CRF Model for Recipe Ingredient Entity Recognition

Problem Statement

The objective of this assignment was to design and implement a Conditional Random Fields (CRF) model to identify and label key entities in recipe ingredient text. The goal was to extract quantities, measurement units, and ingredient names accurately, enabling structured representation of unstructured recipe data.

Methodology

1. **Data Ingestion & Preparation:** Loaded recipe data, performed quality checks, and interpreted key variables.
2. **Train-Validation Split:** Split dataset into 70% training and 30% validation.
3. **Exploratory Data Analysis:** Flattened tokens, analyzed label frequencies, created bar plots, and derived insights.
4. **Feature Extraction:** Implemented token-level and contextual features with domain knowledge (units, methods, regex for numbers).
5. **Model Building & Training:** Trained a CRF model using sklearn-crfsuite with balanced hyperparameters and class weights.
6. **Prediction & Evaluation:** Generated classification report and confusion matrix for validation data.
7. **Error Analysis:** Investigated misclassifications and calculated validation accuracy.

Techniques Used

- Natural Language Processing (tokenization, sequence labeling, feature engineering)

- Conditional Random Fields (CRF) statistical model
- Data Visualization using Seaborn and Matplotlib
- Machine Learning Pipeline: splitting, training, evaluation, error analysis

Key Insights

- Strong performance on frequent labels like **ingredient** and **quantity**.
- Domain-specific lexicons (units, cooking verbs) significantly improved accuracy.
- Most errors involved rare units, plural vs singular confusion, and multi-word ingredients.
- Future improvements: expanding lexicons, using embeddings, and augmenting training data.