

GUIDEBOOK FOR VENDATA PARTICIPANTS

1)Getting Started

Understanding the Problem: Review the hackathon prompt and documentation corpus thoroughly to grasp the problem statement, objectives, and evaluation criteria.

Embrace Learning: Hackathons encourage learning by doing. Don't worry about being an expert; focus on gaining experience and having fun.

Team Collaboration: Form teams with diverse skills, including coding, machine learning, and presentation by abilities, to enhance your solution.

2)Choosing a Model

Exploring Options: Research and explore different large language model architectures, considering factors like model size, computational requirements, and ease of use. We insist you to use nanoGPT.

You can find the resources here: <https://github.com/karpathy/nanoGPT>

3)Data Preparation

Cleaning the Corpus: Preprocess the corpus by cleaning and formatting it effectively for model training, removing noise and inconsistencies.

Augmenting Data: Enhance the corpus by incorporating additional relevant documentation to improve model coverage while maintaining proper data source tracking.

Tokenization and Encoding: Tokenize the text data and encode it into a format suitable for training the chosen model.

4)Model Training

Training Techniques: Train the selected model on the prepared corpus using

appropriate techniques like likelihood maximization, fine-tuning, and transfer learning.

Hyperparameter Optimization: Experiment with hyperparameters such as batch size, learning rates, and epochs to optimize model performance.

Validation Sets: Utilize validation sets to monitor model performance, detect overfitting or underfitting, and determine the optimal stopping point.

Model Interpretability: Explore techniques for interpreting and understanding the model's behavior to gain insights into its decision-making process.

5)Evaluation & Next Steps

Performance Metrics: Evaluate the trained model's performance on provided test queries, focusing on metrics like accuracy, usefulness, and latency.

Identifying Limitations: Identify any limitations or areas for improvement in the solution and provide guidance for future work.

Potential Enhancements: Consider potential extensions or enhancements to the model based on feedback and evaluation results, guiding future iterations.

Deployment Considerations: Explore options for deploying the trained model in a production environment, considering factors like scalability, reliability, and security.

Documentation and Reporting: Document the entire process, including data preprocessing steps, model training configurations, evaluation results, and insights gained. Prepare a comprehensive report or presentation to communicate your findings and solution effectively.

NOTE: If you're new to LLM training or encounter any difficulties, mentors from VenData are available nearby to assist you throughout the hackathon. Embrace the learning experience and have fun building your solution!