GPT-3 的 BPE(Byte-Pair Encoding) tokenizer

Link: https://github.com/latitudegames/GPT-3-Encoder

这是 GPT-3 使用的 BPE (Byte-Pair Encoding) tokenization 的基础字符表。 （整体的 GPT-3 vocabulary 50257 个 token 是这 256 个字符的组合 ）

通俗讲英语有 26 个字母，GPT 有 256 个“字母”. (本质是个 base256 编码)

有了这 256 个基础字母，根据统计频率还能继续 merge 多个字母来扩充 vocabulary 中的 token 数量到 50257 token。 如 “T”、 “h” 、 “e” 显然的可合并为 "The"

merge map 是在文件 vocab.bpe

vocabulary-50257 是在文件 encoder.json

构建基础 256 关键的 bytes_to_unicode 函数计算过程如下

```python
def bytes_to_unicode():

    bs = (
        list(range(ord("!"), ord("~") + 1))
        + list(range(ord("¡"), ord("¬") + 1))
        + list(range(ord("®"), ord("ÿ") + 1))
    )
    cs = bs[:]
    n = 0
    for b in range(2 ** 8):
        if b not in bs:
            bs.append(b)
            cs.append(2 ** 8 + n)
            n += 1
    cs = [chr(n) for n in cs]
    return dict(zip(bs, cs))
```

1）256 个 base code(bytes/uint8/ascii)：

[33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,161,162,163,164,165,166,167,168,169,170,171,172,174,175,176,177,178,179,180,181,182,183,184,185,186,187,188,189,190,191,192,193,194,195,196,197,198,199,200,201,202,203,204,205,206,207,208,209,210,211,212,213,214,215,216,217,218,219,220,221,222,223,224,225,226,227,228,229,230,231,232,233,234,235,236,237,238,239,240,241,242,243,244,245,246,247,248,249,250,251,252,253,254,255,0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,127,128,129,130,131,132,133,134,135,136,137,138,139,140,141,142,143,144,145,146,147,148,149,150,151,152,153,154,155,156,157,158,159,160,173]

2）新的 256个 base code (去除不能显示的 ascii，拼接了后续 n 个)：

[33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,161,162,163,164,165,166,167,168,169,170,171,172,174,175,176,177,178,179,180,181,182,183,184,185,186,187,188,189,190,191,192,193,194,195,196,197,198,199,200,201,202,203,204,205,206,207,208,209,210,211,212,213,214,215,216,217,218,219,220,221,222,223,224,225,226,227,228,229,230,231,232,233,234,235,236,237,238,239,240,241,242,243,244,245,246,247,248,249,250,251,252,253,254,255,256,257,258,259,260,261,262,263,264,265,266,267,268,269,270,271,272,273,274,275,276,277,278,279,280,281,282,283,284,285,286,287,288,289,290,291,292,293,294,295,296,297,298,299,300,301,302,303,304,305,306,307,308,309,310,311,312,313,314,315,316,317,318,319,320,321,322,323]

3）显示为 unicode (上面是 10 进制的转为 hex 显示)：

["21","22","23","24","25","26","27","28","29","2a","2b","2c","2d","2e","2f","30","31","32","33","34","35","36","37","38","39","3a","3b","3c","3d","3e","3f","40","41","42","43","44","45","46","47","48","49","4a","4b","4c","4d","4e","4f","50","51","52","53","54","55","56","57","58","59","5a","5b","5c","5d","5e","5f","60","61","62","63","64","65","66","67","68","69","6a","6b","6c","6d","6e","6f","70","71","72","73","74","75","76","77","78","79","7a","7b","7c","7d","7e","a1","a2","a3","a4","a5","a6","a7","a8","a9","aa","ab","ac","ae","af","b0","b1","b2","b3","b4","b5","b6","b7","b8","b9","ba","bb","bc","bd","be","bf","c0","c1","c2","

c3","c4","c5","c6","c7","c8","c9","ca","cb","cc","cd","ce","cf","d0","d1","d2","d3","d4","d5","d6","d7","d8","d9","da","db","dc","dd","de","df","e0","e1","e2","e3","e4","e5","e6","e7","e8","e9","ea","eb","ec","ed","ee","ef","f0","f1","f2","f3","f4","f5","f6","f7","f8","f9","fa","fb","fc","fd","fe","ff","100","101","102","103","104","105","106","107","108","109","10a","10b","10c","10d","10e","10f","110","111","112","113","114","115","116","117","118","119","11a","11b","11c","11d","11e","11f","120","121","122","123","124","125","126","127","128","129","12a","12b","12c","12d","12e","12f","130","131","132","133","134","135","136","137","138","139","13a","13b","13c","13d","13e","13f","140","141","142","143"]

4）显示为 unicode 代表的字符：

["!","\"","#","$","%","&","'","(",")","*","+",",","-",".","/","0","1","2","3","4","5","6","7","8","9",":",";","<","=",">","?","@","A","B","C","D","E","F","G","H","I","J","K","L","M","N","O","P","Q","R","S","T","U","V","W","X","Y","Z","
["","\\","]","^","_","`","a","b","c","d","e","f","g","h","i","j","k","l","m","n","o","p","q","r","s","t","u","v","w","x","y","z","
{","|","}","~","¡","¢","£","¤","¥","¦","§","¨","©","ª","«","¬","®","¯","°","±","²","³","´","µ","¶","·","¸","¹","º","»","¼","½","¾","¿","À","Á","Â","Ã","Ä","Å","Æ","Ç","È","É","Ê","Ë","Ì","Í","Î","Ï","Ð","Ñ","Ò","Ó","Ô","Õ","Ö","×","Ø","Ù","Ú","Û","Ü","Ý","Þ","ß","à","á","â","ã","ä","å","æ","ç","è","é","ê","ë","ì","í","î","ï","ð","ñ","ò","ó","ô","õ","ö","÷","ø","ù","ú","û","ü","ý","þ","ÿ","Ā","ā","Ă","ă","Ą","ą","Ć","ć","Ĉ","ĉ","Ċ","ċ","Č","č","Ď","ď","Đ","đ","Ē","ē","Ĕ","ĕ","Ė","ė","Ę","ę","Ě","ě","Ĝ","ĝ","Ğ","ğ","Ġ","ġ","Ģ","ģ","Ĥ","ĥ","Ħ","ħ","Ĩ","ĩ","Ī","ī","Ĭ","ĭ","Į","į","İ","ı","IJ","ij","Ĵ","ĵ","Ķ","ķ","ĸ","Ĺ","ĺ","Ļ","ļ","Ľ","ľ","Ŀ","ŀ","Ł","ł","Ń"]

例子：

中文的 “中” 字

转换为 utf8 是三个字节 \xE4\xB8\xAD 或者表示为 [228, 184, 173]

变为 Unicode 为 \u00e4\u00b8\u143, 代表的字符为 "ä¸Ń"

在 gpt3-vocabulary-50257 中查阅到

"\u00e4\u00b8\u0143": 40792

所以 “中” 的 token 数字为 40792.

BPE 可编码任何字符串，不会出现 <unk> symbol