Tokenization & BPE Tokenizer — Q&A

Topic: NLP Fundamentals | One-Page Reference

Q1. What is tokenization in NLP?

A. Tokenization is the process of splitting raw text into smaller units called tokens (words, subwords, or characters) that a model can process numerically.

Q2. Why is tokenization important?

A. Models cannot process raw strings; tokenization converts text into integer IDs that map to embeddings, making text mathematically usable.

Q3. What are the main types of tokenization?

A. Word-level (split on spaces), character-level (each character is a token), and subword-level (BPE, WordPiece, Unigram) — subword is most common in modern LLMs.

Q4. What is a vocabulary in tokenization?

A. A fixed set of all possible tokens. Each token is assigned a unique integer ID. Common vocab sizes range from 30K to 100K tokens.

Q5. What is Byte-Pair Encoding (BPE)?

A. BPE is a subword tokenization algorithm that starts with individual characters and iteratively merges the most frequent adjacent pair of symbols until the desired vocabulary size is reached.

Q6. What are the steps of the BPE algorithm?

A. (1) Initialize vocab with all characters. (2) Count all adjacent symbol pairs in the corpus. (3) Merge the most frequent pair into a new symbol. (4) Repeat steps 2–3 until vocab size is reached.

Q7. Give a simple BPE example.

A. Corpus: "low lower lowest". Start: l,o,w,e,r,s,t. Most frequent pair: (l,o) → merge to "lo". Next: (lo,w) → "low". Continue until vocab limit.

Q8. What problem does BPE solve?

A. It handles out-of-vocabulary (OOV) words. Rare or unseen words are broken into known subword pieces rather than mapped to .

Q9. How does BPE balance vocabulary size vs. sequence length?

A. Larger vocab = shorter sequences but more memory; smaller vocab = longer sequences but better generalization. BPE lets you tune this trade-off.

Q10. What is the difference between BPE and WordPiece?

A. BPE merges the most frequent pair; WordPiece (used in BERT) merges the pair that maximizes the likelihood of the training data using a language model score.

Q11. What is tokenizer special tokens?

A. Special tokens like [CLS], [SEP], , , and are added by the tokenizer to mark sentence boundaries, padding, or classification signals for the model.

Q12. What is a token vs. a word?

A. A word is a human-readable unit; a token is a model-defined subword unit. One word can produce multiple tokens (e.g., 'unhappiness' → 'un', 'happiness') or vice versa.