Topic 2: Tokenization in LLMs

1. The foundational paper introducing subword tokenization (Byte Pair Encoding): https://aclanthology.org/P16-1162.pdf
2. SentencePiece: A simple and language independent subword tokenizer: https://aclanthology.org/D18-2012.pdf
3. A Deep Dive into Subword Models: https://medium.com/@hexiangnan/byte-pair-encoding-vs-unigram-tokenization-a-deep-dive-into-subword-models-4963246e9a34
4. Explains how word, character, and subword tokenization work: https://christophergs.com/blog/understanding-llm-tokenization
5. Tiktoken Library Tutorial: A practical introduction to OpenAI's tiktoken library: https://www.datacamp.com/tutorial/tiktoken-library-python
6. Short video explaining tokenization basics and comparing BPE, WordPiece, and SentencePiece visually: https://www.youtube.com/watch?v=VFp38yj8h3A
7. HuggingFace BPE: https://huggingface.co/learn/llm-course/en/chapter6/5