BPE

Split words per character

There may be <w> and <\w> tags to say beginning and end of word

"<w>abc</w> <w>babcba</w>" 🡪 split all

"<w> a b c </w> <w> b a b c b a </w>"

Look for characters that occour in sequence multiple times

* Here we have “a b” – “b a” – “b c” that occour 2 times each, so we can choose the first merge amongst the 3

Word piece

"abc babcba" 🡪 split all

“a b c b a b c b a”

“a b” score = 1/1\*2

“b c” score = 2/3\*2