Syntactic Processing Assignment – Report

by Uvaraj Thulasiram

The goal of this project was to identify and extract important information from recipe ingredient lists. I have used **Named Entity Recognition (NER)** to detect and label three key entities in each recipe sentence: quantities (eg: “2”, “1/4”), units (eg: “cups”, “teaspoon”), and ingredients (eg: “rice”, “onion”).

I started by reading the dataset in JSON format, where each sentence was paired with its associated POS tags. This file was provided by Upgrad. I then split the text into tokens and ensured that the number of tokens matched the number of labels. There were a few records with unequal number of tokens and POS tags, which were removed as part of the cleaning process.

Next, I performed **exploratory data analysis (EDA)** to understand the most common ingredients and units. Using seaborn library, I visualized the top 10 ingredients and units. I extracted features from each word using the spaCy NLP library. Features included parts of speech, whether the word was a number, contained a slash (like “1/2”), or matched known quantity and unit keywords. I also added context by including the previous and next words as part of the feature set. Created a regex pattern for quantity.

To address label imbalance (where "ingredient" was far more common), I applied **class weights**, giving more importance to less frequent labels like "quantity". I trained a **Conditional Random Fields (CRF)** model using sklearn-crfsuite and evaluated it on both training and validation sets. The model performed well (99.63% accuracy), with most predictions accurate. Some common errors included confusing units with quantity.

Finally, I did basic error analysis to see where and why the model made mistakes.