

Module 2

Reading and Short Answer Assignment

Assigned: Wednesday, 05/29/2024

Due: Wednesday, 06/05/2024 by 11:59:59 PM

This week your assignment is to watch the Computerphile video on Unicode:

<https://youtu.be/MijmeoH9LT4?si=tK4wr8rp-PfKZMLN>

and watch an overview of byte-pair encoding (BPE) from Huggingface:

<https://youtu.be/HEikzVL-lZU?si=Ntr9pULZVvZww0Yu>

Questions

Please write a short response (4-5 sentences / short paragraph) to each of the following questions / the following question. Your responses will be graded for accuracy, critical thinking, and clarity. You may use any common word processing or text format. Please upload your answers by the due date.

Question 1: Explain how a byte-level BPE is able to tokenize a character it has never seen before (i.e. an emoji).