Byte Pair Encoding (Sennrich et al., 2016)

- Given a dictionary of token types and frequences.
  1. Replace the most frequent pair of characters with a $\boxed{\text{new unit}}$. (Record this “merge” operation.)
  2. Repeat until the desired number of merge operations is reached.

Current vocabulary	The new merge
lo**we**r lo**we**st ne**we**r widest	we → $\boxed{\text{we}}$
lo$\boxed{\textbf{we}}$**r** lo$\boxed{\text{we}}$st ne$\boxed{\textbf{we}}$**r** widest	$\boxed{\text{we}}$r → $\boxed{\boxed{\text{we}}\text{r}}$
lo$\boxed{\boxed{\text{we}}\text{r}}$ lo$\boxed{\text{we}}$st ne$\boxed{\boxed{\text{we}}\text{r}}$ widest	