Assessment Task: Supplier Data Standardization for Metal Trading

Background
Vanilla Steel is at the forefront of revolutionizing the metal trading industry by providing digital solutions that streamline transactions and enhance market liquidity. A critical component of our operation involves integrating supplier data from various formats into our centralized system. This task will evaluate your ability to process, clean, and standardize supplier inventory data from multiple formats into an unified data structure.

Objective
Develop a prototype solution to automatically extract and standardize supplier data from different file formats into a unified structure.

Task Description
1. **Data Understanding and Exploration**
  - Review the provided datasets to understand the structure and type of data contained in each file.
  - Identify key attributes that need to be extracted and standardized across all datasets.
  - Implement methods to extract text data from the CSV files.
  - You can use any 3rd party libraries such as Use libraries such as `pandas`, `tabula-py`.

2. **Data Processing and Cleaning**
  - Handle inconsistencies in data formats, such as varying column names, data types, and missing values.
  - Clean the data by addressing any anomalies and ensuring consistency in units and descriptions.

3. **Tokenization and Feature Extraction**
  - Tokenize the extracted text using i.e. `spaCy` or similar NLP libraries.
  - Extract relevant features from tokens (e.g., text, position, context).

4. **Classification**
  - Train a machine learning model to classify tokens into predefined categories such as `material_id`, `material_name`, `quantity`, `unit`, `price_per_unit`, `supplier`, `dimensions`, `weight` (feel free to adjust the list, we care about the quality of the solution rather than quantity)
  - Use pre-trained models (e.g., `BERT`, `spaCy's NER`) or train a custom model.

5. **Post-processing**
  - Aggregate the classified tokens into a structured format.
  - Ensure the final output aligns with the defined schema.

6. **Documentation and Presentation**
  - Prepare a brief report or presentation outlining your approach, findings, challenges and recommendations for further improvement.

Deliverables
- Python script containing your solution.
- Python test script validating your solution.
- The standardized dataset in CSV format.
- Report or presentation summarizing your approach and results.
- Clear instructions how to run your script.
- Plan for the future expansion of the algorithm and the potential most difficult challenges.

Evaluation Criteria
- **Technical Proficiency**: Ability to use appropriate tools and libraries for data extraction and classification.
- **Problem-Solving Skills**: Effectiveness of the solution in handling unstructured and inconsistent data.
- **Machine Learning Application**: Ability to train and evaluate a model on the provided datasets.

Next steps
- If you need any additional information or clarification, feel free to reach out to us. We are more than happy to assist.
- If you are unable to fully complete your work, dont worry. Submit everything you have done to the point and provide us written description what have you achieved, what is still TODO and how would you resolve it (written format or pseudo code).

Good luck! We look forward to seeing your innovative solutions.

Files Provided for the Assessment
- [source1](./resources/source1.xlsx)
- [source2](./resources/source2.xlsx)
- [source3](./resources/source3.xlsx)