Entity Recognition in Recipe Data – Project Reports

1. Problem Statement

The primary objective of this project is to build a Named Entity Recognition (NER) model using Conditional Random Fields (CRF) to identify and extract key entities from unstructured recipe ingredient text. This will allow us to structure recipe data into fields like:
- Quantity (e.g., "2", "1/2")
- Unit (e.g., "cup", "tablespoon")
- Ingredient (e.g., "onion", "turmeric powder")

Such structuring facilitates downstream tasks such as automated nutrition analysis, recipe standardization, inventory planning, and grocery list generation.

2. Assumptions Made

- The text tokens are space-separated; multi-word ingredients like "bitter gourd" are split and tagged individually.
- Only three classes are present: quantity, unit, and ingredient.
- No punctuation handling is required, as the raw text appears to be sanitized.
- Labels are token-aligned to the input text.
- The ordering of tokens in "input" and "pos" are aligned 1:1

3. Methodology

a. Data Loading and Prepossessing:
- JSON data is parsed and converted into a tabular structure using pandas.
- Tokens and their corresponding tags are extracted for model training.
- Data is split into training and test sets using train_test_split.

b. Model Training & Evalution:
- A Conditional Random Field (CRF) model is trained using the sklearn-crfsuite package.
- Class weights are computed to handle class imbalance.:
- Precision, Recall, F1-score
- Confusion matrix
- Label-wise classification report using flat_classification_report

4. Techniques Used

- NER Modeling: Using CRF, a popular method for sequence labeling tasks.
- Feature Engineering: Manually derived contextual and morphological features.
- Data Handling: JSON parsing, tokenization, and transformation using pandas and re.
- Visualization: Data distribution plots using matplotlib and seaborn.
- Evaluation: Metrics from sklearn and sklearn_crfsuite.

5. Key Insights

- The CRF model shows high performance in identifying quantity and unit tags due to their consistent patterns.
- Ingredient labels are more challenging due to high variability and multi-word entities.
- Proper feature design, especially around token context, significantly boosts model accuracy.
- The structured output can be readily used for downstream applications such as structured recipe generation or integration into diet tracking systems.

6. Recommendations and Improvements

- Introduce data augmentation with synthetic recipes to increase diversity.
- Explore BiLSTM-CRF models for improved performance using contextual embeddings.
- Address multi-word entity linking, e.g., tagging "bitter gourd" as a single ingredient.

7. Conclusion

This project successfully demonstrates how to structure unstructured recipe ingredient text using a CRF-based Named Entity Recognition approach.
With clean data, thoughtful feature engineering, and proper evaluation, the system achieves effective results for a real-world NLP task in the culinary domain.