Tokenizer篇：

LLM的基本运作原理是：将输入内容进行分词映射，变为数字之后输入到模型当中，最后将计算结果编码生成文本。

Tokenizer则是用于将输入内容分词，并且映射到词表中。

目前主流的分词算法就那么几个，我们这里使用BPE进行分词。

借助huggingface的tokenizer库，调用现有分词器进行微调或是从零开始训练都是可以的。

参阅代码以及注释，修改源数据格式，然后即可使用Qwen分词器进行重新生成。