Lecture-12: LLM Data processing

4 steps

(1) Tokenization → word based
→ subword based (BPE)
→ character based

(2) Token embeddings → converting token IDs to vector

(3) positional embedding → Encoding information about position.

(4) Input embedding =

(token embedding
+
positional embedding)