Besides character n-grams, word n-grams are also important for identifying languages, because different languages have unique patterns of word combinations. The sequence of words often gives a clearer indication of the language than individual words or characters.

Word n-grams also gives information about the grammar, for example, if a verb shows up at the end of a sentence, maybe it’s German.

Word n-grams can also deal with the disturbing words, like borrowed words from one language to another language.

I set the n-gram range to 2-5, because it offers more information than unigram and the computation is not inefficient.

When trying to find the best parameter, I set the penalty to L1 loss and L2 loss, learning rate to 0.01 and 0.1, solver to stochastic average gradient descent. After running the Grid Search, I got the best parameters: Best parameters: {'clf\_\_C': 0.1, 'clf\_\_penalty': 'l2', 'clf\_\_solver': 'sag'}

The advantage of Grid Search Cross Validation