Identifying Key Entities in Recipe Data

Objective

This project aims to build a Named Entity Recognition (NER) system tailored for culinary data. Using Conditional Random Fields (CRF), the model is designed to identify and classify tokens in recipe text into categories such as ingredients, quantities, and units. The goal is to transform unstructured ingredient descriptions into structured data, enabling applications like smart recipe organization, nutritional tracking tools, and automated grocery list generation. The dataset consists of annotated ingredient lines, where each token is labeled (e.g., '2' as a quantity, 'cups' as a unit, 'flour' as an ingredient), providing a rich foundation for training a domain-specific NER model.

Approach

- Data Handling
  - The input data is in JSON format, containing labeled ingredient strings.
  - Each token is tagged as either quantity, unit, or ingredient.
- Preprocessing
  - Text cleaning was performed using regular expressions and normalization techniques.
  - Tokenization and POS tagging were applied to enhance structure and consistency.
- Feature Design
  - Lowercased tokens
  - Common suffixes (e.g., '-ed', '-ly')
  - Numeric and fractional detection
  - Contextual features (previous and next tokens)
  - Capitalization and punctuation indicators
- Model Training
  - A CRF model was trained using the sklearn_crfsuite library.
  - CRFs are well-suited for sequence labeling tasks due to their ability to model dependencies between adjacent labels.
- Evaluation
  - Model performance was assessed using precision, recall, F1-score, confusion matrix, and classification reports.

Tools and Techniques

- Natural Language Processing:
  - Tokenization, POS tagging, and domain-specific entity labeling
- Machine Learning:
  - CRF for sequence modeling
  - Evaluation using flat classification metrics
- Data Engineering:
  - JSON-to-DataFrame conversion
  - Visual analysis using Matplotlib and Seaborn
- Model Deployment:
  - The trained model was serialized using joblib for reuse and deployment

Observations and Insights

- Data Exploration
  - Visualizations revealed token frequency distributions and label imbalances
  - Common mislabeling patterns were identified through exploratory analysis
- Model Accuracy
  - The CRF model achieved an impressive accuracy of 99.8%, demonstrating strong performance in identifying entities in recipe text
- Evaluation Highlights
  - Label Confusion: Ambiguous terms like “little” and “taste” were often misclassified, especially between quantity and unit
  - Ingredient Errors: Non-ingredient words (e.g., “is”, “taste”) were sometimes incorrectly labeled as ingredients
  - Class Weights: Adjusting class weights helped mitigate imbalance, though ingredient remained a challenging label
  - Contextual Gaps: Misclassifications often stemmed from insufficient context, especially for tokens like “per” or “of”
  - Boundary Handling: Use of BOS and EOS markers improved segmentation, though edge cases still posed challenges.

Conclusion

This project demonstrates the effectiveness of CRF-based NER systems in extracting structured information from informal, domain-specific text like recipes. With carefully crafted features and domain-aware preprocessing, the model delivers high accuracy and is adaptable to other domains such as:

- Medical prescriptions
- Product catalogs
- Shopping lists

By structuring recipe data, this approach opens the door to advanced applications like personalized meal planning, automated grocery list generation, and nutritional analysis.