Part I. (folder-Unigram)

Warm-up: Unigram.py shows the size of the vocabulary and percentage of out of vocabulary tokens

Part II. (folder-model1)

Task: Character based LSTM model

1. Preprocessing data by data.py, including padding, mapping characters to ids and mapping language to ids.
2. Train and validate the model by char_LSTM.py, which calculates the cross entropy and perplexity for training and validation data.

Part III. (folder-model2)

Task: Assign the label based on p(text|label)

1. Preprocessing data by data.py
2. Train model (which embeds the language ids) and assign label based on max p(text|label)
3. Report accuracy on validation data and assign labels to test data