Device: Apple M3 MacBook Air with 16 GB memory.

Calculate for each of

1. tinystories validation set 22K docs, size = 22.5 MB (27630 occurrences of "<|endoftext|>")
2. Tinystories train set 2.12M docs, size = 2.23 GB (2717699 occurrences of "<|endoftext|>")
3. Openwebtext data validation set 60K docs, size = 290 MB (59059 occurrences of "<|endoftext|>")
4. Openwebtext data train 2.4M docs, size = 11.92 GB (2399397 occurrences of "<|endoftext|>")

Time taken

5. Calculating max in each step (O(N) where N = len(`pair_counts`))
    a. w/o multiprocessing
    b. ww/ multiprocessing
6. With min-heap (O(log(N) where N = len(`pair_counts`))
    a. w/o multiprocessing
    b. ww/ multiprocessing

w/o multiprocessing

1. Tinystories Validation set
    a. O(N)
        i. File Name: 20260107_173439
        BPE Tokenization Summary:
        Total time: **2.01** seconds (0.03 minutes)
        Pre-tokenization time: 1.80 seconds (89.5%)
        BPE merging time: 0.21 seconds (10.5%)
        Vocabulary size: 500
        Number of merges: 243
        Longest token length (bytes): 13
        Longest token id: 256
        Longest token (utf-8, replace): <|endoftext|>
    b. O(logN)
        i. File Name: 20260107_205813
        BPE Tokenization Summary:
        Total time: **2.00** seconds (0.03 minutes)
        Pre-tokenization time: 1.81 seconds (90.4%)
        BPE merging time: 0.19 seconds (9.6%)
        Vocabulary size: 500
        Number of merges: 243
        Longest token length (bytes): 13
        Longest token id: 256
        Longest token (utf-8, replace): <|endoftext|>

2. Tinystories Train set
    a. O(N)
        i. File Name: 20260107_213604
        BPE Tokenization Summary:
        Total time: **219.54** seconds (3.66 minutes)
        Pre-tokenization time: 187.61 seconds (85.5%)
        BPE merging time: 31.92 seconds (14.5%)
        Vocabulary size: 10000
        Number of merges: 9743
        Longest token length (bytes): 15
        Longest token id: 9379
        Longest token (utf-8, replace): responsibility
    b. O(logN)
        i. File Name: 20260107_212055
        BPE Tokenization Summary:
        Total time: **189.35** seconds (3.16 minutes)
        Pre-tokenization time: 186.48 seconds (98.5%)
        BPE merging time: 2.87 seconds (1.5%)
        Vocabulary size: 10000
        Number of merges: 9743
        Longest token length (bytes): 15
        Longest token id: 9379
        Longest token (utf-8, replace): responsibility

With multiprocessing

1. Tinystories Validation set
    a. O(N)
        i. File Name: 20260108_233744
        BPE Tokenization Summary:
        Total time: **0.74 seconds (0.01 minutes)**
        Pre-tokenization time: 0.52 seconds (70.7%)
        BPE merging time: 0.22 seconds (29.3%)
        Vocabulary size: 500
        Number of merges: 243
        Longest token length (bytes): 13
        Longest token id: 256
        Longest token (utf-8, replace): <|endoftext|>
    b. O(logN)
        i. File Name: 20260108_233857
        BPE Tokenization Summary:
        Total time: **0.72 seconds (0.01 minutes)**
        Pre-tokenization time: 0.52 seconds (72.5%)
        BPE merging time: 0.20 seconds (27.5%)

        Vocabulary size: 500
        Number of merges: 243
        Longest token length (bytes): 13
        Longest token id: 256
        Longest token (utf-8, replace): <|endoftext|>

2. Tinystories Train set
    a. O(N)
        i. File Name: 20260108_234056
        BPE Tokenization Summary:
        Total time: **86.62 seconds (1.44 minutes)**
        Pre-tokenization time: 54.38 seconds (62.8%)
        BPE merging time: 32.24 seconds (37.2%)
        Vocabulary size: 10000
        Number of merges: 9743
        Longest token length (bytes): 15
        Longest token id: 9379
        Longest token (utf-8, replace): responsibility
    b. O(logN)
        i. File Name: 20260108_234358
        BPE Tokenization Summary:
        Total time: **57.99 seconds (0.97 minutes)**
        Pre-tokenization time: 55.20 seconds (95.2%)
        BPE merging time: 2.79 seconds (4.8%)
        Vocabulary size: 10000
        Number of merges: 9743
        Longest token length (bytes): 15
        Longest token id: 9379
3. OWT Valid set
    a. O(N) - 2 runs
        i. File Name: 20260108_172005
        BPE Tokenization Summary:
        Total time: **1983.55 seconds (33.06 minutes)**
        Pre-tokenization time: 6.70 seconds (0.3%)
        BPE merging time: 1976.85 seconds (99.7%)
        Vocabulary size: 32000
        Number of merges: 31743
        Longest token length (bytes): 64
        Longest token id: 28060
        Longest token (utf-8, replace):
        ____
        ii. File Name: 20260108_180116
        BPE Tokenization Summary:

Total time: **2315.59 seconds (38.59 minutes)**
Pre-tokenization time: 7.48 seconds (0.3%)
BPE merging time: 2308.11 seconds (99.7%)
Vocabulary size: 32000
Number of merges: 31743
Longest token length (bytes): 64
Longest token id: 28060
Longest token (utf-8, replace):
_____
_____

b. O(logN)
    i. File Name: 20260108_175735
    BPE Tokenization Summary:
    Total time: **107.74 seconds (1.80 minutes)**
    Pre-tokenization time: 7.15 seconds (6.6%)
    BPE merging time: 100.60 seconds (93.4%)
    Vocabulary size: 32000
    Number of merges: 31743
    Longest token length (bytes): 64
    Longest token id: 28060
    Longest token (utf-8, replace):
    _____
    _____
    ii. File Name: 20260108_184250
    BPE Tokenization Summary:
    Total time: **111.81 seconds (1.86 minutes)**
    Pre-tokenization time: 7.20 seconds (6.4%)
    BPE merging time: 104.61 seconds (93.6%)
    Vocabulary size: 32000
    Number of merges: 31743
    Longest token length (bytes): 64
    Longest token id: 28060
    Longest token (utf-8, replace):
    _____
    _____

4. OWT Train set
    a. O(N)
        i. File Name: 20260108_133945
        BPE Tokenization Summary:
        Total time: 11588.80 seconds (193.15 minutes)
        Pre-tokenization time: 282.16 seconds (2.4%)
        BPE merging time: 11306.64 seconds (97.6%)

Vocabulary size: 32000
Number of merges: 31743
Longest token length (bytes): 64
Longest token id: 25822
Longest token (utf-8, replace):
ÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂ

b. O(logN)
   i. File Name: 20260109_162315
      BPE Tokenization Summary:
      Total time: 21018.43 seconds (350.31 minutes)
      Pre-tokenization time: 345.74 seconds (1.6%)
      BPE merging time: 20672.69 seconds (98.4%)
      Vocabulary size: 32000
      Number of merges: 31743
      Longest token length (bytes): 64
      Longest token id: 25822
      Longest token (utf-8, replace):
      ÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂ