Ronak Mehta
Instructor: Noah A. Smith
Assignment A5 (Coding)

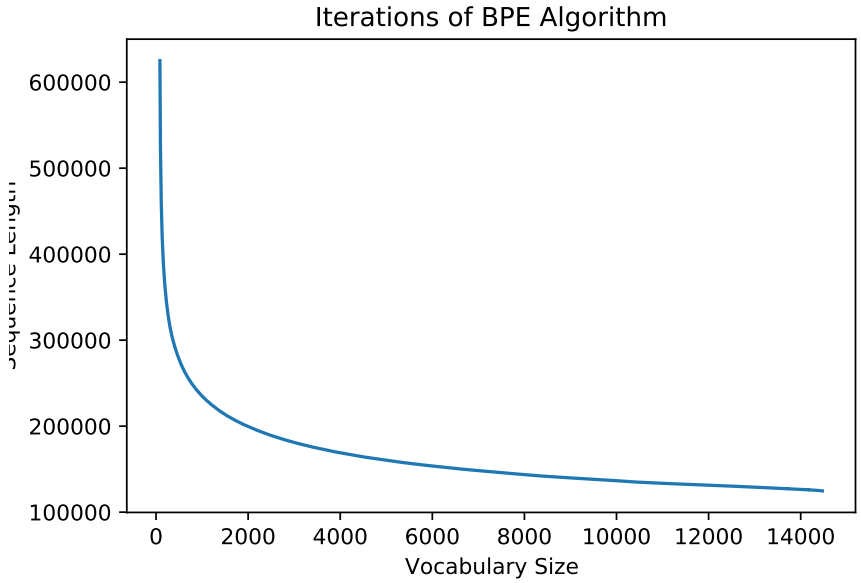
CSE 517: Natural Language Processing
Winter 2021
Due: March 5, 2021

Problem 2

1. It is necessary to append `<s>` to the end of words, as the model needs to learn where to end words if it is to generate readable output. If not, we would not be able to differentiate words in the generated text, and would instead see a sequence of uninterpretable types.

2. No, as there could be ties for the most frequent bigram on any given iteration. In this case, it is ambiguous which bigram to merge, and both merges would be "correct". Consider the string "`i t <s>`". The merges "`it`" and "`<s>`" are both valid, after which merging the entire word together might not be the most frequent bigram.

3. My ending vocabulary size is about $14,000$ types and ending sequence length is about $120,000$ tokens.



4. I used the word "gobbledygook" and did not get an `<unk>`, although I am not sure how we would considering we included all the characters in the vocabulary, so at the very least one encoding with all the characters of the string split exists.

5. I believe the curve depicts the trade-off of the two methods well. The pure character encoding method has the advantage of having a very small vocabulary size, but the sequence length will be very long. If we are using a parallel model like a transformer, this may not be a problem. However, if we are using an autoregressive model such as an RNN, we might prefer BPE's larger vocabulary and shorter sequence length. The same trade-off carries over to the statistical modelling aspect of the problem, as the character-level model will require a longer history but have to decided between fewer classes for the output, and vice versa.