Dataset Description for Internship Task

Overview

The dataset provided for your internship task consists of JSON files where each example follows a structured format. The dataset is designed for training a Named Entity Recognition (NER) model using textual data.

Data Format

Each entry in the dataset contains the following fields:

- **tokens**: A list of words representing a sentence or phrase.
- **ner_tags**: A list of numerical labels corresponding to each token, indicating its entity type.
- **lang**: The language of the sentence (e.g., 'en' for English).
- **sequence**: The original sentence as a single string.

Example Data Entry:

Unset

```
{
  "tokens": ["included", "future", "Rage", "Against", "the", "Machine", "and", "Audioslave", "drummer", "Brad", "Wilk", "."],
  "ner_tags": ["O", "O", "O", "O", "O", "O", "O", "O", "O", "B-PER", "I-PER", "O"],
  "lang": "en",
  "sequence": "included future Rage Against the Machine and Audioslave drummer Brad Wilk ."
}
```

Named Entity Recognition (NER) Tags

The dataset employs the BIO tagging scheme for named entity recognition:

- **O**: Outside any named entity.
- **B-PER**: Beginning of a person's name.
- **I-PER**: Inside a person’s name (continuation of the entity).

Entity Distribution in Training Data

The dataset contains the following distribution of entities for training purposes:

- **O**: 639,055
- **B-PER**: 40,264
- **I-PER**: 29,466
- **B-EMAI**L:0
- **I-EMAIL**:0

Test Dataset (For Final Model Evaluation)

A separate dataset is provided exclusively for testing the model's performance after training is complete. This dataset should **not** be used for validation during training.

Entity Distribution in Test Data:

- **O**: 77,866
- **B-PER**: 5,207
- **I-PER**: 3,959
- **B-EMAI**L:0
- **I-EMAIL**:0

Important Notes

- The test dataset is strictly for **final model evaluation** and should **not** be used as a validation set during training.
- Your task involves training an NER model using the training dataset and later evaluating its performance using the test dataset.
- Ensure proper data preprocessing and model evaluation techniques to optimize the recognition of person names in text.

Good luck with your internship task!