# Syntactic Processing of Recipe Data — Report

This assignment focuses on syntactic processing to identify key entities within unstructured recipe data. The primary objective is to extract structured information—specifically, ingredients, quantities, and units—from free-text cooking instructions. Given the semi-structured nature of recipe inputs, the task poses challenges due to variation in format, spelling, and ordering. The ultimate goal is to design a sequence labeling approach that can accurately annotate tokens in recipe texts with their semantic roles.

The dataset provided consists of 285 annotated recipes. Each entry includes a line of recipe text and a corresponding sequence of labels, where each label is aligned token-by-token. The labeling schema includes tags such as quantity, unit, and ingredient. These annotations are essential for training a model capable of extracting structured information from new recipe inputs. The token labeling enables downstream applications such as ingredient-based search, nutritional analysis, and grocery list automation.

The overall methodology includes data loading, preprocessing, statistical exploration, and pattern analysis. The recipes were first parsed into tokens, and their corresponding labels were analyzed to understand frequency and distribution. A key step involved computing statistics on the token count per recipe. The average number of tokens per recipe was found to be approximately 35.6, with most recipes falling between 25 and 45 tokens. The shortest recipe contained 7 tokens, while the longest had 93. This distribution was visualized using a histogram, revealing a bell-shaped curve, indicating consistent formatting with some outliers.

Label frequency analysis showed that ingredient is by far the most frequent label, followed by quantity and unit. This reinforces the observation that recipes often list multiple ingredients, each potentially preceded by a quantity and a unit of measure. To further investigate syntactic patterns, frequent label sequences (n-grams from 2 to 4 tokens) were extracted. The most common sequences included “quantity ingredient,” “ingredient ingredient,” and “quantity unit ingredient.” These findings confirm a regular structure in recipes: quantities and measurement units precede ingredients, and multiple ingredient tokens often appear consecutively.

This assignment demonstrates the potential of syntactic processing in transforming unstructured culinary data into structured formats. By identifying recurring token-label patterns and analyzing distributional properties, valuable insights were derived for building robust NLP models. The results highlight clear trends in recipe structure, which can be leveraged to improve sequence labeling accuracy in future implementations. Further enhancements could include incorporating context-aware models (e.g., BiLSTM or Transformers) and evaluating performance metrics such as precision, recall, and F1-score.