning, and deep learning techniques. Identify the difficulties of doing clustering in current status for non-standardized address formats.

2. Problem Definition Establish

   Find the problem statement based on the information gained from research and reviews. Focus on how to use Machine Learning and Deep Learning to optimize address clustering.

Research & Development Plan

3. Dataset Preparation

   Prepare the clean, structured address dataset for analysis and make it suitable for Machine Learning and Deep Learning models.

4. Methodology Development

   By going through experiments, develop and refine the Machine Learning Model and Deep Learning Model for clustering addresses based on similarity and accuracy. Develop a fully functional UI that is able to run the selected models in the best way.

Research & Development Plan

5. Implementation and Testing

   To implement the chosen models on the prepared dataset and evaluate their performance. By going through refinement and development, make the model achieve optimal performance.

6. Results Analysis and Optimization

   Analyze the results provided by selected models, identify the areas of improvement and update the models accordingly.

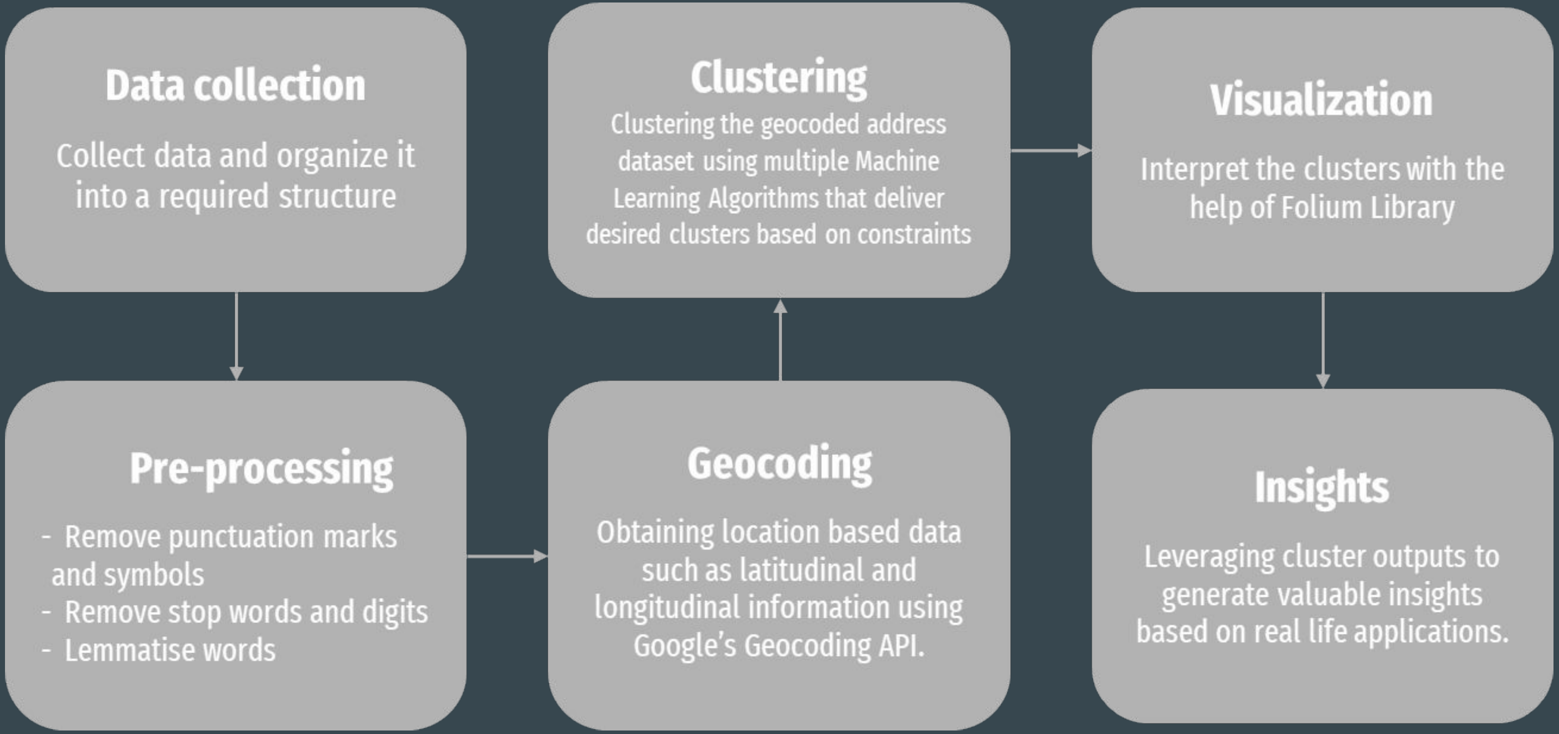
7. Conclusion and Future Work

   Summarize the project findings and discuss the possible applications in the real world.

Design of Experiments and Evaluation Plan

- We will evaluate the performance of our models by using metrics like accuracy, CH Index, and silhouette coefficient.
- Accuracy determines how well the cluster labels match with true labels in supervised settings.
- CH Index evaluates clusters' variance ratios to find the optimal cluster count by maximizing between-cluster variance and minimizing within-cluster variance.
- Silhouette Coefficient measures how well-separated and cohesive clusters are, providing a holistic view of cluster quality and model effectiveness.
- The best performing model will be selected based on its ability to cluster similar addresses together in an efficient manner accurately.

Design of Experiments and Evaluation Plan



Data collection

Collect data and organize it into a required structure

Pre-processing

- Remove punctuation marks and symbols
- Remove stop words and digits
- Lemmatise words

Geocoding

Obtaining location based data such as latitudinal and longitudinal information using Google's Geocoding API.

Clustering

Clustering the geocoded address dataset using multiple Machine Learning Algorithms that deliver desired clusters based on constraints

Visualization

Interpret the clusters with the help of Folium Library

Insights

Leveraging cluster outputs to generate valuable insights based on real life applications.

Project Plan

Task	Description	Deadline	Division of Work
Reference & Problems	Collecting information and reference. Analyze the difficulties that we are facing.	02/19/2024	Xinyuan Wang
Challenge Definition	Identify the specific logical challenges. Get corresponding solutions for the challenges.	02/26/2024	Aditya Mehta
Dataset Preparation	Gather and download datasets we need. Perform data sorting and organizing.	03/11/2024	Manav Kamlesh Parmar
Method Development	Select and evaluate ML and DL models. Optimize models through tests and develop functional UI.	03/25/2024	Soumya Gupta
Implementation & Test	Ready for Project Presentation on April 1, 2024. Apply selected ML and DL models to cluster address datasets.	04/01/2024	Vanshaj Gupta
Result Analysis	Ready for Group Demo Presentation on April 22, 2024. Optimize the models based on analysis to improve performance.	04/22/2024	Everyone Presentation
Conclusion	Ready to turn in Final Project Report on May 1, 2024. Propose directions for future research including possible applications.	04/29/2024	Andi Lian

References

1. Kumar, P. Santhosh, et al. “A Comprehensive Review on Deep Learning Algorithms and Its Applications.” *The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings*, The Institute of Electrical and Electronics Engineers, Inc. (IEEE), 2021, https://doi.org/10.1109/ICESC51422.2021.9532767.
2. Zheng, Baokun, et al. “Identifying the Vulnerabilities of Bitcoin Anonymous Mechanism Based on Address Clustering.” *Science China. Information Sciences*, vol. 63, no. 3, 2020, pp. 132101-, https://doi.org/10.1007/s11432-019-9900-9.
3. Jackson, Jo. *Machine Learning : Proceedings of the Ninth International Workshop (ML92)*. Edited by D. Sleeman et al., Morgan Kaufmann Publishers, 1992.
4. Hahsler, Michael, and Matthew Bolaos. “Clustering Data Streams Based on Shared Density between Micro-Clusters.” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 6, 2016, pp. 1449–61, https://doi.org/10.1109/TKDE.2016.2522412.
5. Küçük Matci, Dilek, and Uğur Avdan. “Address Standardization Using the Natural Language Process for Improving Geocoding Results.” *Computers, Environment and Urban Systems*, vol. 70, 2018, pp. 1–8, https://doi.org/10.1016/j.compenvurbsys.2018.01.009.