CSE 447 Final Project Milestone 1

Mike He (dh63) Shaoqi Wang (shaoqi) Jiuru Li(lij93)

Winter 2022

1 Dataset

We are going to use the datasets provided by `nltk`, more specifically the brown corpus and the twitter corpus. We will be able to access these data using the `nltk` Python module.

2 Method

We will use the following methods:

- `identify_lang(data)`
  This method will predict the langauge of a given piece of text. We will use a Python libarary called `langid` for it.
- `load_corpus()`
  This method loads our dataset using the `nltk` library.
- `train(model, num_epoch, batch_size, sequence_length, lr, checkpoint_path)`
  This method trains our `pytorch` RNN model using the given prameters (sequence length, batch size, learning rate), with checkpointing support.
- `pred(model, inputs, tokenizer, prefix=None)`
  This method returns a prediction using the given `pytorch` RNN model, a tokenizer, and a history. If `prefix` is `None`, this method finds the three words with the highest probability; if `prefix` is not `None`, this method finds three words with the highest probability while having the prefix.

To run a prediction, we will first figure out which language the sentence is in. Then we choose the corresponding RNN model to get a probability for each word in the vocabulary. If the history contains an incomplete word, we will only pick three words with the prefix and the highest probability to get the next word. Otherwise we find three words out of all words in the vocabulary. The next word will give us the next character.