Date Subject

(IV) Test-time tokenization :

- when processing a new word (unseen) :
  - Break it into subwords or letters.
  - merge subwords according to learned rules

Date Byte-pair Encoding Subject BPE

① purpose of BPE
- problem: NLP models often encounter unknown words.
- solution: Represent unknown words as subwords units.
- BPE is a data driven algorithm.

② Vocabulary = set of tokens the model knows.
Tokens can be:
- chracters (a, b, c)
- subwords (low, er)
- Full words (sunflower)

③ How BPE works:
(I) preprocessing:
- start with a corpus.
- normalization
- Tokenization
- lemmatization

(II) before reading further!!
represent it as a sequence of characters
eg: "sunflower" → ['s', 'u', 'n', 'f', 'l', 'o', 'w', 'e', 'r']

(III) Counting pairs:
- count adjecent chracter pairs in our corpus.
- Find the most frequent pairs of tokens
- merge
- add to vocabulary, and replace.