Homework 2

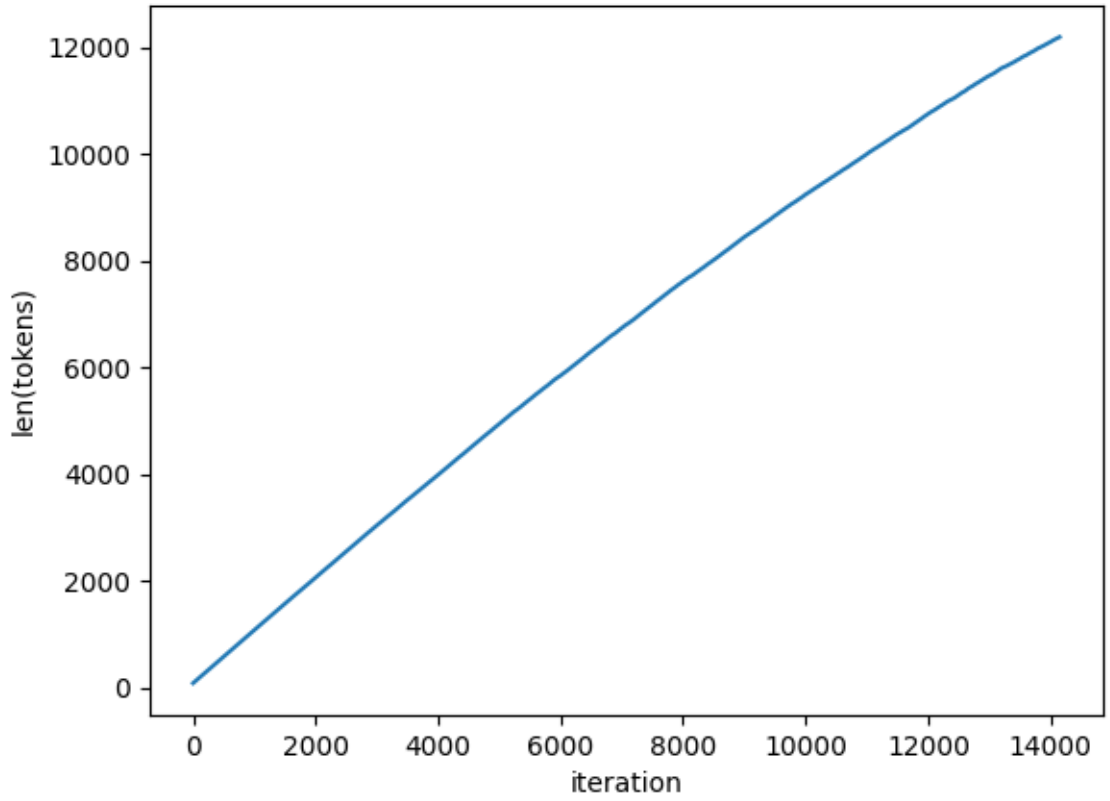

图 1: 随着迭代过程的 token 数量变化

1 Byte-Pair Encoding

1.1 Q1

训练过程中的 token 数量变化如图1所示。最终会有 12196 种 token。在这种词表下，训练数据的长度是 28532。

1.2 Q2

对测试数据进行分词，得到的测试数据长度有 1744。并不会有 <unk> 的情况出现。这是因为，BPE 在分词的过程中，如果遇到没有识别过的词汇，就会跳过这个词汇，而不是用 <unk> 来取代他，这体现了 BPE 很好的扩展性。

2 LLM Tokenizers

接下来我会用同样的两段文本来测试不同的分词器，他们分别是：

“William Whitworth, who worked as a writer and editor at The New Yorker for fourteen years and then served as editor-in-chief of The Atlantic Monthly from 1980 to 1999, died last Friday in Arkansas, the state where he was born. He was a brilliant and intuitive editor who could see around corners and beyond writers’ horizons and deep into thorny manuscripts. Everyone who worked with him will also tell you that he was a prince of a fellow. Throughout publishing you could not find anybody more beloved.”

“这是一段示例文本，包含了一段圆周率的小数部分，3.14159265358979323846264338 3”

2.1 BertTokenizer

分词的结果是：

[’william’, ’w’, ’##hit’, ’##worth’, ’,’, ’who’, ’worked’, ’as’, ’a’, ’writer’, ’and’, ’editor’, ’at’, ’the’, ’new’, ’yorker’, ’for’, ’fourteen’, ’years’, ’and’, ’then’, ’served’, ’as’, ’editor’, ’-’, ’in’, ’-’, ’chief’, ’of’, ’the’, ’atlantic’, ’monthly’, ’from’, ’1980’, ’to’, ’1999’, ’,’, ’died’, ’last’, ’friday’, ’in’, ’arkansas’, ’,’, ’the’, ’state’, ’where’, ’he’, ’was’, ’born’, ’.’, ’he’, ’was’, ’a’, ’brilliant’, ’and’, ’intuitive’, ’editor’, ’who’, ’could’, ’see’, ’around’, ’corners’, ’and’, ’beyond’, ’writers’, ’’ ’, ’horizons’, ’and’, ’deep’, ’into’, ’thorn’, ’##y’, ’manuscripts’, ’.’, ’everyone’, ’who’, ’worked’, ’with’, ’him’, ’will’, ’also’, ’tell’, ’you’, ’that’, ’he’, ’was’, ’a’, ’prince’, ’of’, ’a’, ’fellow’, ’.’, ’throughout’, ’publishing’, ’you’, ’could’, ’not’, ’find’, ’anybody’, ’more’, ’beloved’, ’.’]

[’[UNK]’, ’[UNK]’, ’ —’, ’[UNK]’, ’ 示’, ’[UNK]’, ’ 文’, ’ 本’, ’， ’, ’[UNK]’, ’[UNK]’, ’[UNK]’, ’ 一’, ’[UNK]’, ’[UNK]’, ’[UNK]’, ’[UNK]’, ’ 的’, ’ 小’, ’[UNK]’, ’ 部’, ’ 分’, ’， ’, ’3’, ’.’, ’141’, ’##59’, ’##26’, ’##53’, ’##58’, ’##9’, ’##7’, ’##9’, ’##32’, ’##38’, ’##46’, ’##26’, ’##43’, ’##38’, ’##3’]

对于中文，该分词器会直接按照汉字断开。对于没有见过的词，他会用 <unk> 来代替。对于数字，他会按照几个小的位数断开。

3 BBPE(GPT2)

分词的结果是:

['William', 'ĠWhit', 'worth', ',', 'Ġwho', 'Ġworked', 'Ġas', 'Ġa', 'Ġwriter', 'Ġand', 'Ġeditor', 'Ġat', 'ĠThe', 'ĠNew', 'ĠYorker', 'Ġfor', 'Ġfourteen', 'Ġyears', 'Ġand', 'Ġthen', 'Ġserved', 'Ġas', 'Ġeditor', '-', 'in', '-', 'chief', 'Ġof', 'ĠThe', 'ĠAtlantic', 'ĠMonthly', 'Ġfrom', 'Ġ1980', 'Ġto', 'Ġ1999', ',', 'Ġdied', 'Ġlast', 'ĠFriday', 'Ġin', 'ĠArkansas', ',', 'Ġthe', 'Ġstate', 'Ġwhere', 'Ġhe', 'Ġwas', 'Ġborn', '.', 'ĠHe', 'Ġwas', 'Ġa', 'Ġbrilliant', 'Ġand', 'Ġintuitive', 'Ġeditor', 'Ġwho', 'Ġcould', 'Ġsee', 'Ġaround', 'Ġcorners', 'Ġand', 'Ġbeyond', 'Ġwriters', 'âĢ', 'Ļ', 'Ġhor', 'izons', 'Ġand', 'Ġdeep', 'Ġinto', 'Ġthorn', 'y', 'Ġmanuscripts', '.', 'ĠEveryone', 'Ġwho', 'Ġworked', 'Ġwith', 'Ġhim', 'Ġwill', 'Ġalso', 'Ġtell', 'Ġyou', 'Ġthat', 'Ġhe', 'Ġwas', 'Ġa', 'Ġprince', 'Ġof', 'Ġa', 'Ġfellow', '.', 'ĠThroughout', 'Ġpublishing', 'Ġyou', 'Ġcould', 'Ġnot', 'Ġfind', 'Ġanybody', 'Ġmore', 'Ġbeloved', '.']

['è¿', 'Ļ', 'æĺ', 'ä¸Ģ', 'æ', '®', 'µ', 'ç', '¤', 'º', 'ä', '¾', 'ĭ', 'æ ', 'ĩ', 'æľ', '¬', 'ï', '¼', 'Į', 'åĮ', 'ħ', 'åIJ', '«', 'äº', 'Ĩ', 'ä¸Ģ', 'æ', '®', 'µ', 'åľ', 'Ĩ', 'åij', '¨', 'ç', 'İ', 'ĩ', 'çļĦ', 'å°', 'ı', 'æķ', '°', 'éĥ', '¨', 'åĪ', 'Ĩ', 'ï', '¼', 'Į', '3', '.', '14', '159', '265', '35', '89', '793', '238', '46', '264', '33', '83'] 对于汉字，该分词器也就是按照逐字断开。由于处理的是 unicode，基本上不会有 unk 的情况出现。对于数字，也是按照两三位断一次的方法。