CS 335: Introduction to Large Language Models

Habib University

Activity Sheet 08

Name: ____ ID: ______________

Question 01: Byte Pair Encoding

The Byte Pair Encoding (BPE) algorithm operates by identifying the most frequent adjacent token pairs (bigrams) in a corpus and merging them iteratively into single tokens. Each merge step introduces a new token and a corresponding rule, which is later used for tokenizing words during inference.

Given the pre-tokenized corpus below:

Word	Frequency
low	4
power	3

Apply the BPE algorithm to learn merge rules until the vocabulary reaches a total of 8 tokens. Finally, use the learned merge rules to tokenize the word "lowest" accordingly.